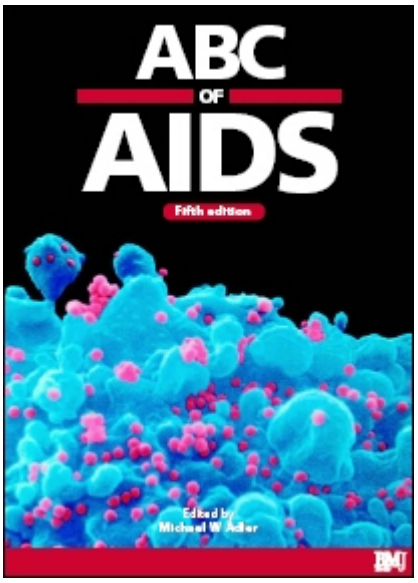
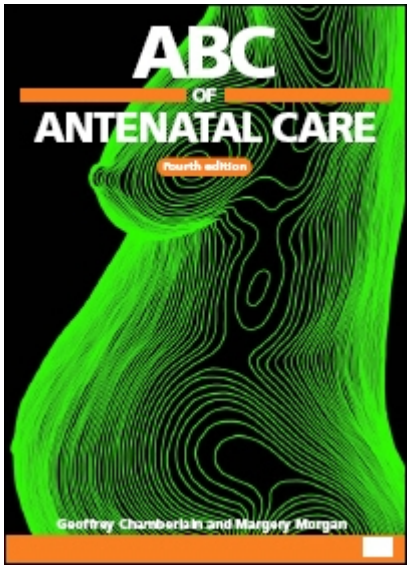
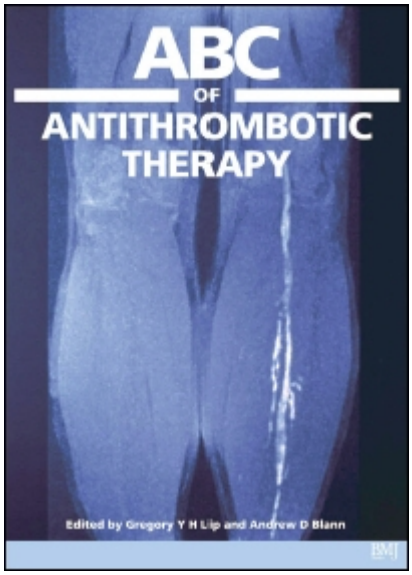
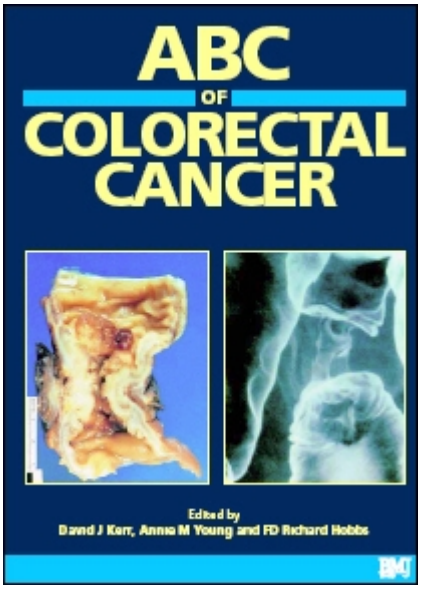
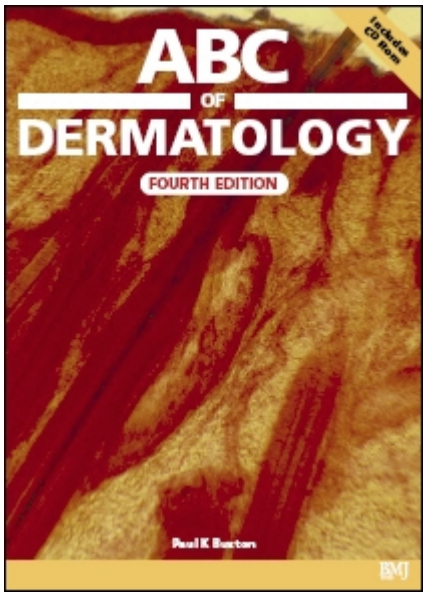
12 IN 1 ABC COLLECTION

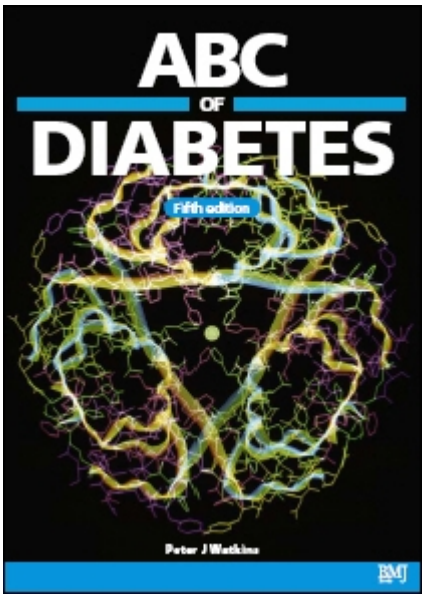


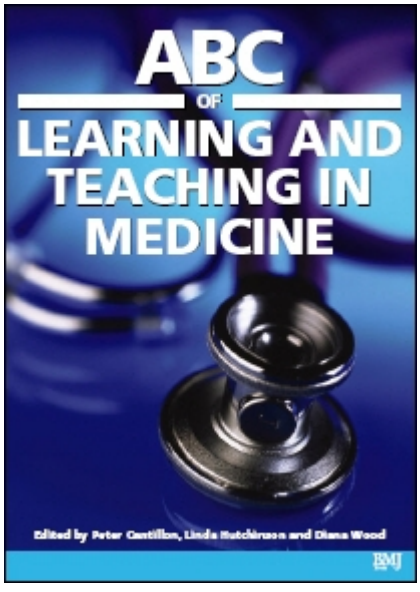


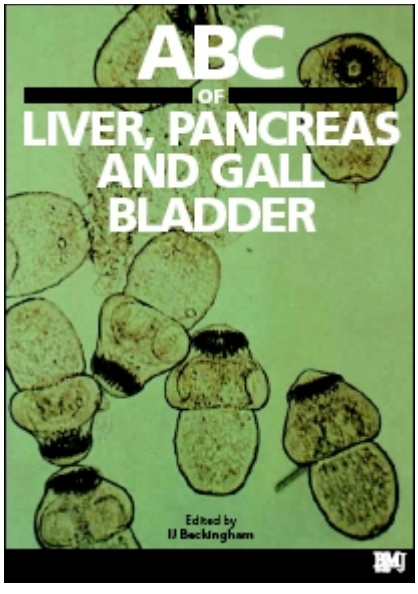


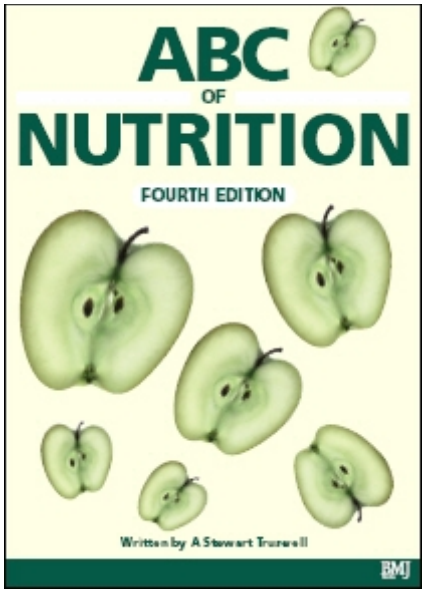


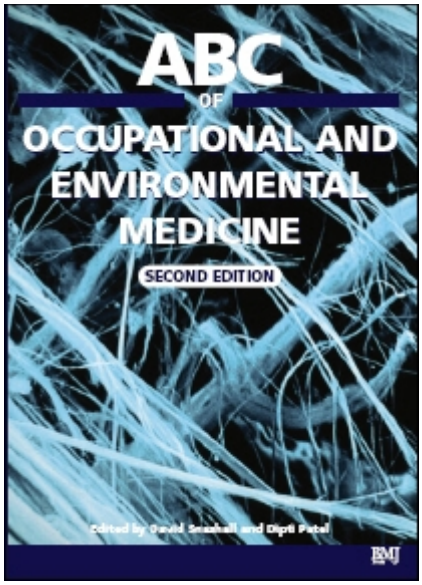


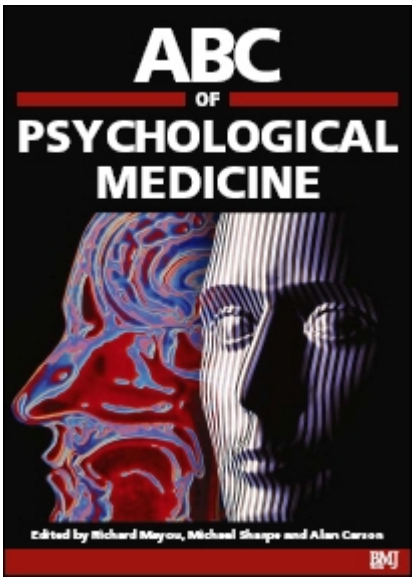









ABC
OF
OCCUPATIONAL AND ENVIRONMENTAL MEDICINE
SECOND EDITION
Edited by David Snashall and Dipti Patel
BMJ


ABC
OF
PSYCHOLOGICAL MEDICINE
Edited by Richard Mayou, Michael Sharpe and Alan Carson
BMJ


ABC
OF
THE FIRST YEAR
Fifth edition
Bernard Valman and Roslyn Thomas
BMJ

ABC OF AIDS

Fifth edition

Edited by
Michael W Adler

BMJ Books

ABC OF AIDS

Fifth Edition

ЗНАНИЕ

ABC OF AIDS

Fifth Edition

Edited by
MICHAEL W ADLER
Professor, Department of Sexually Transmitted Diseases,
Royal Free and University College Medical School, University College, London

BMJ Books

First published in 1987
by the BMJ Publishing Group, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

First edition 1987
Second impression 1987
Third impression 1988
Fourth impression 1988
Fifth impression 1990
Second edition 1991
Third edition 1993
Fourth edition 1997
Sixth impression 1998
Seventh impression 2000
Fifth edition 2001

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0-7279-1503-7

Cover image: NIBSC/Science Photo Library. The image depicts AIDS virus.
Coloured scanning electron micrograph of the surface of a T-lymphocyte (blue) infected with Human Immunodeficiency Virus (HIV).

Cover design by Marritt Associates, Harrow, Middlesex
Typeset by FiSH Books, London
Printed and bound in Spain by Graphycems

Contents

Contributors

Michael W Adler
Professor, Department of Sexually Transmitted Diseases, Royal Free and University College Medical School, University College London, UK

Celia Aitken
Department of Virology, St Bartholomew's and The Royal London, London, UK

Peter Beverley
The Edward Jenner Institute for Vaccine Research, Newbury, UK

RP Brettle
Consultant Physician, Regional Infectious Diseases Unit, Western General Hospital, Edinburgh; Reader in Medicine, University of Edinburgh, UK

Caroline H Bridgewater
Meyerstein Institute of Oncology, Middlesex Hospital, London, UK

Sarah Chippindale
Head of Health Adviser Services HIV/AIDS/GUM, Health Advisers Department, Mortimer Market Centre, London, UK

Kevin M De Cock
Director, CDC Kenya; Visiting Professor of Medicine and International Health, London School of Hygiene and Tropical Medicine, UK

Chris Farnham
Palliative Care Centre, Camden and Islington Community Trust and Royal Free and University College Medical School, University College London, UK

Lesley French
Clinical Psychologist, Camden and Islington CHSNHS Trust, London, UK

Rob George
Palliative Care Centre, Camden and Islington Community Trust and Royal Free and University College Medical School, University College London, UK

Diana Gibb
Senior Lecturer in Epidemiology/Consultant Paediatrician, Institute of Child Health, London, UK

Alison D Grant
Clinical Senior Lecturer, Clinical Research Unit, London School of Hygiene and Tropical Medicine, UK

IJ Hart
Department of Virology, St Bartholomew's and The Royal London, UK

Matthew Helbert
Senior Lecturer, Department of Immunology, St Bartholomew's Hospital, London, UK

John Imrie
Senior Research Fellow, Department of Sexually Transmitted Diseases, Royal Free and University College Medical School, University College London, UK

Anne M Johnson
Professor of Epidemiology, Department of Sexually Transmitted Diseases, Royal Free and University College Medical School, University College London, UK

C Loveday
Professor, Department of Retrovirology, Royal Free and University College Medical School, University College London, UK

Hadi Manji
Consultant Neurologist, National Hospital for Neurology and Neurosurgery and Ipswich Hospital, UK

Ian McGowan
Senior Director Clinical Science, Intrabiotics Pharmaceuticals, California, USA

Rob Miller
Reader in Clinical Infection, Department of Sexually Transmitted Diseases, Royal Free and University College Medical School, University College London, UK

Adrian Mindel
Director of the Sexually Transmitted Infection Research Centre and Professor of Sexual Health Medicine, Westmead Hospital, Sydney, Australia

PP Mortimer
Consultant Virologist, Central Public Health Laboratory, Virus Reference Division, London, UK

Louise Schofield
Clinical Nurse Specialist, Palliative Care Centre,Camden and Islington Community Trust and Royal Free and University College Medical School, University College London, UK

Margaret F Spittle
Consultant Clinical Oncologist, Meyerstein Institute of Oncology, Middlesex Hospital, London, UK

Melinda Tenant-Flowers
Consulted Physician, Department of Sexual Health, The Caldecot Centre, King's Healthcare NHS Trust, London, UK

Gareth Tudor-Williams
Senior Lecturer in Paediatric Infectious Diseases, Imperial College School of Medicine at St Mary's, London, UK

Ian VD Weller
Professor, Department of Sexually Transmitted Diseases, Royal Free and University College Medical School, University College, London, UK

IG Williams
Senior Lecturer, Department of Sexually Transmitted Diseases, Royal Free and University College Medical School, University College, London, UK

Preface

By December 2000 there were 17 538 adult and paediatric patients with AIDS in the UK and 43 774 screened and infected with HIV. Many of those with the virus are well, asymptomatic, and even unaware that they are infected, but others, although they have not yet developed AIDS, have physical, psychological, social, and occupational problems and require as much care as those with AIDS. We therefore need to be concerned not with "a few cases" but with a large number of people infected with the virus, who will be making demands on every part of the health and social services. New infections will occur, and the public health education campaign will need to continue. None of us should feel that the problem of HIV infection and AIDS is unimportant and that it will go away because of the campaign and the possible magic bullet of a cure or vaccine.

We can all hope for these things but it would be a mistake to be lulled into a state of inertia and complacency. All of us will be concerned with AIDS for the rest of our professional lives. This book, originally written as weekly articles for the *BMJ*, attempts to give those doctors and other health care workers, who currently have had little experience of AIDS and HIV, some idea of the clinical, psychological, social and health education problems that they will become increasingly concerned with.

Patients with HIV infection and AIDS spend most of their time out of hospital in the community. Admission is required only when an acute clinical illness supervenes. General practitioners and domiciliary and social services do not always feel skilled and knowledgeable enough to look after them. With the increase in the number of cases, the community services will have to be able and willing to cope. Again, I hope that this book will help to make people feel more skilled and comfortable about caring for patients with HIV and AIDS.

This is the fifth edition of the *ABC of AIDS*; each chapter has been updated or rewritten.

Michael W Adler

1 Development of the epidemic

Michael W Adler

Box 1.1 Early history of the epidemic

1981	Cases of *Pneumocystis carinii* pneumonia and Kaposi's sarcoma in the USA
1983	Discovery of the virus. First cases of AIDS in the UK
1984	Development of antibody test

The first recognised cases of the acquired immune deficiency syndrome (AIDS) occurred in the summer of 1981 in America. Reports began to appear of *Pneumocystis carinii* pneumonia and Kaposi's sarcoma in young men, who it was subsequently realised were both homosexual and immunocompromised. Even though the condition became known early on as AIDS, its cause and modes of transmission were not immediately obvious. The virus now known to cause AIDS in a proportion of those infected was discovered in 1983 and given various names. The internationally accepted term is now the human immunodeficiency virus (HIV). Subsequently a new variant has been isolated in patients with West African connections – HIV-2.

The definition of AIDS has changed over the years as a result of an increasing appreciation of the wide spectrum of clinical manifestations of infection with HIV. Currently, AIDS is defined as an illness characterised by one or more indicator diseases. In the absence of another cause of immune deficiency and without laboratory evidence of HIV infection (if the patient has not been tested or the results are inconclusive), certain diseases when definitively diagnosed are indicative of AIDS. Also, regardless of the presence of other causes of immune deficiency, if there is laboratory evidence of HIV infection, other indicator diseases that require a definitive, or in some cases only a presumptive, diagnosis also constitute a diagnosis of AIDS.

In 1993 the Centers for Disease Control (CDC) in the USA extended the definition of AIDS to include all persons who are severely immunosuppressed (a CD4 count $<200 \times 10^6/l$) irrespective of the presence or absence of an indicator disease. For surveillance purposes this definition has not been accepted within the UK and Europe. In these countries AIDS continues to be a clinical diagnosis defined by one or more of the indicator diseases mentioned. The World Health Organisation (WHO) also uses this clinically based definition for surveillance within developed countries. WHO, however, has developed an alternative case definition for use in sub-Saharan Africa (see chapter 10). This is based on clinical signs and does not require laboratory confirmation of infection. Subsequently this definition has been modified to include a positive test for HIV antibody.

Box 1.2 AIDS-defining conditions without laboratory evidence of HIV

- Diseases diagnosed definitively
  - Candidiasis: oesophagus, trachea, bronchi or lungs
  - Cryptococcosis: extrapulmonary
  - Cryptosporidiosis with diarrhoea persisting >1 month
  - Cytomegalovirus disease other than in liver, spleen, nodes
  - Herpes simplex virus (HSV) infection
    - mucocutaneous ulceration lasting >1 month
    - pulmonary, oesophageal involvement
  - Kaposi's sarcoma in patient <60 years of age
  - Primary cerebral lymphoma in patient <60 years of age
  - Lymphoid interstitial pneumonia in child <13 years of age
  - *Mycobacterium avium*: disseminated
  - *Mycobacterium kansasii*: disseminated
  - *Pneumocystis carinii* pneumonia
  - Progressive multifocal leukoencephalopathy
  - Cerebral toxoplasmosis

Box 1.3 AIDS-defining conditions with laboratory evidence of HIV

- Diseases diagnosed definitively
  - Recurrent/multiple bacterial infections in child <13 years of age
  - Coccidiomycosis – disseminated
  - HIV encephalopathy
  - Histoplasmosis – disseminated
  - Isosporiasis with diarrhoea persisting >1 month
  - Kaposi's sarcoma at any age
  - Primary cerebral lymphoma at any age
  - Non-Hodgkin's lymphoma: diffuse, undifferentiated B cell type, or unknown phenotype
  - Any disseminated mycobacterial disease other than *M. tuberculosis*
  - Mycobacterial tuberculosis at any site
  - Salmonella septicaemia: recurrent
  - HIV wasting syndrome
  - Recurrent pneumonia within 1 year
  - Invasive cervical cancer
- Diseases diagnosed presumptively
  - Candidiasis: oesophagus
  - Cytomegalovirus retinitis with visual loss
  - Kaposi's sarcoma
  - Mycobacterial disease (acid-fast bacilli; species not identified by culture): disseminated
  - *Pneumocystis carinii* pneumonia
  - Cerebral toxoplasmosis

These case definitions are complex and any clinician who is unfamiliar with diagnosing AIDS should study the documents describing them in detail.

CDC Definition of AIDS

Effective 1 January 1993:
All those with confirmed HIV infection with CD4 T lymphocyte count $<0{\cdot}2 \times 10^6/1 \pm$ indicator disease

Transmission of the virus

HIV has been isolated from semen, cervical secretions, lymphocytes, cell-free plasma, cerebrospinal fluid, tears, saliva, urine, and breast milk. This does not mean, however, that these fluids all transmit infection since the concentration of virus in them varies considerably. Particularly infectious are semen, blood, and possibly cervical secretions. The commonest mode of transmission of the virus throughout the world is by sexual intercourse. Whether this is anal or vaginal is unimportant. Other methods of transmission are through the receipt of infected blood or blood products, donated organs, and semen. Transmission also occurs through the sharing or reuse of contaminated needles by injecting drug users or for therapeutic procedures, and from mother to child. Transmission from mother to child occurs *in utero* and also possibly at birth. Finally, the virus is transmitted through breast milk.

Box 1.4 Transmission of the Virus

- Sexual intercourse
  - anal and vaginal
- Contaminated needles
  - intravenous drug users
  - needlestick injuries
  - injections
- Mother → child
  - *in utero*
  - at birth
  - breast milk
- Organ/tissue donation
  - semen
  - kidneys
  - skin, bone marrow, corneas, heart valves, tendons etc.

The virus is not spread by casual or social contact. Health care workers can, however, be infected through needlestick injuries, and skin and mucosal exposure to infected blood or body fluids. Prospective studies in health care workers suffering percutaneous exposure to a known HIV seropositive patient indicate a transmission rate of 0.32%. As of December 1999 there have been 96 reported cases of documented seroconversion after occupational exposure in such workers.

Table 1.1 HIV Transmission: Global Summary

Type of exposure	Percentage of global total
Blood transfusion	3–5
Perinatal	5–10
Sexual intercourse	70–80
(vaginal)	(60–70)
(anal)	(5–10)
Injecting drug use (sharing needles, etc.)	5–10
Health care (needlestick injury, etc.)	<0.01

The precautions and risks for such groups are covered in detail in chapter 15. Finally, there is no evidence that the virus is spread by mosquitoes, lice, bed bugs, in swimming pools, or by sharing cups, eating and cooking utensils, toilets, and air space with an infected individual. Hence, HIV infection and AIDS are not contagious.

Growth and size of the epidemic

Even though North America and Europe experienced the first impact of the epidemic, infections with HIV are now seen throughout the world, and the major focus of the epidemic is in developing/resource-poor countries.

Table 1.2 End-2000 global estimates: children and adults

Categories	Estimate ($\times 10^6$)
People living with HIV/AIDS	36.1
New HIV infections in 2000	5.3
Deaths due to HIV/AIDS in 2000	3.0
Cumulative number of deaths due to HIV/AIDS	21.8

Worldwide
The joint United Nations programme on AIDS (UNAIDS) has estimated that by the end of 2000 there were 36.1 million people living with HIV/AIDS (34.7 million adults and 1.4 million children <15 years). The new infections during that year were 5.3 million, approximately 16,000 new infections per day.

Table 1.3 Regional HIV/AIDS statistics and features, end of 2000

Region	Epidemic started	Adults and children living with HIV/AIDS	Adults and children newly infected with HIV	Adult prevalence rate(*)	% of HIV-positive adults who are women	Main mode(s) of transmission (†) for adults living with HIV/AIDS
Sub-Saharan Africa	late 1970s to early 1980s	25.3 million	3.8 million	8.8%	55%	Hetero
North Africa and Middle East	late 1980s	400 000	80 000	0.2%	40%	Hetero, IDU
South and South-East Asia	late 1980s	5.8 million	780 000	0.56%	35%	Hetero, IDU
East Asia and Pacific	late 1980s	640 000	130 000	0.07%	13%	IDU, hetero, MSM
Latin America	late 1970s to early 1980s	1.4 million	150 000	0.5%	25%	MSM, IDU, hetero
Caribbean	late 1970s to early 1980s	390 000	60 000	2.3%	35%	Hetero, MSM
Eastern Europe and Central Asia	early 1990s	700 000	250 000	0.35%	25%	IDU
Western Europe	late 1970s to early 1980s	540 000	30 000	0.24%	25%	MSM, IDU
North America	late 1970s to early 1980s	920 000	45 000	0.6%	20%	MSM, IDU, hetero
Australia and New Zealand	late 1970s to early 1980s	15 000	500	0.13%	10%	MSM
Total		**36.1 million**	**5.3 million**	**1.1%**	**47%**	

* The proportion of adults (15–49 years of age) living with HIV/AIDS in 2000, using 2000 population numbers.
† Hetero, heterosexual transmission; IDU, transmission through injecting drug use; MSM, sexual transmission among men who have sex with men.

Currently, 95% of all infections occur in developing countries and continents, the major brunt of the epidemic being seen in sub-Saharan Africa and south-east Asia. It is now recognised that cases of AIDS were first seen in Central Africa in the 1970s even though at that time it was not recognised as such. Current surveys from some African countries show that the prevalence of infection is high amongst certain groups – 50–90% of prostitutes, up to 60–70% of those attending departments for sexually transmitted diseases and antenatal clinics. In the developing world, HIV is spread mainly by heterosexual intercourse.

At a family level, UNAIDS estimated that by the end of 1999 the epidemic had left behind a cumulative total of 13.2 million AIDS orphans (defined as those having lost their mother or both parents to AIDS before reaching the age of 15 years). Many of these maternal orphans have also lost their father. Orphans in Zimbabwe are expected to total 1 million by 2005 and 2 million in South Africa by 2010. Traditional family structures and extended families are breaking down under the strain of HIV. Population growth and death rates are increasingly affected. Life expectancy in countries with adult prevalences of over 10% (for example Botswana, Kenya, Zimbabwe, South Africa, Zambia, Rwanda) are expected to see an average reduction in life expectancy of 17 years by 2010–2015. Young, highly productive adults die at the peak of their output, which has a considerable impact on a country's economy.

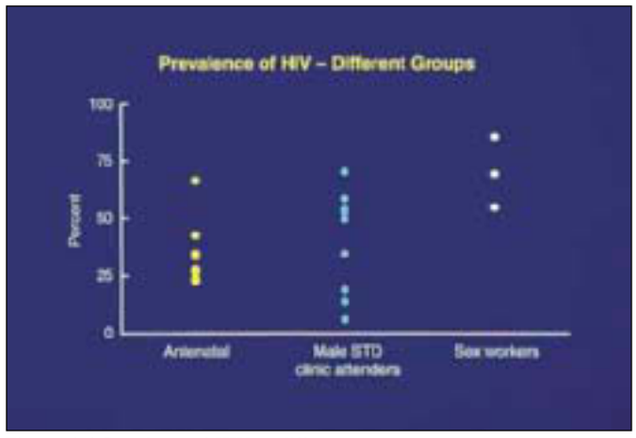


Figure 1.1 Prevalence of HIV — different groups

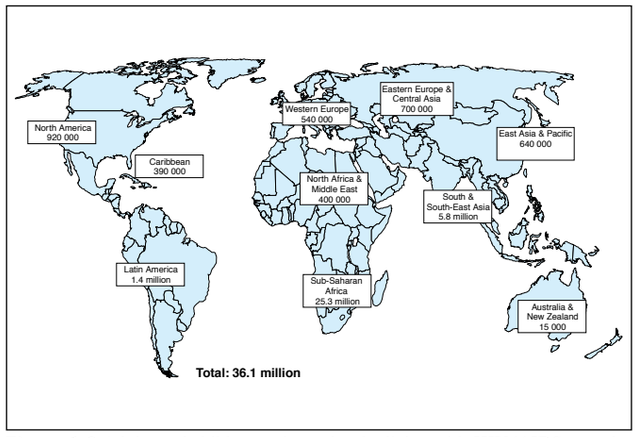


Figure 1.2 Adults and children estimated to be living with HIV/AIDS at end of 2000

USA, UK and Europe

By June 1999, 702 748 adult cases of AIDS had been reported in the USA. In addition there were 8596 paediatric cases (<13 years old). Most of the cases in children (91%) occur because a patient suffered from HIV or belonged to a group at increased risk of HIV; 4% occurred through blood transfusion; 3% in children with haemophilia. Information on risk factors for the remaining 2% of the parents of these children is not complete.

Adult cases in Europe totalled 234 406 by June 2000, and those in the UK 17 151 (December 2000). There are five times more people infected with HIV at any one time than have AIDS. The rate for AIDS cases varies throughout Europe, with particularly high rates in Italy, Portugal, Spain, France and Switzerland, where the commonest mode of infection is through intravenous drug use and the sharing of needles and equipment.

In North America and the UK the first wave of the epidemic occurred in homosexual men. In the UK, proportionally more homosexual men have been notified than in America: 67% of cases compared with 48% respectively. Even though infections amongst men who have sex with men still arise, an increasing proportion of new infections in the USA is occurring amongst intravenous drug users sharing needles and equipment. There is also an increase amongst heterosexuals in both the USA and the UK. Currently in the USA, 16% of cases of AIDS have occurred amongst women, and although the commonest risk factor amongst such women is injecting drug use (42%), the next most common mode of transmission is heterosexual contact (40%).

The nature of the epidemic within the UK is changing with more heterosexual transmission. In the UK 12% of adult cases of AIDS have occurred in women, 70% of which have resulted from heterosexual intercourse. In 2000 there were more new annual infections of HIV than ever before and for the first time more occurring as a result of heterosexual sex than men having sex with men. Most heterosexually acquired infections are seen in men and women who have come from or have spent time in Sub-Saharan Africa.

The advent of an effective antibody test in 1984 has allowed for a clearer understanding of the changing prevalence and natural history of HIV infection. Surveys show that the proportion of individuals infected needs to be high before cases of AIDS start to become apparent. It also underlines the importance of health education campaigns early in the epidemic, when the seroprevalence of HIV is low. Once cases of AIDS start to appear the epidemic drives itself and a much greater effort is required in terms of control and medical care.

Within countries one finds considerable variation in seroprevalence levels for HIV. Over 70% of cases of AIDS and HIV infection within the UK occur and are seen in the Thames regions (London and the surrounding area). Among different groups one also finds geographical differences. For example, the rates among drug users is higher in Edinburgh than London, and for gay men higher in London than anywhere else in the UK. This is also found in the developing world; for example, in Tanzania and Uganda, the urban level of HIV infection in men and women can be five times higher than rural rates.

The use of highly active antiretroviral therapy (HAART) in resource-rich countries has resulted in an increase in life expectancy. This, in combination with the increase in new HIV infections, means that the prevalent pool of those infected, and potentially infectious, is increasing. This presents a continuing challenge for health promotion and a re-statement of the importance of safe sex techniques, particularly condom use (see chapter 16).

Table 1.4 AIDS: adult patient groups in the USA and UK

Patient groups	USA (June 99) n	%	UK (Dec. 00) n	%
Men who have sex with men	334 073	48	11 345	66
Intravenous drug user	179 228	26	1095	6
Men who have sex with men and IV drug user	45 266	6	307	2
Received blood/haemophilia	13 440	2	828	5
Heterosexual contact	70 582	10	3391	20
Other/undetermined	60 159	8	185	1
Total	**702 748**	**100**	**17 151**	**100**

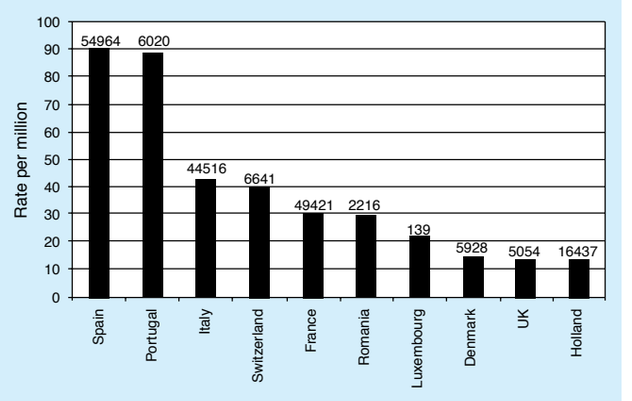


Figure 1.3 AIDS in Europe — top ten countries 1999

Table 1.5 Three main exposure categories (AIDS): % total for various countries in Europe, 1999

	Homosexual/bisexual men	Injecting drug users	Heterosexual exposure
Spain	14.0	65.0	13.0
Italy	14.0	61.0	15.0
Portugal	20.0	47.0	26.0
France	45.0	24.0	20.0
UK	68.0	6.5	18.0
Denmark	67.0	8.0	17.0

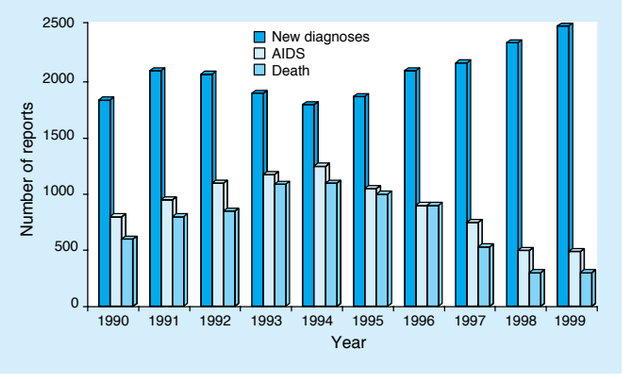


Figure 1.4 New diagnoses, AIDS cases and deaths reported in the year in which they occurred — United Kingdom

AIDS results in a considerable cost not only in human suffering also to health services. Other costs include time off work and the effect of the deaths of young people on national productivity. AIDS represents a major public health problem in the world. A clear understanding of the epidemiology forms the basis of developing a strategy of control ranging from health education to research.

The data on AIDS/HIV in the UK is reproduced with permission from the Communicable Disease Surveillance Centre (CDSC) and the United Nations AIDS Programme.

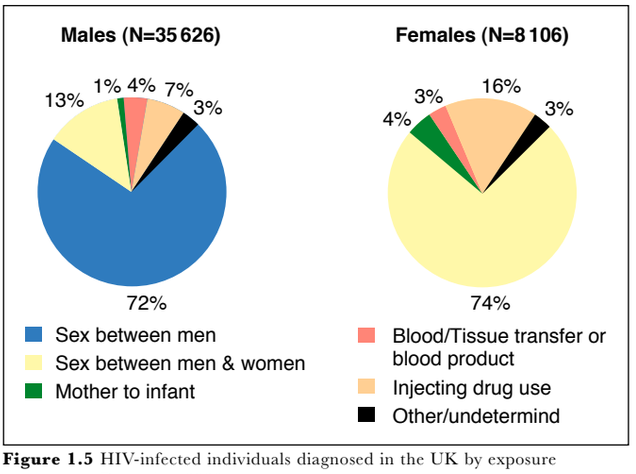


Figure 1.5 HIV-infected individuals diagnosed in the UK by exposure category: to December 2000

2 The virus and the tests

PP Mortimer, C Loveday

Introduction

Although it is clear that HIV is the underlying cause of AIDS and AIDS-related disease, its origin remains obscure. There is firm serological evidence of infection on the east and west coasts of the USA from the mid 1970s, and HIV infection in central Africa may have antedated infection in North America. Phylogenetic analysis of the HIV-1 genome has suggested an origin in chimpanzees while, in the case of HIV-2, similarity to the simian immunodeficiency virus (SIV) genome may point to an origin in sooty mangabey monkeys. In both cases the butchery and consumption of these "bush meats" has been incriminated in transmissions to the human host. Like some other RNA viruses, HIV appears to have mutated and shifted its host range and virulence, explaining how a new pathogenic retrovirus could arise in man. Its virulence may since have been amplified as a result of travel, population dislocation and promiscuous sexual contact, with rapid passage of the virus.

Retroviruses are so named because their genomes encode an unusual enzyme, reverse transcriptase, which allows DNA to be transcribed from RNA. Thus, HIV can make copies of its own genome, as DNA, in host cells such as the human CD4 "helper" lymphocyte. The viral DNA becomes integrated in the lymphocyte genome, and this is the basis for chronic HIV infection. Integration of the HIV genome into host cells is a formidable obstacle to any antiviral treatment that would not just suppress but also eradicate the infection. Nevertheless, modern treatment with combinations of nucleoside analogues and protease inhibitors has transformed the prognosis for carriers of HIV, usually achieving a sustained fall in virus concentration in blood and restoration of the main target cell (CD4 lymphocyte) to near normal levels.

By contrast, the inherent variability of the HIV genome and the failure of the human host to produce neutralising antibodies to the virus, as well as technical difficulties and concerns about safety, have continued to frustrate attempts to make an effective vaccine. This must not, however, allow efforts to develop and evaluate candidate vaccines to slacken. A particular concern is that a useful candidate vaccine (probably a recombinant envelope vaccine developed in North America or Europe against the locally prevalent HIV-1 B subtype) would be ineffective in those parts of the world where other subtypes predominate.

WHO estimates that in the year 2000 there are 36 million carriers of HIV worldwide, and only a small fraction of them have access to suppressive treatment. Both their contacts, their dependants and possibly they themselves would have their life prospects transformed by an effective, or even partially effective, vaccine, and successful application of antiviral treatment in developed countries should in no way be allowed to deflect attention from the necessity of developing and delivering an effective vaccine and of promoting "safe sex" behaviour.

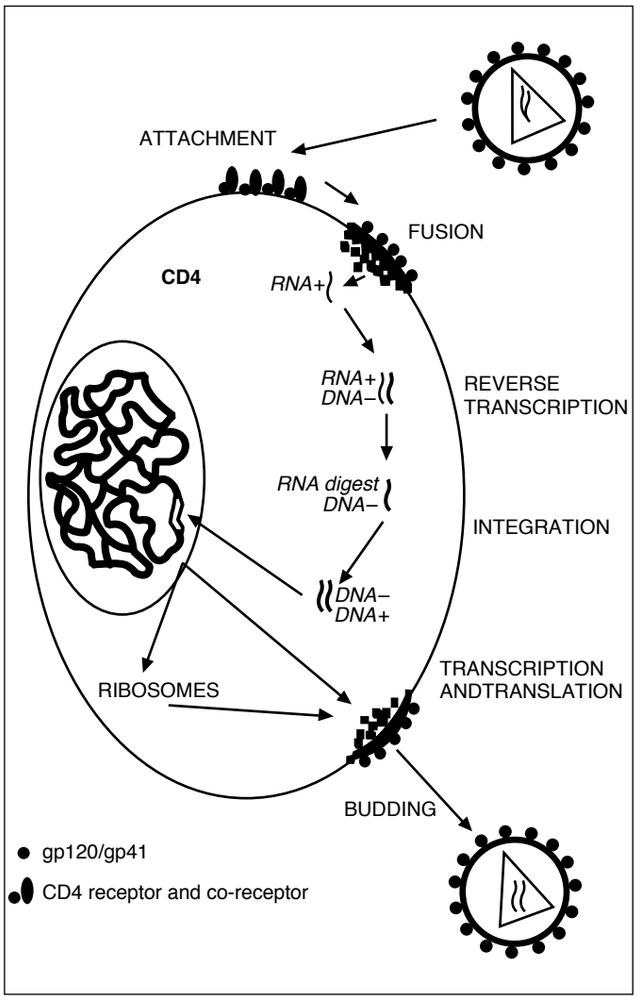


Figure 2.1 HIV replication

HIV and related viruses

HIV was discovered by Barré-Sinoussi, Montagnier, and colleagues at the Institut Pasteur, Paris, in 1983 and given the name lymphadenopathy associated virus (LAV). In 1984 Popovic, Gallo, and co-workers described the development of cell lines permanently and productively infected with the virus. In line with two previously described retroviruses, HTLV-I and HTLV-II, they designated this virus HTLV-III. Other virus isolates from patients with AIDS and AIDS-related disease in America, Europe and Central Africa have proved to be all the same virus, now referred to as HIV-1. Eight subtypes of HIV-1, alphabetically designated, have so far been described.

Around 1985 another human retrovirus, different from HIV-1, was recognised in patients from West Africa. This virus, referred to by the Paris investigators as LAV-2 and more recently as HIV-2, is also associated with human AIDS and AIDS-related disease. It is closely related to the simian retrovirus, SIV, carried by healthy African green monkeys, and the cause of an AIDS-like disease in captive rhesus monkeys. Though potentially important worldwide, HIV-2 infections remain uncommon outside West Africa and they have proved far less virulent than HIV-1 infections.

Transmission of HIV infection

HIV-1 and HIV-2, the major and minor human AIDS viruses, are transmitted in ways that are typical for all retroviruses – "vertically" – that is from mother to infant, and "horizontally" through sexual intercourse and through infected blood. The lymphocytes of a healthy carrier of HIV replicate, and eliminate, over one billion virions each day and the circulating virus "load" may exceed ten million virions per millilitre. At these times viraemia can be recognised by measuring the p24 antigen of HIV in blood and quantifying viral DNA or RNA (see below). Transmission also depends on other factors, including the concentration of HIV secreted into body fluids such as semen, secondary infection of the genital tract, the efficiency of epithelial barriers, the presence or absence of cells with receptors for HIV, and perhaps the immune competence of the exposed person. All infections with HIV appear to become chronic and many are continuously productive of virus. The ultimate risk of spread to those repeatedly exposed is therefore high.

The stage of infection is an important determinant of infectivity. High titres of virus are reached early in infection, though this phase is difficult to study because symptoms may be mild or absent and any anti-HIV response undetectable; it is nevertheless a time when an individual is likely to infect contacts. When, much later, the cellular immune response to HIV begins to fail and AIDS supervenes the individual may again become highly infectious. In the interval between, there may be periods when except through massive exposures – for example blood donation – infected individuals are much less infectious. Nevertheless, in the absence of reliable markers of infectivity, all seropositive individuals must be seen as potentially infectious, even those under successful treatment. Effective ways are constantly being sought to protect their contacts and this has led to the development of the concept of "safe sex". Ideally, this should inform sexual contact between all individuals regardless of whether they are known to be infected with HIV.

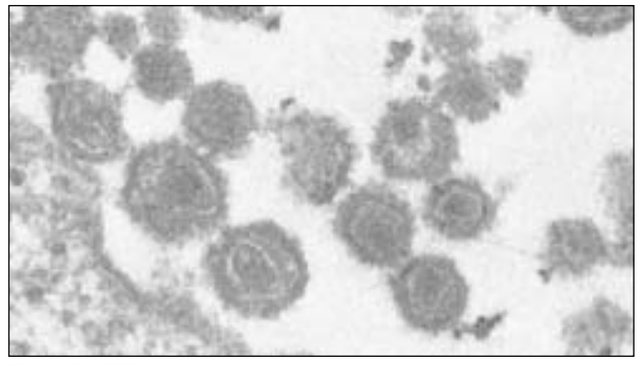


Figure 2.2 HIV particles, many showing typical lentivirus morphology (×118 000)

Box 2.1 Nomenclature of human retroviruses

(a) Two lentiviruses causing AIDS, HIV-1 (previously LAV, HTLV-III) with subtypes A–K, and outliers, HIV-10 and HIV-2
(b) Two oncoviruses causing lymphoma and leukaemia
HTLV-I
HTLV-II

Box 2.2 Transmission factors

- Phase of infection and virus titre
- Local trauma and epithelial damage
- Concurrent sexually transmitted infection
- Intensity of exposure
- Absence of antiretroviral treatment

Figure 2.3 Latex condoms emerging from a dripping tank in the factory
Source: Seohung Industrial Company

Tests for anti-HIV-1 and HIV-2

Anti-HIV tests have transformed our understanding of the epidemiology of AIDS in the years since they were introduced in 1984, and they are still the bedrock of clinical diagnosis and much epidemiological research. Anti-HIV appears three weeks to three months after exposure to HIV and thereafter is invariably detectable in spite of any detrimental effect the virus may have on lymphocyte function and therefore antibody production. Neutralising antibodies to HIV are also measurable, but their titres are low. An inability to mount a neutralising response to HIV antigens together with the mutability of the virus are the most likely reasons why conventional approaches to preparing a vaccine have so far failed.

At first HIV antigen was prepared from infected cell lines. However, antigens can now be made by DNA cloning and expression or by synthesis of viral polypeptides. Several types of anti-HIV test exist, but most use a similar enzyme conjugate and give a colour signal due to the reaction between an enzyme specifically bound onto a polystyrene surface, membrane or inert particles and a substrate that then changes colour. Other tests depend on the binding of a fluorescein or chemiluminescent conjugate, or the visible agglutination of HIV-coated gelatin or latex particles.

Since anti-HIV tests became commercially available in 1985 they have been widely used in diagnostic and transfusion laboratories in the developed world. The accuracy – both sensitivity and specificity – of the antibody assays is continually being improved, and in competent hands the occurrence of false positive and false negative results is less and less frequent. The proportion of true to false positive results depends on the population studied, but even in low risk groups such as volunteer blood donors it is now very high in well conducted laboratories. Human, not test, errors cause most false results, and the key to avoiding these mistakes is continuous review with repeat testing where necessary. All positive reactions should both be confirmed by additional assays and succeeded by a test on a follow-up specimen (see below). The use of several screening tests in parallel on proven positive specimens also acts as a check on the possibility of false negativity in these assays (which it is otherwise difficult to guard against).

More discriminating tests can recognise the components of the antibody response. The serological response to individual HIV proteins can be studied by Western blot, and the immunoglobulin class response to HIV in blood and other fluids can also be investigated. The IgM response slightly proceeds the IgG response early in infection and is indicative of recent infection. Other test procedures, which employ both a highly sensitive and a "detuned" assay for anti-HIV are designed to detect infection within the previous few months and may therefore be used epidemiologically to measure incidence. The IgA anti-HIV response is a feature of infection in infancy.

Simple and non-invasive tests, confirmatory tests, follow-up tests

Simple anti-HIV screening tests have been developed for use in clinics, in unfavourable laboratory conditions and close to the patient. When results are needed urgently, for instance before transplantation procedures and to select a blood donor in the field, they are quick and practical. Saliva (oral fluid) and urine can conveniently be used as specimens to investigate for anti-HIV when venepuncture is difficult, hazardous or unacceptable to the patient. These simple rapid and non-invasive tests are attractive options and may lead to developments such as home

Box 2.3 Anti-HIV

- Appears 3 weeks to 3 months after exposure
- Indicative of infection, except in infants of HIV-positive mothers
- Has weak neutralising capacity
- Persists throughout HIV infection

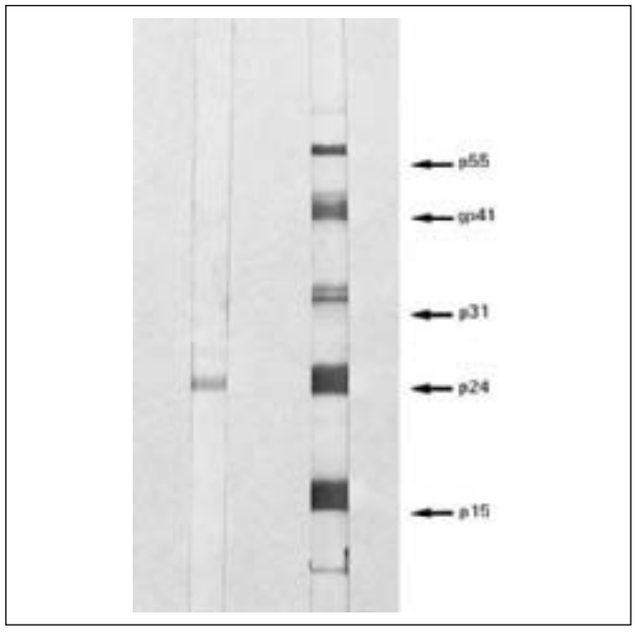


Figure 2.4 The left strip, a Western blot result from a serum specimen collected soon after HIV infection, shows antibody to p24 without other bands being clearly visible. The right strip, a result on a serum sample collected from the same patient 3 months later, shows antibody to many viral proteins, including p15, p24, p31, gp41, and p55

Modern screening kits detect antibody to both HIV-1 and HIV-2. Anti-HIV-2, which is mostly encountered in West Africans and in Europeans who have lived in West Africa, has also been reported in the Indian subcontinent, but it is rare in the Americas. In the UK blood donations and clinical specimens are routinely tested for both infections.

testing. However, few of these tests are quite as accurate as the conventional assays on serum, and follow-up confirmatory tests are essential before a positive diagnosis is made by these means.

In many countries, including the UK, formal procedures have been put in place to secure accurate testing. The most important is that when there is a positive anti-HIV finding the test is repeated and the implicated specimen is tested by other, methodologically independent, anti-HIV assays. Another specimen should then be sought. Although this may cause some delay in confirming a positive finding, anti-HIV testing is as a consequence more precise. A few infected individuals may have little or no detectable anti-HIV when first tested or there may have been technical or clerical mistakes, including specimen misidentifications and transcription errors. Follow-up at an interval of one to four weeks greatly diminishes the chance of either a false negative or a false positive anti-HIV result, and follow-up specimens are the most important element in the accurate laboratory diagnosis of HIV infection. When newly infected individuals are followed up, they show an increase in the titre and range of HIV antibodies. By contrast, persistently weak anti-HIV reactions are usually non-specific. Sometimes PCR (see below) will resolve a difficult-to-confirm antibody reaction. Follow-up procedures also guard against specimen misidentification and transcription errors.

If specimen is anti HIV-positive repeat test using other assay. If HIV-positive take 2nd specimen to confirm.

Test for the virus: antigen, viral DNA and RNA, subtypes, mutants

Viral antigens are present in serum, in particular the HIV core antigen, p24. This is only detectable for as long as it is in excess of antibody to p24, typically at the outset of infection. Tests for this HIV antigen are commercially available, and they assist in the diagnosis of early infection and the recognition of infection in infants. In practice, however, tests for HIV antigen have proved of limited value due to lack of sensitivity, although this may be enhanced by preliminary acid or alkali dissociation of immune complexes in the specimen. Viraemia may also be recognised by isolation of HIV from plasma in cultured lymphocytes, but this is time consuming and not especially sensitive. Essentially it has become a research tool.

HIV can also be detected in specimens in the form of genome sequences. Though only rare lymphocytes carry the HIV genome, the polymerase chain reaction (PCR) can be used greatly to amplify chosen HIV genome sequences in those clinical specimens that contain these small numbers of infected lymphocytes. To a large extent, therefore, viral culture has been superseded by PCR amplification of HIV DNA extracted from mononuclear cells in the circulation. Even more commonly, reverse transcription and amplification of HIV RNA is now being used to detect and quantify virus present in blood. While these procedures are no more accurate than anti-HIV assays and much more expensive, they may be useful in diagnosis, for example in infancy when any anti-HIV detected may be of maternal origin. PCR amplification also provides rapid access to the HIV genome and can lead to characterisation of an HIV isolate to strain level. The (semi) quantification of viraemia (i.e. to within about 0.5 $\log^{10}$) is an important determinant of the need for, and the effect of treatment. It is especially useful as the choice of antiviral combinations widens. Targets for genome amplification include the genes coding for the main envelope, core and transcriptase proteins. On the basis, particularly, of analysis of the sequences of amplified sections of the envelope gene, HIV-1 has been subtyped – so far from A to K. In some cases the sequences found in the various HIV genes are not concordant, showing that recombination occurs in HIV.

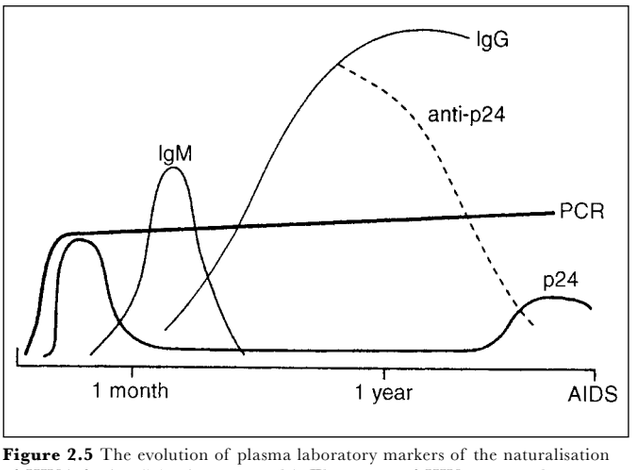


Figure 2.5 The evolution of plasma laboratory markers of the naturalisation of HIV infection ('x' axis not to scale). The course of HIV may now be modified by combined antiviral treatment which will suppress HIV PCR reactivity but not usually modify the anti-HIV response

Sequencing of PCR "amplicons" is also the basis for proving HIV transmission events in special settings, for example, health care.

The growing use of antiretroviral drugs, especially singly, has encouraged the emergence of resistance. This is usually associated with point mutations in the HIV genome. As the common resistance mutations have become better known, testing for them has begun to be used to guide changes in therapy. There is also growing interest in the epidemiology of those mutations that confer resistance for the obvious reason that a highly transmissible resistant mutant might be untreatable and assume an epidemic character.

Testing of patients and blood donors

Tests for anti-HIV-1 and -2, HIV-1 antigen and HIV-1 genome are widely available in the UK. Anti-HIV tests are carried out daily in most public health laboratories and in blood transfusion centres. The facilities in transfusion centres emphatically do not exist to provide testing for those at risk, however. The primary means by which the blood supply is protected from contamination with HIV is through those individuals at increased risk of HIV infection refraining from volunteering to give blood (see chapter 16). Those who wish to be tested for anti-HIV should instead consult their general practitioner or attend a sexually transmitted diseases (genitourinary medicine) clinic, where the advisability of HIV testing can be discussed. If a decision to test is made the necessary investigations are readily and freely available. In some localities "open access" facilities exist to encourage self-referral for counselling and testing. Other innovations, such as home testing on the patient's own initiative, are being considered in the USA and might be introduced into the UK.

As testing becomes more common, and as kits with which people can test themselves are now technically feasible and might be introduced in the future, it is important to be aware of the psychological impact of test findings on those who are tested. While the emergence of effective drug treatment for HIV carriers makes testing for anti-HIV desirable for those who think they may have been put at risk, there should remain an element of medical supervision to respond to patients' questions and anxieties. Telephone helplines have been proposed to provide this support.

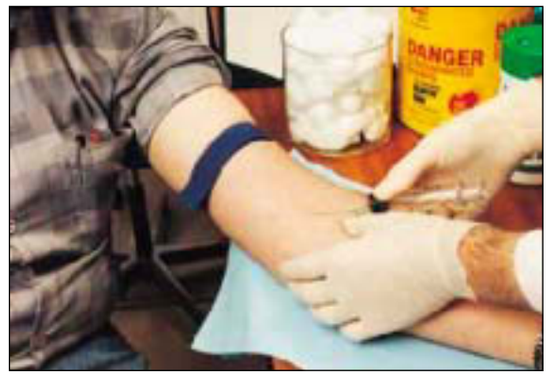


Figure 2.6 Venepuncture to collect a diagnostic specimen. Note the gloved hands and the yellow needleproof container for safe disposal of the needle and syringe.

Important precautions

The desirability of discussing investigations for HIV infection with patients beforehand and of interpreting the results to them afterwards is discussed in Chapter 13. When patients are tested for anti-HIV in a healthcare setting, permission to collect a sample should always have been sought by the doctor and given by the patient. An exception to this is when serum residues, already irreversibly anonymised, are tested for anti-HIV as part of an epidemiological study. Such studies have become a basis for monitoring the epidemic and predicting future trends and resource needs. They have shown, for instance, that in the UK approximately a third of the HIV-infected population (total about 30 000 in year 2000) are unaware of their infection or have not disclosed it at the time of the medical contact.

Clotted blood for testing should be obtained by careful venepuncture without spillage or risk of inoculation accident. The needle and syringe should be disposed of safely and the blood placed in a leakproof container, properly identified, and sent by a secure route to the laboratory. PCR testing requires a fresh EDTA specimen such as commonly used for

Box 2.4 Prevalent HIV infection diagnosed/undiagnosed

30 000 people living with HIV and AIDS in the UK
34% undiagnosed:

Homosexual men	28%
Heterosexual men/women	49%
Injecting drug users	6%

haematological investigations. Oral fluid can be collected from the gum/tooth margin and anti-HIV detected in this fluid. Anti-HIV can also be detected in urine.

The patient's identity and the suspected diagnosis should not be exposed to public gaze, and use of numbers or codes rather than names may be preferred. However, the risk of misidentification may thereby be increased. Patient information should only be shared over the telephone between individuals who know each other, and written reports should be sent to named members of staff, under confidential cover. Positive results should be checked on a fresh newly-drawn specimen. The consequences of breaches of these well-tried procedures may be very serious for patients and damaging to the reputation of doctors. Because of the implications of positive laboratory findings for the health of the patient and his or her family and contacts, and for the patient's social and professional life, a high level of competence and sensitivity is to be expected from all who are concerned in instigating investigation for HIV infection. Testing patients without their informed consent is unacceptable.

Laboratory tests for HIV have increased understanding of AIDS and greatly facilitated diagnosis, management, treatment and control measures. However, to derive most benefit from them and do least harm, tests must be used wisely, with proper regard to all the possible consequences for those who are being tested. Any changes to what are now well-established procedures must be carefully considered, piloted, evaluated for cost-effectiveness, and, if introduced, periodically audited to ensure that they are yielding the benefits promised.

Figure 2.2 was provided by the late JE Richmond and Figure 2.4 by JP Clewley.

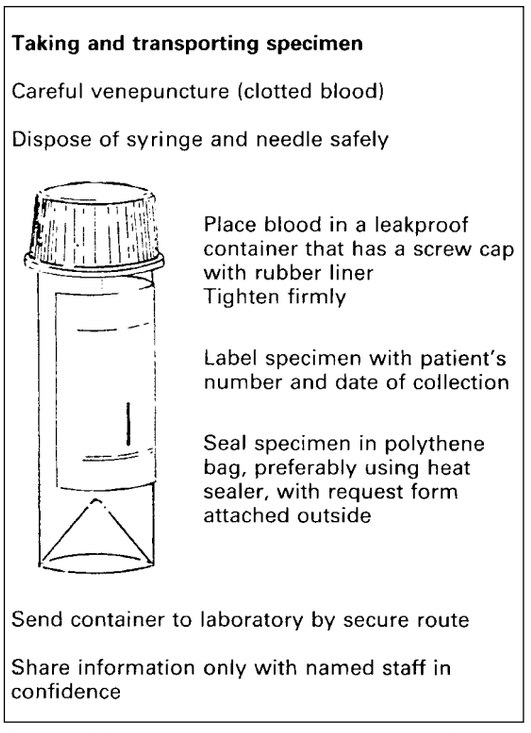


Figure 2.7 Containing and transporting the specimens. Consult the laboratory about appropriate specimens usually clotted blood or blood collected in EDTA

3 Immunology of AIDS

Peter Beverley, Matthew Helbert

Infection with HIV

Many of the clinical features of HIV infection can be ascribed to the profound immune deficit that develops in infected individuals. HIV is immunosuppressive because it infects cells of the immune system and ultimately destroys them. An understanding of this process is helpful in interpreting tests used in monitoring the disease and may explain the failure of immunotherapy and the difficulties in developing vaccines for HIV.

The most obvious target of the virus is a subset of thymus-derived (T) lymphocytes carrying the surface molecule CD4, which has been shown to bind the envelope glycoprotein of HIV (gp120). CD4 is also present on a large proportion of monocytes and macrophages, Langerhans' cells of the skin and dendritic cells of all tissues. More recently it has also become clear that virus entry also requires co-receptors, most of which are members of the seven transmembrane-spanning G protein-coupled receptor family. In the immune system these principally function as receptors for chemokines that orchestrate the migration, differentiation and function of leucocytes during immune responses. Two receptors, CCR5 and CXCR4, are particularly important. CCR5 (R5) is widely expressed on lymphocytes, macrophages, dendritic cells and cells of the rectal, vaginal and cervical mucosae. Virus strains able to infect primary macrophages (macrophage (M) or R5 tropic viruses) use CCR5 as a co-receptor. Only R5 strains are detected early after infection, while both R5 viruses and strains that infect T cells and use CXCR4 (T or X4 tropic viruses) are found late in infection. These data suggest that R5 strains are important for transmission of HIV while X4 variants arise during the course of infection and may be responsible for T-cell loss and disease progression. Even stronger evidence that CCR5-using M tropic viruses transmit infection, comes from the observation that individuals homozygous for a 32 base pair deletion of CCR5 show greatly increased resistance to HIV infection. Several other chemokine receptors have been shown capable of acting as co-receptors *in vitro* and polymorphisms in CCR2 as well as CCR5 and SDF1 (the ligand for CXCR4), are associated with different rates of progression to AIDS.

Table 3.1 Co-receptors and their ligands. A large number of seven transmembrane-spanning receptors which can act as co-receptors have been identified. In most cases their importance *in vivo* remains to be determined. Many are present on some CD4 cells or macrophages

	Type	Ligand
Chemokine receptor		
CXCR4	CXC	SDF-1
CCR2	CC	MCP-1, MCP-2, MCP-3
CCR3	CC	Eotaxin, RANTES, MIP-1α, MCP-3, MCP-4
CCR5	CC	MIP-1α, RANTES, MIP-1β
CCR8	CC	I-309
CCR9	CC	CC chemokines
CX_3CR1	CX_3C	Fractalkine
Orphan receptors		
AJP	?	?
ChemR23	?	?
GPR15/BOB	?	?
STRL33/Bonzo	?	?
Other receptors		
BLTR		Leukotriene B_4
US28	Viral (CMV)	RANTES, MIP-1α, MIP-1β, MCP-1

Immunopathology

CD4 lymphocytes (T helper cells) have been termed "the leader of the immunological orchestra" because of their central role in the immune response, and their destruction accounts at least in part for the immunosuppressive effect of the virus. When these cells are stimulated by contact with an antigen they respond by cell division and the production of lymphokines, such as interferons, interleukins, tumour necrosis factor and the chemoattractant chemokines. Lymphokines act as local hormones controlling the growth, maturation and behaviour of other lymphocytes, particularly the cytotoxic/suppressor (CD8) T-cells and antibody-producing B lymphocytes. Lymphokines also affect the maturation and function of monocytes, tissue macrophages and dendritic cells.

Macrophages and particularly dendritic cells are important antigen-presenting cells for initiating immune responses of lymphocytes. Not only do they act as a reservoir for the virus

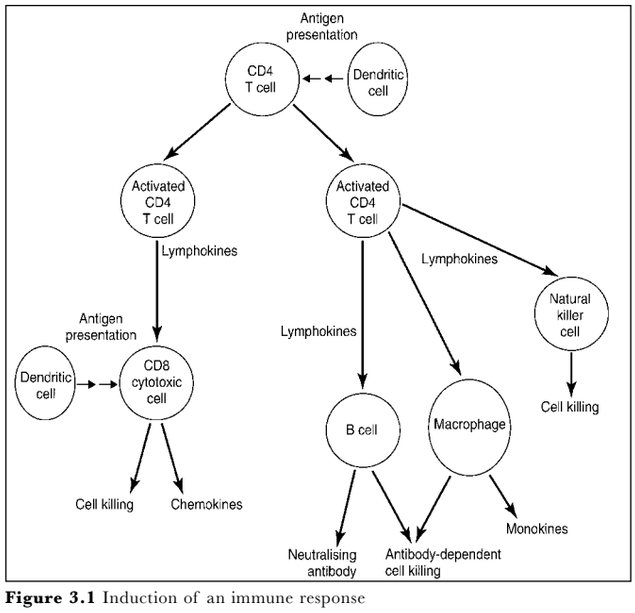


Figure 3.1 Induction of an immune response

but their antigen-presenting function is impaired, with secondary effects on lymphocytes. Monocytes are the precursors to some glial cells and abnormal lymphokine production after HIV infection may have harmful effects on neural tissue and result in HIV encephalopathy.

Early after HIV infection antibody responses are not impaired; indeed, development of antibodies to the virus envelope and core proteins is the principal evidence for HIV infection and persists until death. In adults, massive activation of B lymphocytes is manifested by a rise in serum immunoglobulin concentration, perhaps due to direct activation of B cells by HIV. This polyclonal activation explains why a variety of false positive serological tests are seen in HIV infection. In young children, the reverse pattern may be seen, with extremely low levels of immunoglobulin sometimes requiring intravenous replacement therapy.

Within days or weeks after infection there may be a transient fall in CD4 lymphocyte numbers and a more sustained rise in the number of CD8 cytotoxic/suppressor cells. Among the CD8 cells, expanded oligoclonal populations are frequently seen and as in other acute virus infections, some of these represent a specific response to HIV. Following this acute reaction, healthy seropositive individuals may have normal numbers of lymphocytes, although the numbers of CD8 cells frequently remain high. Even at this stage, however, *in vitro* testing may show a lowered response to previously encountered (recall) antigens (tetanus toxoid or purified protein derivative, for example). This seems to be due to poor production of the lymphokine interleukin 2. Individuals may remain healthy for long periods, but a hallmark of disease progression, often prior to the development of new clinical symptoms, is a fall in the number of CD4 lymphocytes. In AIDS the number of CD8 lymphocytes also falls.

Biopsy of the lymph nodes in patients with persistant generalised lymphadenopathy shows many enlarged follicles, often infiltrated by CD8 lymphocytes, with depletion of CD4 cells. Even in clinically silent HIV infection, lymph nodes are the site of remarkably active HIV replication. Uninfected cells may also die by apoptosis, initiated by unexplained mechanisms. In the later stages lymph nodes return to normal size and follicles become "burnt out", with loss of normal architecture and progressive cellular depletion.

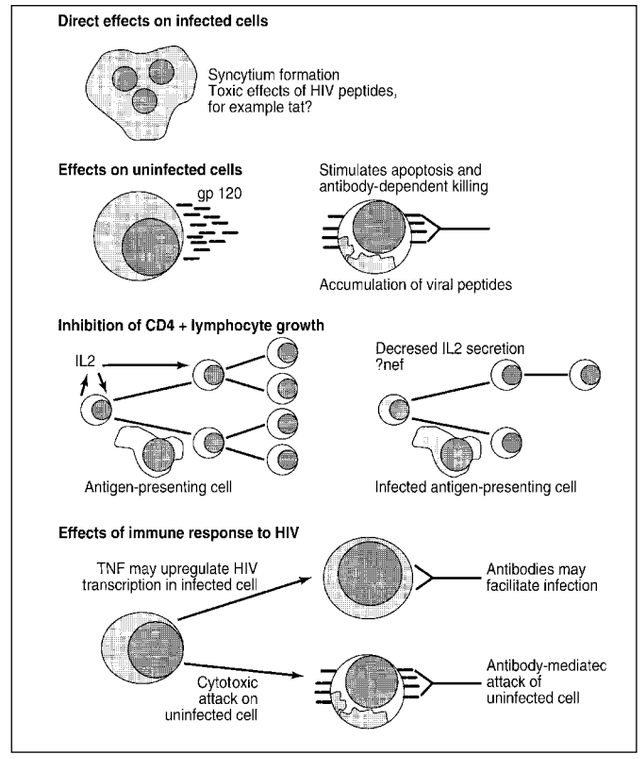
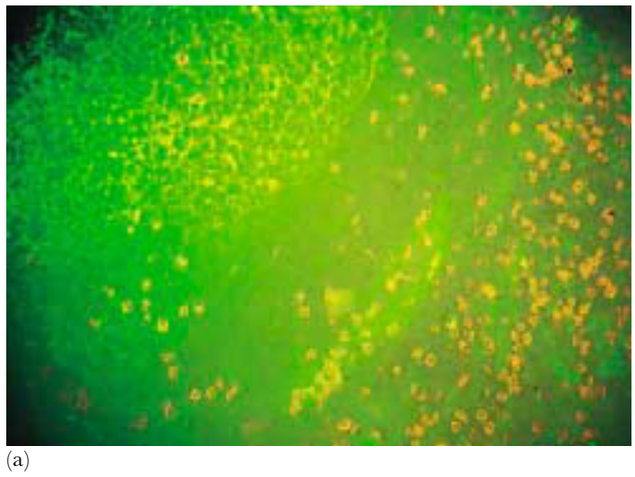
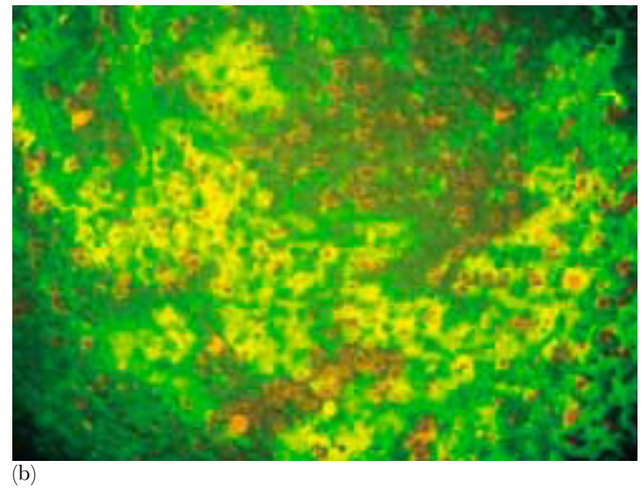


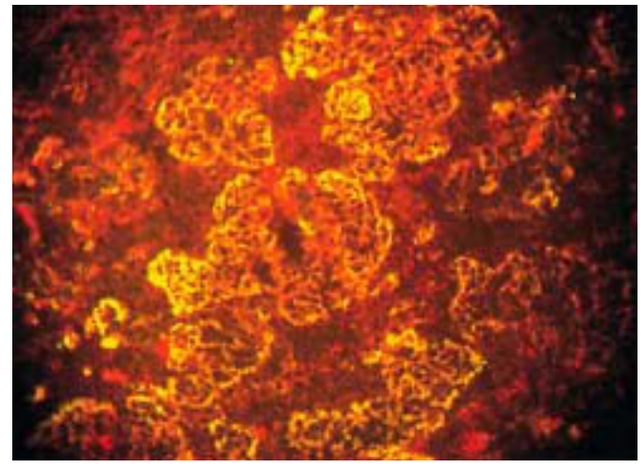
Figure 3.2 Mechanisms of CD4 lymphocyte loss in HIV infection



(a)



(b)



(c)

Figure 3.3 (a): Normal lymph node in which B lymphocytes and follicular dendritic cells (green) form a regular network and suppressor/cytotoxic CD8 T-cells (red) populate the paracortical areas. (b): Node from HIV-positive patient with persistent generalised lymphadenopathy which has been infiltrated by many CD8 cells and in which the regular structure has been destroyed. (c): Same section as middle picture showing complexes of HIV core antigen (orange) and immunoglobulin (red) deposited in germinal centre

Specific immune responses to HIV

In spite of the fact that HIV-infected individuals show the gross abnormalities of immune function described above, they are able to mount a specific immune response to HIV itself. Although serum reactivity to all the viral proteins is detectable, virus neutralising titres are generally low and directed against the immunising virus strain (type specific immunity). Passive transfer of antibody from asymptomatic to symptomatic patients is claimed to be beneficial, but this requires confirmation. Antibodies to HIV may even facilitate infection of cells bearing immunoglobulin (Fc) receptors, such as monocytes. In AIDS a fall in the titre of antibodies to core protein (p24) is often associated with disease progression. p24 antigen, which is detectable in the serum of some patients, may show a rise at the same time and has been used as a marker of disease progression.

CD8 cytotoxic lymphocytes (CTL) capable of killing HIV-infected targets are detected in most HIV-infected individuals and may be beneficial. This is suggested by the observation that viraemia declines at the time that CTL are first detected following infection, and in patients with stable disease, a high frequency of CTL is detectable in the peripheral blood. In addition, in individuals who have been regularly exposed to HIV while remaining seronegative and without detectable virus, HIV-specific CTL have been detected. As well as killing infected cells directly, CD8 lymphocytes may contribute to protection by producing several chemokines and CAF (CD8 T-cell antiviral factor), which strongly inhibit viral replication in CD4 cells. All this has led to the suggestion that CTL are an effective protective mechanism. However, because reverse transcription is an error-prone process, virus mutants arise, which evade the CTL response (escape mutants). These mutants may not only evade recognition themselves but also inhibit recognition of un-mutated virus.

There is some evidence to suggest that a minority of patients mount a specific CD4 T-cell response to HIV and that this is associated with effective control of virus replication. In animal experiments CD4 cells have been shown to be important for the maintenance of an effective CTL response, which may explain this association.

Box 3.1 Positive and negative effects of immune responses

Antibody

Beneficial effects

- Neutralising antibody (demonstrated *in vitro* only) might prevent primary infection and destroy some infectious particles
- Evidence for beneficial effect of passive transfer of antibody in man requires confirmation

Harmful effects

- Antibody may also help the virus to enter cells with Fc receptors
- Immune complexes may cause tissue damage, anemia and neutropenia

Cellular immune responses

Beneficial effects

- A strong CD8 response is correlated with primary resistance in some individuals and with long-term survival
- Cytotoxic T-cells may delay the progress of disease by killing infected cells.
- They produce CD8 T-cell anti-viral factor (CAF) which inhibits viral replication and may be important in slowing disease progression

Harmful effects

- They may kill uninfected cells which take up shed gp120
- Abnormal cytokine secretion may cause immunopathology (perhaps including encephalopathy)

Table 3.2 Protective mechanisms of CD8 T-cells

	Cytotoxic	Non-cytotoxic
Property or mechanism	Death of infected cells	Inhibition of viral replication
Antigen specificity	Specific for epitopes of viral proteins	Non-specific
Cell contact needed?	Yes	No
Mechanism or CAF	Perforin or fas/fas L	CC chemokines
Induction by vaccination?	Yes	Not known

Monitoring HIV infection

Counting CD4 lymphocyte numbers (the "CD4 count") is an important part of monitoring HIV infection. A progressive downward trend in CD4 cells reflects disease progression and decreased life expectancy, even in the absence of symptoms. Epidemiological studies have firmly correlated distinct ranges of CD4 cell counts with risk of particular opportunist infections. Recent data show that monitoring either the absolute CD4 lymphocyte count or the ratio of CD4 to CD8 cells, the 4:8 ratio, are both equally good at monitoring progression in HIV infection. β_2 microglobulin and neopterin are molecules shed from activated lymphocytes; serum levels increase with progressive HIV infection and can be a useful adjunct to CD4 counts in monitoring.

CD4 lymphocyte numbers have a diurnal variation and delays in the sample reaching the immunological laboratory (for example, when a sample is held overnight) also cause profound changes. Because CD4 lymphocyte counting is a lengthy process, most consistent results are obtained when samples are taken at a set time in the morning and sent straight to the lab. In case of unavoidable hold ups, samples should not be refrigerated.

Box 3.2 Causes of CD4 lymphopenia

- HIV infection: seroconversion illness and during disease progression
- Acute viral infections*
- Tuberculosis*
- Sarcoidosis*
- Corticosteroid therapy
- Purine metabolism defects; ADA and PNP deficiency
- SLE

* Reduce CD4 counts when not associated with HIV and can further reduce levels in HIV infection. ADA, Adenosine deaminase; PNP, Putine nucleoside phosphorylase

CD4 counts should never be used as a substitute for an HIV test because low peripheral blood counts are seen in other conditions. The classic examples are sarcoidosis and tuberculosis (without HIV). Used inappropriately in these settings, a CD4 lymphocyte count may incorrectly suggest a diagnosis of HIV infection. CD4 counts may be low during seroconversion illness but usually recover initially during the asymptomatic phase. Hence there is a need to carry out several baseline CD4 counts if subsequent monitoring is to be useful.

Vaccine development

Immunisation against an organism whose target is an important component of the immune system presents particular difficulties. In addition, HIV has already been shown to be perhaps the most variable virus yet discovered, and HIV-2 differs greatly from all HIV-1 isolates. So far, efforts to immunise against the virus have concentrated on the use of cloned gp120 because all strains of virus so far tested use gp120 to bind to the CD4 molecule, implying that a part of the envelope is similar in all strains. In experimental animals gp120 does induce a neutralising antibody response to the virus but restricted to the immunising strain of virus (type specific immunity) and these neutralising sera do not provide reliable protection against virus challenge *in vivo* in animal experiments. More recently it has been shown that gp120 and its anchor gp41 exist in the viral envelope as a trimer of heterodimers. Because of this and because gp120 is heavily glycosylated, much of the antibody response is to the variable V2 and V3 loops. Furthermore, primary isolates have been shown to be less susceptible to neutralisation than the tissue culture-adapted strains, from which the recombinant gp120 used as immunogen in most experiments derives. Thus new immunogens are needed to raise broadly reactive neutralising antibody and a variety of oligomeric and deglycosylated forms of gp120, lacking the V2 and V3 loops, are being tried.

High levels of CTL are seen in the early stages of HIV infection and the demonstration of CTL escape mutants suggests that they play a role in controlling the virus. That individuals exposed to HIV but with no evidence of infection exhibit CTL responses, reinforces the view that this type of response is important in protection. An effective vaccine might therefore contain components able to stimulate both neutralising antibody, CD4 T-cells and strong CTL responses.

A key factor in generating immune responses is the way in which the antigens are presented to the immune system. For the generation of effective CTL responses attenuated live viruses are effective and attenuated (nef deleted) simian immunodeficiency virus (SIV) has been shown able to protect monkeys against challenge with virulent virus. While such a strategy is unlikely to be used in humans because of worries about the safety of such a virus, it suggests that live viral vectors may be an effective means of immunising against HIV. HIV genes have been inserted into several possible vectors (vaccinia, canary pox, adenovirus) and a number of phase 1 trials are in progress. Alternate means of delivery capable of inducing both antibody and cellular immunity, such as peptides or proteins in novel adjuvants, naked DNA, or the use of different methods of antigen administration in sequence (prime/boost regimes) are under active investigation.

Clearly neither antibody- nor cell-mediated responses prevent the progression of disease in most patients, but they may delay it. However, strong pre-existing humoral and cellular immunity induced by a vaccine might still be protective. Results of vaccination experiments in monkeys and the existence of individuals who appear to be resistant to HIV infection, provide grounds for cautious optimism with regard to the feasibility of

Box 3.3 Strategies for vaccine development

- A good vaccine should induce neutralising antibody, helper T-cells and cytotoxic T-cells.
- Since antibodies bind to three dimensional structures, induction of neutralising antibody requires native envelope.
  Problem: Native envelope is trimeric.
- T-cells recognise 8–15 amino acid-long peptides bound to Major Histocompatibility Complex (MHC) class I and II molecules.
  Problem: Antigen needs to enter antigen presenting cells, usually dendritic cells, to be broken down to peptides.
- Peptides with novel adjuvants can generate good T-cell responses.
  Problem: Different peptides bind to each MHC allele so a large cocktail of different peptides may be needed.
- Adjuvants are needed to induce large responses.
  Problem: There are very few adjuvants available for unrestricted use in humans. Alum is mainly good for induction of antibody responses.
- DNA immunisation can generate antibody, helper and cytotoxic responses and allows incorporation of adjuvant molecules into the vaccine.
  Problem: So far DNA vaccination has not proved as effective in man as in experimental animals.
- HIV is very variable and escape variants arise rapidly in infected individuals. Prophylactic immunisation may tip the balance in favour of the host and prevent escape. Some parts of the virus sequence are relatively invariant, these should be targeted if possible.
- In experimental animals immunisation with different immunogens appears promising. DNA vaccination followed by immunisation with antigen in a recombinant viral vector seems particularly effective. This is now under trial in man.

Table 3.3 Immunotherapy for AIDS

Treatment	Outcome
α and γ interferons	Inconclusive
Interleukin-2	Inconclusive
Cyclosporin A	Not beneficial
Anti-HIV antiserum	Possible transient improvement
Bone marrow transplantation	Transient improvement in lymphocyte count and skin anergy
Anti-CD3 or IL-2 after HAART	Under investigation

producing HIV vaccines. Adequate testing of an HIV vaccine will be difficult in man, although the SIV model provides a model for vaccine development.

Possibilities for immunotherapy

Attempts at immune reconstitution have been made using interleukin 2, interferons, thymic factors or bone marrow transplantation. These have not been notably successful and remain potentially harmful, since the very factors which activate T-cells will also activate HIV replication. *In vivo*, activation of CD4 cells is caused by stimulation with antigens in the form of micro-organisms or vaccines. This suggests that it is sensible to treat intercurrent infections promptly and provides a rationale for prophylactic chemotherapy for pneumocystis. In some studies, vaccination (for example with influenza vaccine) has been shown to be enough of an antigenic stimulus to increase HIV replication.

The advent of highly activated antiretroviral therapy (HAART) has enabled the viral load to be enormously reduced, but the difficulty of maintaining this type of therapy over long periods has led to a search for strategies to complement drug treatment. Two observations are pertinent, the first is that even after 2–3 years of HAART treatment, latent virus can still be detected and the second is that antiviral immune responses decline during treatment. It has therefore been suggested first that latent virus should be "flushed out" by activation of the immune system with anti-CD3 antibody or interleukin 2 while still continuing drug treatment. Secondly vaccination against HIV should be instituted to prevent recrudescence of low level infection. Both strategies are being actively investigated.

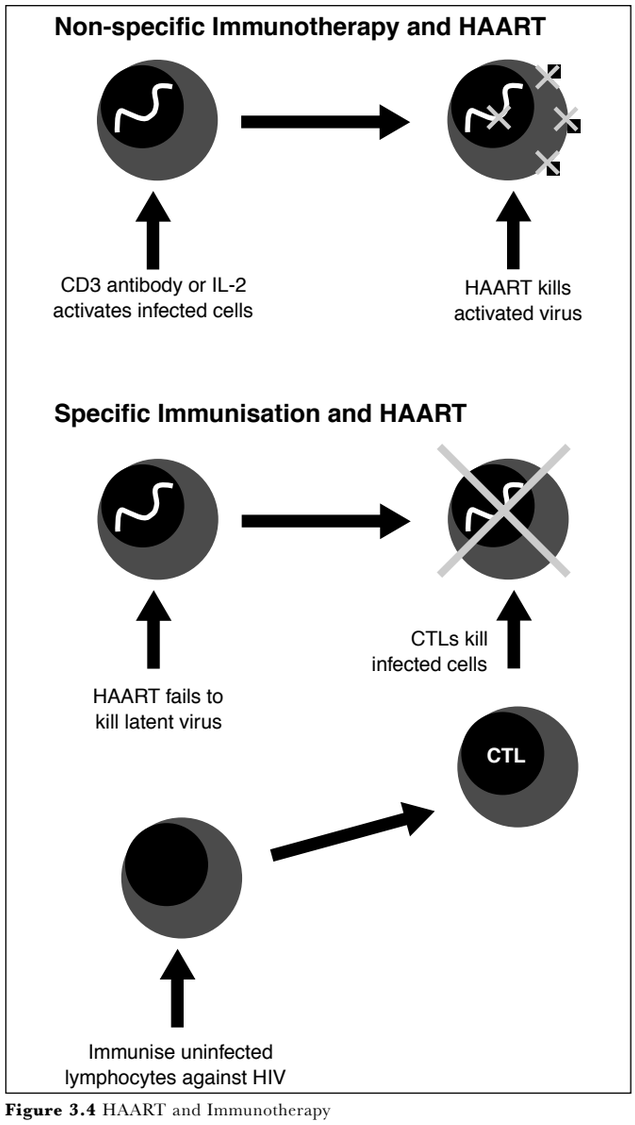


Figure 3.4 HAART and Immunotherapy

4 Natural history and management of early HIV infection

Adrian Mindel, Melinda Tenant-Flowers

Introduction

Infection with HIV causes a spectrum of clinical problems beginning at the time of seroconversion (primary HIV) and terminating with AIDS and death. It is now recognised that it may take 10 years or more for AIDS to develop after seroconversion. The Centers for Disease Control (CDC) in the USA developed the most widely used classification for HIV disease based on the presence of clinical symptoms and signs, the presence of certain conditions and investigative findings, the availability of HIV screening and the degree of immunosuppression as measured by the CD4 lymphocyte count. The infection is divided into four groups (Box 4.1):

Group I Primary HIV infection
Group II Asymptomatic phase
Group III Persistent generalised lymphadenopathy
Group IV Symptomatic infection

Group IV is subdivided into several subgroups and some of these (groups IVA, B, C1 and D) are AIDS-defining conditions (Box 4.1).

In 1993 the CDC included all HIV-infected persons with CD4 lymphocyte counts of <200 cells/mm^3 as fulfilling an AIDS defining diagnosis. However, this additional classification is not widely used outside the USA.

A second classification also combines clinical and CD4 count information. Symptoms and clinical findings are graded in severity from A to C_0 and CD4 counts as they fall from 1 to 3 (Table 4.1).

Group I Primary HIV infection

Primary HIV infection (PHI) is also called the seroconversion illness or acute HIV infection. It represents the stage of infection after the acquisition of the virus when antibodies are developing as shown in Figure 4.1. Between 25% and 65% of people have been found to present with symptoms at the time of seroconversion. These can range from a mild, glandular fever-like illness to an encephalopathy. Common symptoms and signs are shown in Box 4.2. The severe symptoms are rare. The differential diagnosis of the mild seroconversion illness is protean and, without a high index of suspicion and a history indicating relevant risk behaviours or factors, the diagnosis may be missed. Investigations that may be useful in reaching a diagnosis are set out in Table 4.2.

The appropriate diagnostic tests for PHI, which should be carried out on serial blood samples, include tests for HIV antibodies and antigen. If these are negative and PHI is suspected, the definitive test is an HIV RNA PCR, which is the most sensitive test for the detection and quantification of the virus. Some of these assays are not routine and the interpretation of investigation results during PHI is difficult, therefore close consultation with colleagues in virology is strongly advised.

At the time of PHI there is sometimes a high rate of viral replication, leading to a transient rise in HIV viral load and concomitant immunosuppression due to a short-lived fall in the CD4 count. This may result in manifestations of HIV disease

Box 4.1 Summary of CDC 1992 classification system for HIV disease

Group I	Primary HIV
Group II	Asymptomatic infection
Group III	Persistent generalised lymphadenopathy
Group IV	Symptomatic infection
Group IVA	HIV wasting syndrome (AIDS) and constitutional disease
Group IVB	HIV encephalopathy (AIDS) and neurological disease
Group IVC1	Major opportunistic infections specified as AIDS-defining
Group IVC2	Minor opportunistic infections
Group IVD	Cancers specified as AIDS-defining
Group IVE	Other conditions

Table 4.1 Summary of CDC 1993 classification system for HIV disease

	CD4 lymphocyte count $\times 10^6$/l		
	(1)	(2)	(3)
	>500	200–499	<199
(A) Asymptomatic including Groups I, II and III	A1	A2	A3
(B) Symptomatic not A or C	B1	B2	B3
(C) AIDS-defining conditions	C1	C2	C3

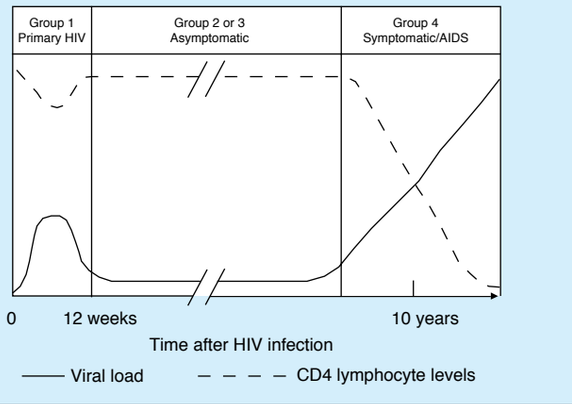


Figure 4.1 Association between virological, immunological and clinical events and time course of HIV infection

Box 4.2 Clinical manifestations of primary HIV infection

- Glandular fever-like illness
- Fever, malaise, diarrhoea, neuralgia
- Arthralgia, sore throat, headaches
- Lymphadenopathy
- Macular papular rash
- Ulceration
  - Oropharynx
  - Anogenital area
- Neurological symptoms
  - Meningitis
  - Neuropathy
  - Myelopathy
  - Encephalopathy

which are normally seen later in the infection, for example oral candida. Diagnostic confusion as to the stage of HIV infection may arise, which can only be resolved by following up the patient for long enough to see the symptoms and signs resolve, HIV antibodies appear, the viral load fall and the CD4 count rise. Treatment should be directed at alleviating any symptoms, and there is considerable interest in the possible use of antiretroviral agents at this time because the virus may be more susceptible due to the relatively low numbers of virus particles which can replicate, the reduced ability of the predominantly non-syncytium-inducing strains of virus to infect a wide variety of cell types and the enhanced immune response seen in PHI.

Such treatment may decrease long-term damage to the immune system and delay or even prevent the development of AIDS. However, if not started within 12–18 months of PHI the theoretical advantage may be lost and, in any case, has to be balanced against the uncertain outcome, drug toxicity, adherence difficulties and the possibility of developing resistant virus, limiting future treatment options.

Group II Asymptomatic infection

After PHI, HIV antibodies continue to be detectable in the blood. The amount of virus in blood and lymphoid tissues falls to very low levels and the rate of HIV replication is slow although it does not cease. CD4 lymphocyte counts are within normal limits or generally above 350 cells/mm^3. This phase may persist for 10 years or more (Figure 4.1). The role of antiretroviral therapy during asymptomatic infections is discussed in chapter 9. The decision to treat is made on the basis of the CD4 count and the viral load. The aim of therapy is to maintain immune function by suppressing viral replication to prevent further damage to the immune system. As for PHI treatment, the potential gain of therapy must be weighed against the potential risks and uncertainties.

Group III Persistent generalised lymphadenopathy

Persistent generalised lymphadenopathy may be a presenting feature of HIV infection in a person who is otherwise well. HIV-related lymphadenopathy persists for at least three months, in at least two extra-inguinal sites and is not due to any other cause. The differential diagnosis of this lymphadenopathy is shown in Table 4.3.

A lymph node biopsy in HIV disease is not recommended as a routine procedure as the findings are non-specific and the presence of lymphadenopathy due to HIV alone does not worsen the prognosis. The indications for a biopsy are the same in HIV and non-HIV-related conditions (Box 4.3).

Group IV Symptomatic HIV infection before the development of AIDS

The progression of HIV infection is a result of a decline in immune competence that occurs due to increased replication of HIV from sites where it has been latent. The exact triggers for this reactivation are poorly understood. As the disease progresses, infected persons may suffer from constitutional symptoms, skin and mouth problems and haematological disorders, many of which are easy to treat or alleviate. A decrease in viral load in response to the introduction of antiretroviral therapy often corresponds to a complete or partial resolution of these symptoms.

Table 4.2 Differential diagnosis of glandular fever-like illness

Condition	Test
Viral	
Infectious mononucleosis	Paul-Bunnell
Cytomegalovirus	Serology/culture
Rubella	Serology
Herpes simplex	HSV culture
Adenovirus	Serology
Hepatitis B/C	Serology
HIV	HIV, Ab, Ag, PCR
Protozoal	
Toxoplasmosis	Serology
Bacterial	
Syphilis	Serology
Streptococcal pharyngitis	Bacterial culture
Brucellosis	Serology
Neoplastic	
Lymphoma or leukaemia	Full blood count/diff Lymph node biopsy Bone marrow

Table 4.3 Common causes of generalised lymphadenopathy

Condition	Test
Infections	
Bacterial	
Syphilis	Serological tests (Venereal Diseases Research Laboratory), *Treponema pallidum* haemagglutination and Fluorescent Antibody tests
Brucellosis	Serological tests
Viral	
Infectious mononucleosis (Epstein–Barr virus)	Paul–Bunnell
Cytomegalovirus	CMV cultures or antibodies
Hepatitis A	Serology
Hepatitis B	Serology
Rubella	Serology
Parasites	
Toxoplasmosis	Toxoplasma serology
Tumours	
Lymphomas, leukaemia's or other tumours	Full blood count, lymph node biopsy, CT or MRI scans etc.
Miscellaneous	
Sarcoidosis	Clinical features, Kviem test

Box 4.3 Indications for lymph node biopsy

- Constitutional symptoms
- Painful nodes
- Asymmetrical enlargement
- Sudden increase in size
- Hilar lymphadenopathy

Constitutional symptoms

Common constitutional symptoms associated with Group IVA HIV infection include malaise, fevers, night sweats, weight loss and diarrhoea. Serious constitutional symptoms are set out in Box 4.4. The exact criteria for diagnosing the AIDS-defining HIV wasting syndrome are, the combination of 10% weight loss from baseline and one of the other serious symptoms set out in Box 4.4. Many patients find these symptoms worrying and debilitating and they should be investigated to diagnose treatable causes other than HIV. Once other causes have been excluded, symptomatic treatment can include antipyretics, antidiarrhoeal agents and, if all else fails, steroids.

Box 4.4 Constitutional symptoms in HIV infection

- Weight loss >10% baseline
- Fever lasting at least 1 month
- Diarrhoea lasting at least 1 month

Skin and mouth problems

Many skin problems occur in patients with HIV infection (Box 4.5). These may represent exacerbations of previous skin disease, or a new problem. Identical skin conditions occur in HIV-negative persons. However, in the immunocompromised, these common conditions may be more severe, persistent and difficult to treat. Many minor opportunistic infections (Group IVC2) manifest themselves on the skin and in the mouth. Seborrhoeic dermatitis is frequently seen and usually presents as a red scaly rash affecting the face, scalp and sometimes the whole body. This condition often responds well to 1% hydrocortisone and antifungal cream.

Box 4.5 Skin and mouth problems associated with HIV

Skin problems

Miscellaneous
Seborrhoeic dermatitis
Fungal
Tinea
  Cruris
  Pedis
  Other
Candida
  Genital
  Perianal
  Other
Pityriasis versicolor
Bacterial
Staphylococcal infection (impetigo)
Acneform folliculitis
Viral
Herpes simplex (types 1 and 2)
  Oral
  Genital
  Perianal
  Other
Varicella zoster
Human papilloma virus
Molluscum contagiosum
Neoplastic
Cervical dysplasia

Mouth problems

Hairy oral leukoplakia
Dental abscesses/caries
Gingivitis
Candidiasis
Ulceration
  Bacterial
  Herpetic
  Aphthous

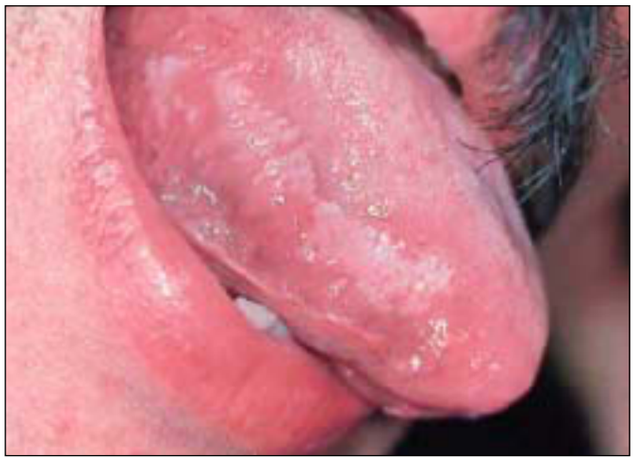


Figure 4.2 Hairy leukoplakia

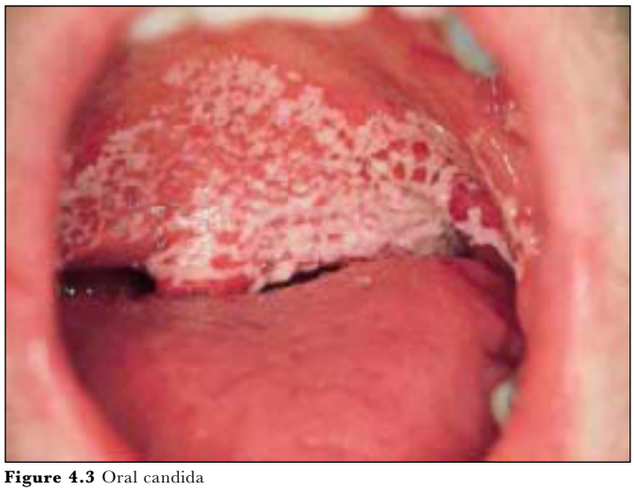


Figure 4.3 Oral candida

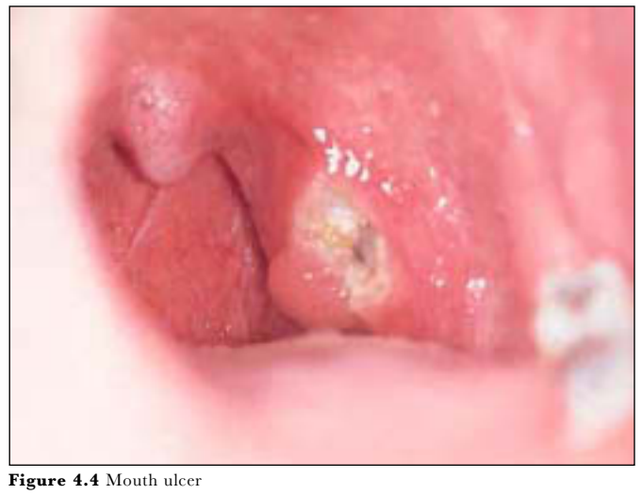


Figure 4.4 Mouth ulcer

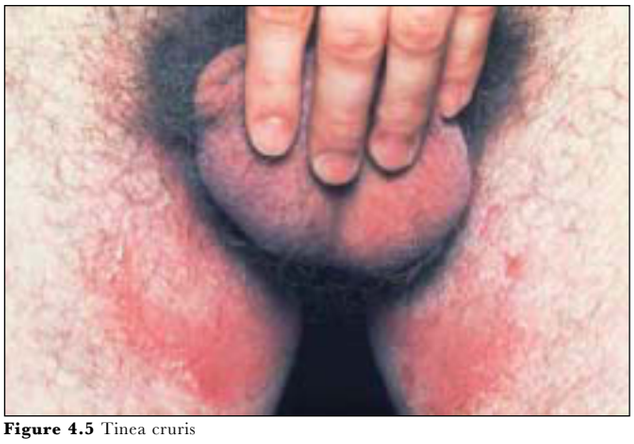


Figure 4.5 Tinea cruris

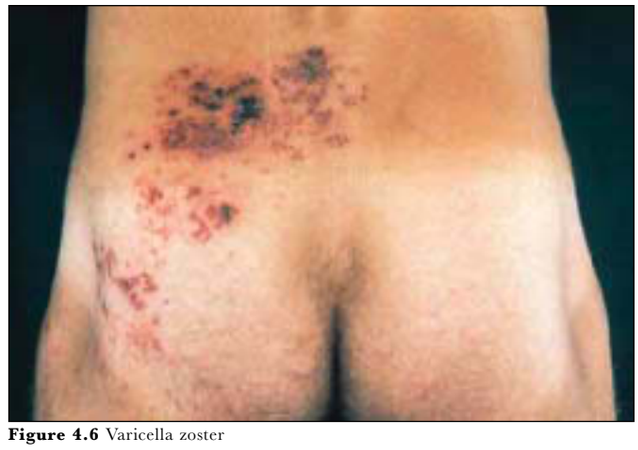


Figure 4.6 Varicella zoster

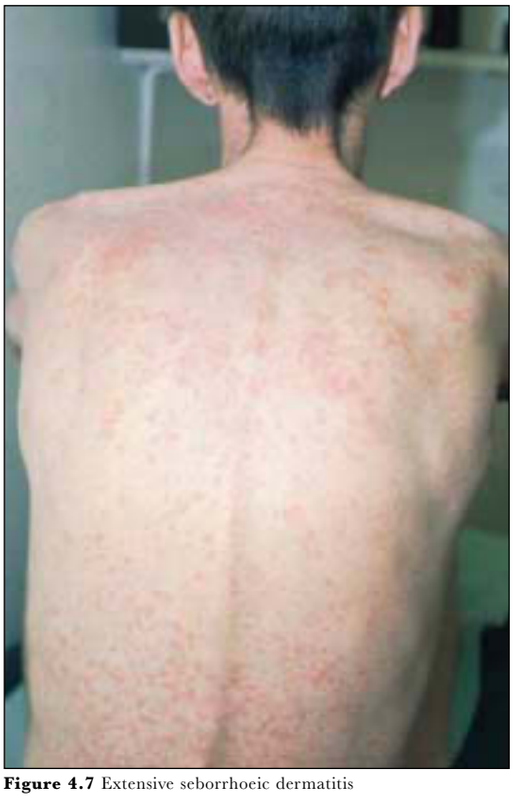


Figure 4.7 Extensive seborrhoeic dermatitis

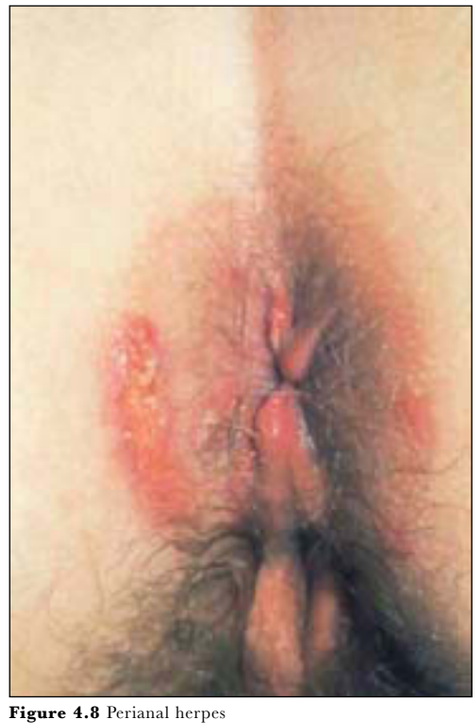


Figure 4.8 Perianal herpes

Other common dermatoses that respond to antifungal creams (for example Clotrimazole) include tinea cruris and pedis and candidiasis. Folliculitis often responds to 1% hydrocortisone and antifungal cream, impetigo to antibiotics and shingles to aciclovir, valaciclovir or famciclovir. Recurrent perianal or genital herpes may become more troublesome, with recurrences lasting longer and occurring more frequently; if this persists for more than 3 months it is considered an AIDS-defining opportunistic infection (Group IVC1). Treatment with long-term acyclovir, valaciclovir or famciclovir suppression is often required. Genital and perianal warts are common, difficult to treat and frequently recurrent, and high-grade cervical dysplasia is seen more often in HIV-infected women.

Mouth problems are also common, cause considerable distress and when severe may result in difficulty with eating and drinking. Oral candida can be managed with topical or systemic antifungals (eg, nystatin, ketoconazole or fluconazole). If dysphagia develops, oesophageal candidiasis should be suspected and investigated. Oral hairy leukoplakia can be differentiated from oral candida by its characteristic distribution along the lateral borders of the tongue and the fact that it cannot be scraped off. Although unsightly, this condition which is due to Epstein–Barr virus reactivation is painless and temporary remission can be obtained with acyclovir, valaciclovir or famciclovir. Other oral conditions including dental abscesses, caries, gingivitis and oral ulceration (herpetic or bacterial) may occur. Mouth ulcers may be particularly difficult to treat and expert specialist assessment is recommended. Metronidazole, acyclovir, 0.2% chlorhexidine mouthwashes and analgesic sprays may all be effective depending on the cause and, in extreme cases, thalidomide has been used. Maintenance of good oral hygiene and dental care are important.

HIV and haematological problems

Lymphopenia with depression of the CD4 cell subset is a marker for HIV disease. Mild to moderate neutropenia and a

normochromic, normocytic anaemia of unknown origin are often seen but usually have no adverse effect on HIV-infected individuals. Severe anaemia or neutropenia should be investigated for other underlying causes. Thrombocytopenia is common in HIV disease and, only if persistent, causing bleeding and less than 20×10^9/litre warrants treatment with antiretrovirals which is usually effective. Many therapies used to treat HIV may be toxic to bone marrow.

Risk of progression and the value of surrogate markers

One of the hardest problems confronting the physician dealing with an asymptomatic patient with HIV infection is predicting how soon that patient will progress to symptomatic disease or AIDS. This issue is important, firstly in terms of counselling and secondly, to decide which patients may benefit from antiretroviral treatment or prophylaxis to prevent opportunistic infections.

Variables associated with rapid disease progression include a symptomatic PHI, older age at diagnosis and receiving a large inoculum of virus, for example via a contaminated transfusion from a donor with a high viral load. The effect of prophylaxis against opportunistic infections (for example cotrimoxazole for pneumocystis and toxoplasmosis) has been to delay the onset of AIDS and to change the pattern of disease represented by the first AIDS-defining illness. Antiretroviral treatment has independently been shown to increase survival before and after AIDS. Some infected individuals do not progress for many years and work is in progress to determine whether this is due to their genetic makeup, amount of viral inoculum, characteristics of the infective virus or their immune system.

Many laboratory indices have been used as prognostic indicators, both to evaluate disease progression and treatment efficacy. The most widely used are the CD4 absolute lymphocyte count or percentage and the viral load. At least two CD4 measurements should be obtained before initiating prophylaxis for opportunistic infections or antiretroviral therapy, as the CD4 count is subject to diurnal and seasonal variation and reduced by intercurrent infection. A fall in CD4 cells is associated with disease progression, particularly if the rate of decline is rapid. Likewise, at least two viral loads, from the same laboratory using the same assay, should be obtained to avoid interassay variation. Some HIV clades are more difficult to monitor with certain assays and the laboratory should be informed of the country of origin of the patient.

Patients who may need close monitoring include individuals whose CD4 count falls below 350 cells/mm^3, those with a rapidly declining CD4 count, those with a rising viral load and patients who are symptomatic as they may all be candidates for antiretroviral therapy. Patients who present with persistent constitutional symptoms, mouth or skin problems should be considered for antiretroviral therapy irrespective of CD4 count and viral load. These issues are discussed further in the chapters on treatment of infections and antiretroviral agents.

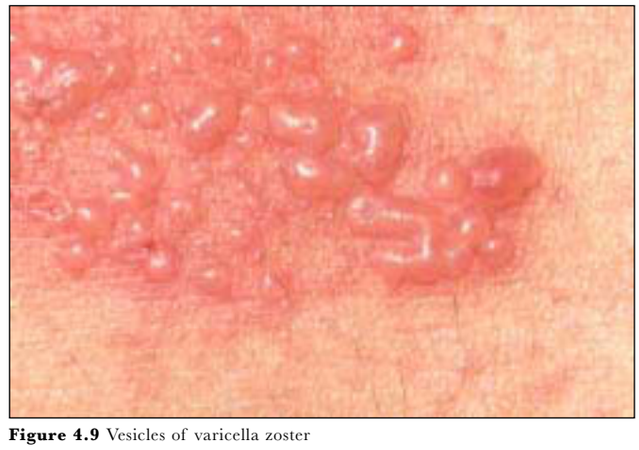


Figure 4.9 Vesicles of varicella zoster

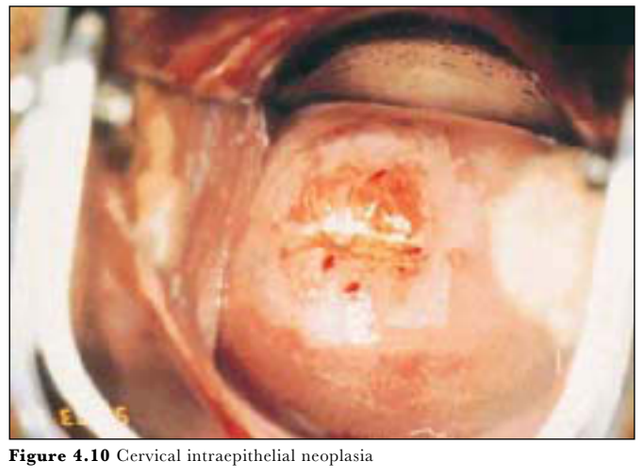


Figure 4.10 Cervical intraepithelial neoplasia

General management of HIV-infected people

One of the most important aspects of dealing with any HIV-infected person is confidentiality (Box 4.6). Maintaining confidentiality might be complicated: for example the patient's family or friends may not know his or her diagnosis or sexual orientation; people at work (or school) may seek medical information (especially if the individual is having time off

Box 4.6 General management of the HIV-infected person

- Protect confidentiality
- Medical issues
- Psychological support (patient, family and friends)
- Avoidance of transmission
- Other issues (dental treatment, insurance, work, or school, etc.)

work); or the person may fear that information may inadvertently be given to third parties. Special precautions may be required, firstly to reassure the patient that confidentiality is protected and, secondly, to limit any unwarranted dissemination of confidential information. Issues related to partner notification are discussed in chapter 13.

The routine medical management of these individuals is usually straightforward. They should be seen regularly, for example every three to six months. At each visit the patient's weight should be recorded and special attention given to mouth or skin problems and, if necessary, they should be referred to the appropriate specialist. Screening for STDs and hepatitis viruses should be offered if the individual is at risk and hepatitis A and B vaccines can be safely given. Repeating a full blood count and measuring the CD4 count and viral load every three to six months allows early detection of actual or imminent immune dysfunction. Patients should be advised to reattend if they develop any symptom, especially those suggestive of opportunistic infections or cancers, for example shortness of breath, cough, haemoptysis, pain or difficulty in swallowing, diarrhoea, weight loss, fevers, headaches, fitting, altered consciousness or purple spots on their skin. Other symptoms may indicate increased viral replication and the need to consider treatment.

Psychological and emotional support of the infected individual, the family and friends are a vital aspect of management (see chapter 13). HIV antibody positive persons should also be advised about reducing the risk of transmitting HIV to others and reducing their own risk of receiving different, possibly drug resistant, strains of HIV. Advice concerning safer sex, safer needle use, pregnancy, breastfeeding and children should also be provided (see chapter 16). Patients should be advised to tell their dentists about their infection, and it may sometimes be necessary to refer them to a dental unit with an interest in HIV-related problems.

The physician may also be asked to advise about insurance, work, immigration, travel passes, housing and disability benefit. Patients should be referred to the relevant legal or benefit agency as soon as possible. Infected individuals will often have considerable difficulty in obtaining life insurance as most insurance companies ask specific questions about the infection and either refuse insurance or charge very high premiums. Finally, patients should be told that being positive is no barrier to employment provided there is no chance of their body fluids entering another person or of them transmitting an opportunistic infection, such as tuberculosis, by coughing. It is worth noting that for notifiable diseases such as TB, standard, confidential public health notification procedures still apply. Because of widespread misconceptions about infectivity which are still prevalent, information about the individual's HIV status should never be divulged to employers without their written consent.

5 Tumours in HIV

Caroline H Bridgewater, Margaret F Spittle

The United States Center for Disease Control recognises three malignancies as AIDS-defining conditions. These are Kaposi's sarcoma, intermediate or high grade B-cell non-Hodgkin's lymphoma (NHL) and cervical carcinoma. Primary central nervous system lymphoma is a rare B-cell NHL that is often considered separately from the other NHLs. Other malignancies are known to have an increased incidence in HIV whilst not being AIDS-defining, for example Hodgkin's disease. All malignancies are more aggressive in HIV positive patients than in the general population and usually present at advanced stages.

The investigation and treatment of suspected malignancy is complicated by unusual presentations and sites of disease, concomitant infections and immunosuppression. Malignancies may occur at different points in the disease process for different individuals and management must be tailored to the patient's overall maximum benefit.

There are many new developments in the understanding of the pathogenesis of AIDS-related malignancies and in the future these will inform new therapies. In the last few years alone highly active antiretroviral therapy (HAART) has had a great impact on the incidence and natural history of some of these malignancies. As opportunistic infections are more easily treated and patients live longer the malignancies are likely to become relatively more common. The incidence of AIDS-related malignancies varies within the different population groups with HIV and as affected groups evolve there will doubtless be a change in the incidence of malignancies seen in the UK. With these rapid changes optimal treatment strategies are controversial and patients should be entered into clinical trials.

Table 5.1 Risk of malignancies in HIV-positive patients

Malignancy	Relative risk compared to HIV-negative population	Viral co-factor
Kaposi's sarcoma	716–972	KSHV (HHV8)
NHL	71–141	EBV
Primary CNS lymphoma (PCNSL)	~100	EBV
Cervical cancer	?	HPV
Hodgkin's disease	5–9	EBV
Anal cancer	3.5–5	HPV
Testicular germ cell tumours	3	?

Box 5.1 Clinical groups of patients with Kaposi's sarcoma (KS)

- "Classic" KS: elderly, predominantly male, Jewish or Eastern European
- "Endemic" or African KS (various types)
- Immunosuppression-related KS (patients with transplants)
- "Epidemic" or AIDS-related KS

Kaposi's sarcoma

Among the first reported illnesses amongst homosexual men in the USA in 1981 was Kaposi's sarcoma (KS), with 20–40% of HIV-infected homosexual men suffering KS. Hitherto KS had been known in three forms. In elderly Jewish or Eastern European patients as "classic" KS, in sub-Saharan Africa as "endemic" KS, and more recently in transplant and other immunosuppressed patients. KS currently remains the most frequent neoplastic condition in AIDS.

Aetiology

The uneven geographical distribution of KS had long suggested that environmental factors were aetiologically important. Epidemiological observations that KS initially occurred in clusters in the HIV population and that it was 20 times more likely in homosexual men than other risk groups suggested a sexually transmitted cofactor. Work in the biological and statistical fields has gone on to establish causality. Whilst no biological pathway has yet been identified, there is now sufficient evidence to state that a DNA virus, Kaposi's sarcoma-associated herpes virus (KSHV) also known as human herpes virus 8 (HHV8), is an essential, although not necessarily a sufficient, cause of KS.

This evidence has primarily come from longitudinal studies showing that KSHV infection precedes KS. This is consistent with analogous evidence of other herpes viruses, for example Epstein–Barr virus (EBV) being oncogenic. KSHV has been

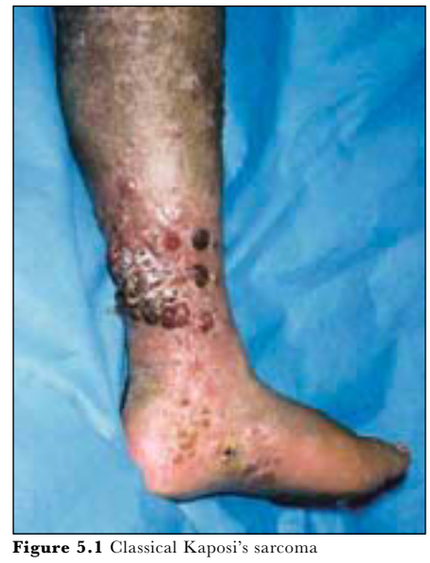


Figure 5.1 Classical Kaposi's sarcoma

detected in tissue biopsies taken from patients with African and classical KS as well as AIDS-related cases.

KSHV can be sexually transmitted and seroconversion has been noted following renal transplantation. In endemic areas non-sexual horizontal and vertical spread are the proposed dominant modes of transmission. The evidence for this is the age-dependent increase in KSHV seroprevalence in prepubescent children in studies from Gambia and Uganda and the greater (29% compared to 0%) seropositivity rate in children born to KSHV seropositive women in South Africa.

Histopathology

The tumours have a characteristic appearance, consisting of groups of spindle cells separated by slits giving a sieve pattern. These spindle cells derive from primitive mesenchymal cells. Red cells are often seen in the slits and early lesions may consist almost entirely of bizarre endothelium-lined vascular spaces in the dermis with few spindle cells.

The tumour stains positive for factor VIII and smooth muscle-specific α-actin on immunocytochemistry staining.

The following cytokines have been shown to promote the growth of KS cells *in vitro*: interleukin 6 (IL-6), tumour necrosis factor (TNF), basic fibroblast growth factor (bFGF), vascular endothelial growth factor (VEGF), platelet-derived growth factor (PDGF), Oncostatin M and granulocyte colony-stimulating factor (GCSF). It may be possible to exploit this therapeutically by inhibiting these cytokines.

Kaposi's sarcoma is a multifocal process rather than a metastatic one.

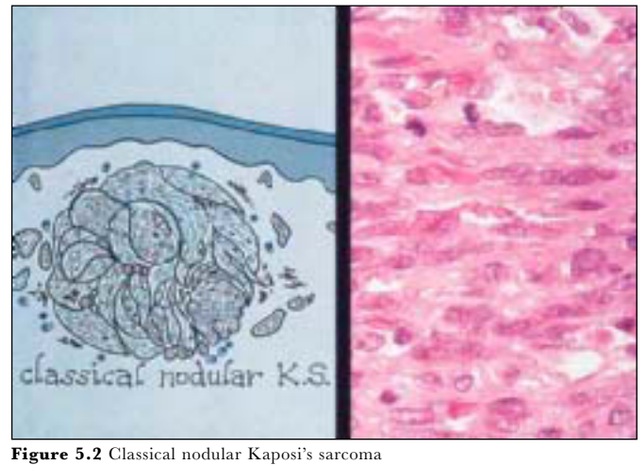


Figure 5.2 Classical nodular Kaposi's sarcoma

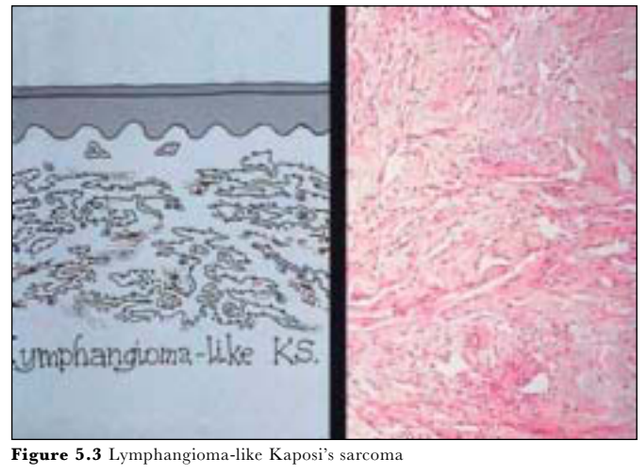


Figure 5.3 Lymphangioma-like Kaposi's sarcoma

Clinical presentation

The classic form tends to follow a very indolent course producing large ulcerated plaques on the lower legs. It shows a strong male preponderance and as most affected individuals are elderly their KS causes significant morbidity but not mortality.

Endemic KS follows a more aggressive course in younger adults with more florid skin lesions and lymph node involvement. Death occurs due to widespread systemic involvement. In young children the lymphadenopathic variant is most commonly seen.

In the non-HIV immunosuppressed patient the lesions of KS may improve with reduction or cessation of the immunosuppression.

In the UK these forms are all rare and most cases of KS are AIDS related.

The presentation of KS in AIDS is variable but the disease tends to become increasingly aggressive and may be lethal. Mucocutaneous lesions begin as flat dusky red papules progressing over weeks or months to vary from a few scattered nodular lesions to large plaques. The legs, trunk, arms, face, hard palate and penis are common sites with associated "woody" oedema and ulceration predominantly affecting the lower limbs. KS on the feet make walking difficult and painful. Other mucocutaneous lesions often cause distress because of their disfiguring appearance.

All organs other than the central nervous system may be affected and the presence of visceral disease is predicted by mucocutaneous disease; one third of respiratory "episodes" in patients with cutaneous KS are due to pulmonary KS. The most common visceral lesions are pulmonary and gastrointestinal. Lymph node disease is also common and may cause venous compression resulting in gross peripheral oedema. Presentation of pulmonary KS is usually with exertional dyspnoea but may be with cough or haemoptysis. Chest radiograph changes are often non-specific with interstitial infiltrates, pleural effusions and mediastinal lymphadenopathy. Further information is

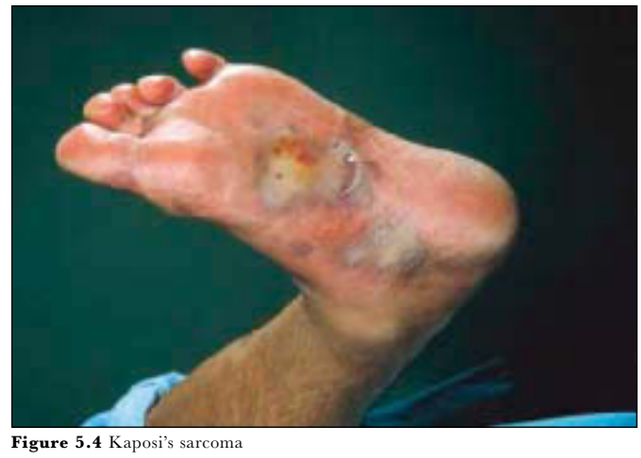


Figure 5.4 Kaposi's sarcoma

gained by bronchoscopy and CT; if possible bronchial biopsy should be avoided as bleeding may be heavy (see chapter 6).

Lesions may occur along the length of the gastrointestinal tract from the palate to the anus and diagnosis is by endoscopy. KS in the oral cavity and oesophagus may cause pain but is usually asymptomatic. Bleeding may occur from lesions throughout the gastrointestinal tract and patients may also suffer protein-losing enteropathy and diarrhoea.

KSHV is also found in two rarer malignancies, primary effusional lymphoma (a subset of B-cell non-Hodgkin's lymphoma) and multicentric Castleman's disease, a lymphoid malignancy which also has an increased incidence in AIDS.

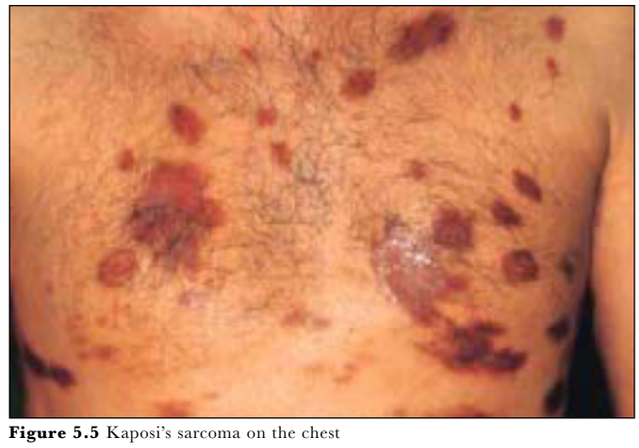


Figure 5.5 Kaposi's sarcoma on the chest

Incidence

KS was the AIDS-defining diagnosis in 30% of patients in the 1980s but this has now fallen. The reduction may be attributed to changes in sexual practices as well as the advent of antiretroviral therapy. KS commonly precedes opportunistic infections and with improvements in treating such infections KS is increasingly common as the cause of death for AIDS patients. Hence whilst the incidence of KS is falling the prevalence is increasing. The deaths of almost 30% of AIDS sufferers are now accounted for by visceral and particularly pulmonary KS.

Treatment

Treatment must be tailored to the site and extent of KS and to the patient's underlying clinical condition. The aim of treatment is resolution of symptoms and prolongation of life. Cure is currently impossible due to the disseminated nature of the condition and its poorly understood pathogenesis as well as the underlying AIDS. HAART has reduced the need for second-line therapies by increasing median time to treatment failure as well as reducing the incidence of KS. There are reports of KS regression with HAART and no other treatment.

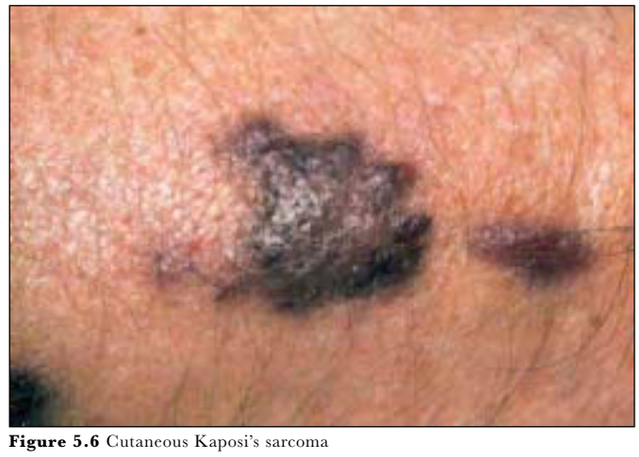


Figure 5.6 Cutaneous Kaposi's sarcoma

Local treatment is important for cosmesis of cutaneous lesions. Superficial radiotherapy is given using 100 kV X-rays applied directly to the skin or palate. A dose of 8 Gy in a single fraction achieves good palliation in 70% of lesions, particularly in early KS with little haemosiderin staining. The area of the lesion is treated with a margin using a lead cutout to protect surrounding tissues. The dose should be given in divided doses on consecutive days (fractionation) to sensitive areas such as the soles or face. Radiotherapy can be repeated if further regression is required or relapse occurs. Alternative treatments include camouflaging with cosmetics and intralesional injection with vinblastine or interferon.

Palatal, bronchial and oesophageal KS can also be treated with radiotherapy. It is particularly useful to stop bleeding.

For extensive mucocutaneous disease or visceral involvement chemotherapy is the preferred option. Several regimens are available and choice of regimen depends on coexistent pathologies and, in some countries, availability and price. Sadly many of the newer treatments with better side-effect profiles are prohibitively expensive for the developing nations where KS is prevalent. In patients with relatively well preserved immune function interferon-α is a useful treatment.

Bleomycin and vincristine in combination was initially the commonest regimen giving a response rate of 50–60% with acceptable side-effects. This has now largely been superseded by the liposomal preparations of doxorubicin and daunorubicin (anthracycline antibiotics) following trials which showed comparable efficacy and reduced toxicity. A liposome is a sphere made of phospholipid bilayers which can be selectively distributed to tumours allowing local drug deposition when the liposome breaks down. Liposomal packaging allows higher doses of these drugs to be delivered with fewer side-effects.

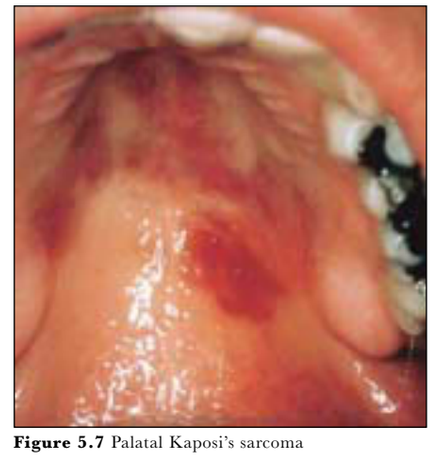


Figure 5.7 Palatal Kaposi's sarcoma

Cumulative doses of non-liposomal anthracyclines are limited at 450–550 mg/m^2 by cardiotoxicity and in a chronic condition like KS long or repeated courses of chemotherapy may be required.

More recently there has been increased interest in the use of antiangiogenics and trials are currently underway on the tyrosine kinase receptor inhibitor SU5416 and thalidomide. This has been fired by studies showing the presence of vascular endothelial growth factor and basic fibroblast growth factor in KS tissues. Paclitaxel, a newer cytotoxic, is also undergoing trials. Common side-effects are those shared with many other cytotoxics including nausea, vomiting, myelosuppression and mucositis. In the AIDS patient, with concomitant diseases and multidrug therapy including HAART, it can be difficult to find the root cause of such symptoms.

Box 5.2 Summary of malignancies

AIDS-defining malignancies
- Kaposi's sarcoma
- High/intermediate grade non-Hodgkin's lymphoma including primary CNS lymphoma
- Cervical carcinoma

Other malignancies with increased incidence
- Hodgkin's disease
- Ano-genital squamous cell carcinoma
- Testicular germ cell tumours

Non-Hodgkin's Lymphoma

The occurrence of non-Hodgkin's lymphoma (NHL) was known to complicate immunodeficiency states before the advent of HIV. Up to 20% of HIV positive people may ultimately develop NHLs and it is the presenting diagnosis in 3% of patients. In immunodeficiency NHLs are commonly extranodal.

Aetiology

Both HIV itself and its related opportunistic infections may cause polyclonal B-cell expansion which is probably cytokine and antigen driven. Patients with AIDS have impaired immunity to EBV when compared to HIV negative EBV-infected individuals and EBV is itself likely to cause polyclonal B-cell proliferation. AIDS lymphomas have modified immunoglobulin variable regions which are consistent with antigen drive as an important factor in lymphomagenesis. Macrophages, acting as antigen-presenting cells, also appear to be clonally expanded. When CD4+ T-cell levels fall, antigen levels rise and the risk of lymphomagenesis increases. Such proliferation allows for sequential genetic errors leading to a monoclonal and hence malignant transformation.

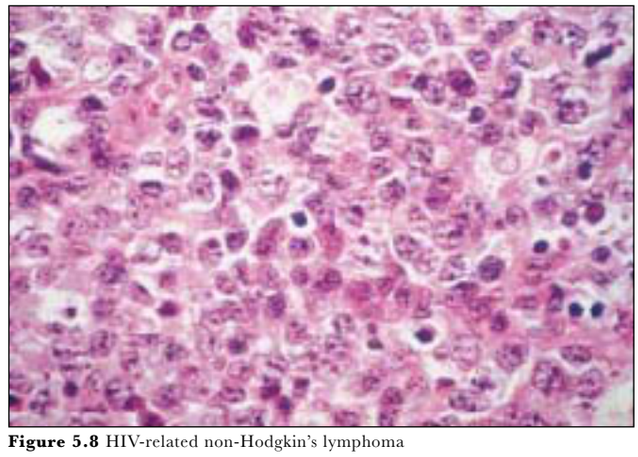


Figure 5.8 HIV-related non-Hodgkin's lymphoma

Pathology

Systemic lymphomas in AIDS are pathologically diverse. A diffuse small non-cleaved subset is unique to HIV patients and is associated with elevated IL-6 and soluble CD23 levels. It is less frequently associated with EBV than the diffuse immunoblastic or diffuse large cleaved cell subtypes. Histological type does not currently affect prognosis although the different subtypes are clinically separate. All these subtypes are high grade.

Immunohistochemistry reveals positive staining for CD20 in 90% of B-cell lymphomas.

Clinical presentation

NHL can occur at any stage of immunodeficiency with approximately one-third of patients with AIDS-related NHL having a previous AIDS diagnosis. Stage III or IV disease accounts for 70%–80% of cases (see Box 5.3) with a majority of patients presenting with extranodal disease. Common sites are the gastrointestinal tract, liver and bone marrow. Bone marrow involvement occurs in 20%–30% of cases and exacerbates chemotherapy-induced bone marrow toxicity.

NHL is associated with B symptoms of sustained fever greater than 38°C, weight loss (greater than 10% of body weight) and night sweats. All of these symptoms may occur in an HIV positive patient without NHL and so are of limited diagnostic and prognostic use in this clinical setting.

Box 5.3 Ann Arbor: classification of lymphoma

Stage I	Single lymph node region +/– local spread to extralymphatic tissue (E)
Stage II	Two or more node regions on same side of diaphragm +/– local spread to extralymphatic tissue (E)
Stage III	Involved nodes both sides of the diaphragm
Stage IV	Diffuse or disseminated involvement of one or more extralymphatic organs

Prognosis is poor with a median survival of 4–6 months in spite of an often good early response to treatment. A previous AIDS-defining illness, a CD4 count $<100 \times 10^6$, bone marrow involvement and poor performance status are all poor prognostic factors. In good prognosis patients the median survival is still only 11–14 months.

Treatment

Treatment of any lymphoma is based on its stage and grade and the patient's ability to withstand the rigors of treatment. For AIDS patients with their high-stage, high-grade disease this means chemotherapy. When faced with patients who are immunosuppressed and have poor bone marrow reserve before treatment the oncologist must make a balanced choice between reduced doses, which may compromise benefit, and quality of life. CHOP combination chemotherapy giving cyclophosphamide, vincristine and doxorubicin with oral prednisolone is delivered three weekly. Alternatively m-BACOD (methotrexate, Bleomycin, Adrianycin, Cyclophosphamide, Vincristine, Dexamethasone), another combination regimen, can be given. These regimens are toxic to bone marrow and in order to allow second and subsequent courses to be given on time patients may require GCSF. Prophylaxis against *Pneumocystis carinii* pneumonia should be considered. Allopurinol should be given to patients with bulky disease to prevent gout occurring when uric acid levels rise as the tumour breaks down. Patients with positive cytology or EBV DNA detected in their cerebrospinal fluid and those with meningeal or extensive sinus or base of skull disease require concomitant intrathecal methotrexate and cytosine arabinoside. Alternative chemotherapy regimens are undergoing trials but few have proved superior to CHOP and often result in worse immunosuppression and opportunistic infections.

For patients who have poor performance status, low CD4 counts and other AIDS diseases, palliative chemotherapy of vincristine plus prednisolone can be given. Radiotherapy is also useful for the palliation of symptoms caused by bulky disease. In the rare cases where NHL in AIDS presents as Stage I or II disease radiotherapy can be used as first-line treatment, avoiding the toxicity of chemotherapy.

Median survival is better in patients obtaining a complete response initially. In most studies half the patient deaths have been due to the lymphoma with remaining deaths being due to opportunistic infections.

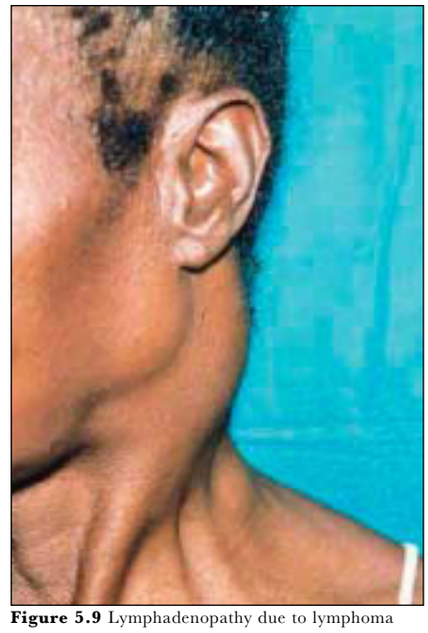


Figure 5.9 Lymphadenopathy due to lymphoma

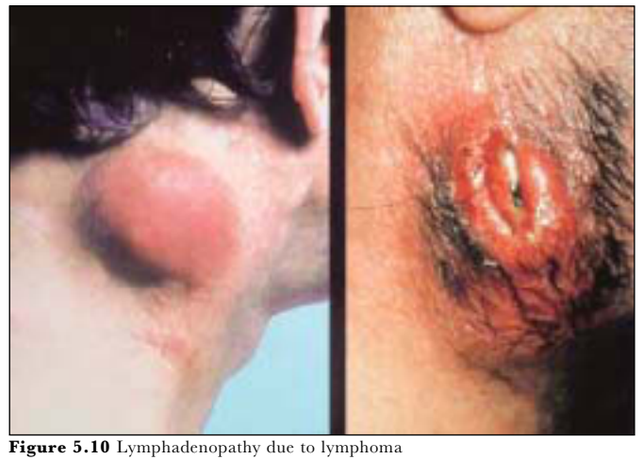


Figure 5.10 Lymphadenopathy due to lymphoma

Primary cerebral lymphoma

This is a strongly EBV-related process which occurs late in the clinical spectrum and accounts for around 15% of AIDS-related lymphomas. It is pathologically very similar to the post-transplant lymphoproliferative syndromes with EBV-latent gene expression. The presence of EBV DNA in the cerebrospinal fluid is highly predictive for primary central nervous system lymphoma (PCNSL). In the general population it is vary rare accounting for only 0.5–1.2% of all intracranial neoplasms which in themselves are rare. The incidence of PCNSL is also increased in other immunosuppressed conditions. PCNSL affected 2–6% of HIV positive individuals in the pre-HAART era. Typically patients have CD4 <50 cells/mm3 and a history of prior opportunistic infections. These lymphomas are always of the diffuse immunoblastic or diffuse large cleaved cell types. Prognosis is even worse than that for systemic lymphoma with median survival being only 1–2 months.

Cerebral lymphoma may be difficult to differentiate from cerebral toxoplasmosis, as both have a variable presentation ranging from subtle personality changes to seizures. Both

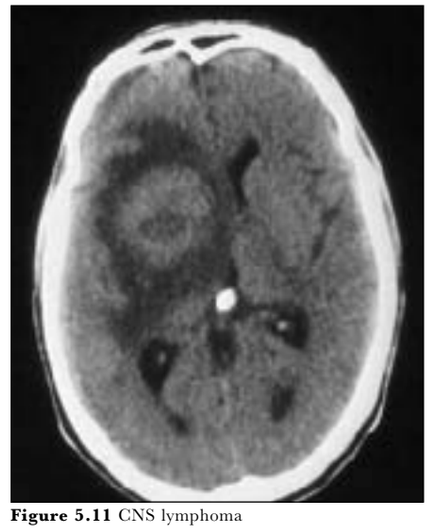


Figure 5.11 CNS lymphoma

commonly appear as multiple enhancing lesions on CT or MRI scanning. Initial treatment is usually empirically directed against toxoplasmosis. If this fails PCNSL can only reliably be confirmed by biopsy and many patients are reluctant to undergo such a procedure when life expectancy is limited. Once a definite diagnosis is made treatment is with high dose steroids and radiotherapy. Patients who respond tend to die of opportunistic infections reminding us that the underlying condition is advanced by the time PCNSL occurs.

Cervical carcinoma

For other malignancies in HIV the main predisposing factor is immune deficiency; however, the relationship between squamous cell neoplasia of the cervix and HIV is unique because of common sexual behaviour risk factors.

Viral DNA from high-risk types of the human papilloma virus (HPV16, 18, 31, 33, and 45) is found in 90% of all cervical cancers irrespective of HIV status. Not every woman with HPV infection develops cervical carcinoma and HPV infection alone is not sufficient for tumour development. Persistence of infection is probably important and other risk factors include smoking, oral contraceptive use and early pregnancy. HPV infection in HIV-infected women may represent reactivation of HPV types acquired in the past rather than recent acquisition of new types.

HIV positive women have a high rate of vulvo-vaginal infection which may make screening unreliable, regular Pap smears are therefore critical. Cervical intraepithelial neoplasia (CIN) is more commonly of a higher grade in HIV positive women and if invasive carcinoma ensues it is also more aggressive. A low threshold for referral for colposcopy is essential. Standard treatment strategies of ablation and excision have yielded disappointing levels of recurrence and patients need to be followed up very closely. If invasive disease ensues treatment is as for immunocompetent patients with surgery, radiotherapy and chemotherapy. Unfortunately treatment reactions are often severe with patients suffering severe vaginal mucositis.

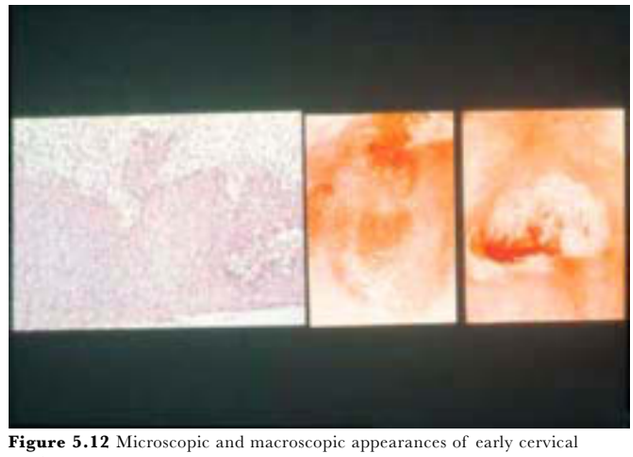


Figure 5.12 Microscopic and macroscopic appearances of early cervical carcinoma

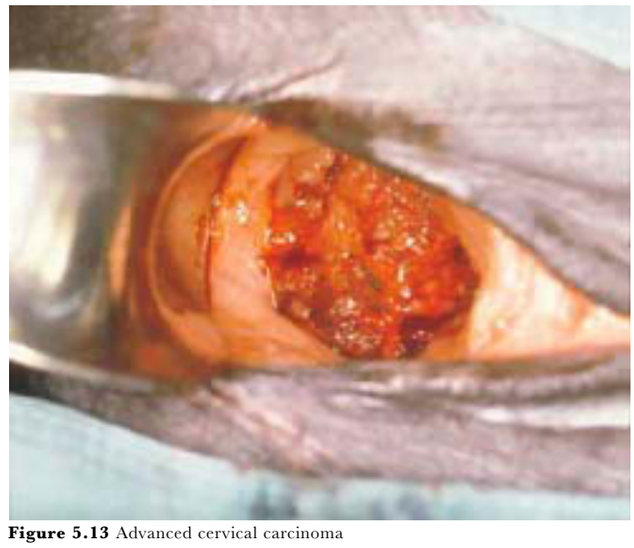


Figure 5.13 Advanced cervical carcinoma

Other associated malignancies

Hodgkin's disease

Hodgkin's disease is three to nine times more common in HIV patients compared with the general population. In Spain and Italy there is a high incidence of HIV amongst intravenous drug abusers who suffer more Hodgkin's disease than HIV sufferers in the UK. Research is needed here as in so many areas to discover the relevance of this observation. Hodgkin's disease tends to occur relatively early in HIV infection with a median CD4 cell count of 300/mm^3.

As with the non-Hodgkin's lymphomas, presentation is usually with bulky or advanced stage disease and 50% have bone marrow involvement. Most patients have B symptoms. In the HIV negative population Hodgkin's disease is typified by contiguous spread, this is not the case in HIV patients. Histologically tumours are usually high-grade mixed cellularity (41-100%) and lymphocyte-depleted subtypes (20%) and behave aggressively. Between 80% and 100% of Hodgkin's disease tissue from HIV-infected individuals is associated with EBV infection and this is probably relevant in pathogenesis. Treatment is with combination chemotherapy using standard regimens such as ABVD (doxorubicin, bleomycin, vinblastine and dacarbazine) and antiretrovirals. Bone marrow toxicity makes GCSF and dose reductions frequent necessities. If patients can continue their antiretroviral therapy throughout

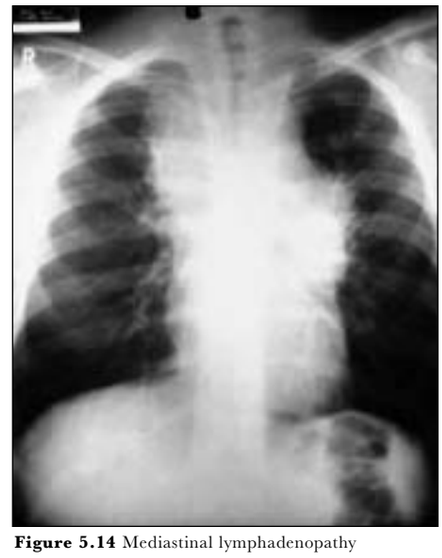


Figure 5.14 Mediastinal lymphadenopathy

chemotherapy they suffer less immunosuppression. Complete responses following chemotherapy are seen in 45–70%. Median survivals are 12–18 months and whilst not being an AIDS-defining condition 94% of patients progress to AIDS by 2 years. Good prognosis is associated with no prior AIDS diagnosis, CD4 $>250 \times 10^6$/l and complete response to treatment.

Ano-genital squamous cell carcinoma

Anal cancer like cervical cancer is related to human papillomavirus. HIV positive patients are two to six times more likely than HIV negative persons to have anal human papillomavirus infection. Persistence of infection is inversely related to CD4 count. Low-grade anal intraepithelial neoplasia is more likely to progress to high grade anal intraepithelial neoplasia in HIV positive patients. However it remains unclear whether HIV directly affects the development of anal carcinoma.

There is a threefold increase in incidence in testicular germ cell tumours in homosexual HIV positive men. Seminoma is much more common than teratoma. Lung cancer of all histological types, non-melanomatous skin cancers, angiosarcomas and paediatric leiomyosarcomas may all be increased in HIV infection. Lung cancers occur at an earlier age and have a poorer prognosis in the HIV positive population.

The effect of HAART

With the improved control of HIV replication brought about by combination antiretroviral therapy, the frequency of AIDS-related malignancy is falling. The incidence of KS has dropped by approximately 75%. Sadly there has been a smaller decline in the incidence of NHL, although primary CNS lymphoma has also markedly declined. The lack of change in the incidence of systemic lymphoma reflects a heterogeneous and complex pathophysiology, not as susceptible to the influence of HAART as KS.

The effect on HPV-associated anogenital squamous cell carcinoma has also been disappointing.

This varied response highlights the continued need for innovative treatments, both in terms of chemotherapy and the manipulation of the immune response.

Current research and the future

The rapidly increasing evidence on viral involvement in AIDS-associated malignancies suggests novel molecular targets for drug discovery using drug screening and molecular modelling. Vaccines for cancers occurring in patients with human papilloma viruses associated with cervical and ano-genital carcinoma and EBV in haematological malignancies are currently being researched. Other therapeutic approaches include biological therapy (for example IL-2, IL-12, IFN-α), immune-based therapy (for example antigen-presenting cells and monoclonal antibodies against B-cell targets) and angiogenesis inhibitors.

New assays to detect KSHV are now in use. Further work is needed on the cofactors influencing the progression of KSHV seropositive individuals to the development of KS. The anti-herpes drug cidofovir has activity against KSHV but it remains to be seen as to whether it is an effective treatment for KS.

To improve existing treatments the effects on the underlying HIV infection and the impact on the immune system of anti-tumour therapy need to be identified. As anti-HIV therapies have a clinical effect on tumour incidence, complex issues of drug–drug interactions and overlapping toxicities must be considered.

In the HAART era NHL is likely to become the most common malignancy associated with AIDS — new treatment strategies are urgently needed as treatment is currently extremely disappointing. Possibilities include the exploitation of cytokine networks, as we already know that these patients have low levels of IL-2 and IFN-γ but elevated IL-6. Treatment with low-dose IL-2 is already undergoing trials.

6 AIDS and the lung

Rob Miller

The lungs are commonly affected in patients infected with HIV, with over 60% of patients having at least one respiratory episode during the course of their disease. When immune responses are relatively well preserved in early HIV infection the pattern of respiratory infections is similar to that found in the general population, although they occur with greater frequency. The risk of opportunistic infections and tumours increases as progressive HIV-induced immunosuppression occurs. Over recent years there have been several changes in the pattern of lung disease seen in those infected with HIV. These changes may be accounted for by the widespread availability and uptake of prophylaxis for *Pneumocystis carinii* pneumonia and combination antiretroviral therapy (also known as highly active antiretroviral therapy or HAART).

Box 6.1 HIV-associated respiratory disease

Infections
Bacterial bronchitis/sinusitis
Bacterial pneumonia
Tuberculosis
P. carinii pneumonia
Fungal pneumonia
Cytomegalovirus pneumonitis

Malignancy
Kaposi's sarcoma
Lymphoma
Lung cancer

Non-malignant conditions
Lymphoid interstitial pneumonitis
Non-specific pneumonitis

Investigations

Symptoms of cough and dyspnoea with or without fever and sweats identify the presence of respiratory disease in HIV positive patients, but these are non-specific and symptomatic patients should be investigated.

Box 6.2 Investigation of respiratory disease

Non-invasive tests
Chest radiograph
Arterial blood gases or oximetry
Pulmonary function tests

Invasive tests
Induced sputum
Fibreoptic bronchoscopy and bronchoalveolar lavage with or without transbronchial biopsy
Open lung biopsy

Non-invasive investigations

These tests should ideally allow a specific diagnosis to be made and a therapeutic response monitored by a quick, cheap and universally available method. Unfortunately, none of these tests fulfils the criteria but they do help to:

- Determine the presence or absence of pulmonary disease.
- Assess disease severity.
- Determine if an invasive test is indicted to make an aetiological diagnosis.

Chest radiology

The chest radiograph may be normal in HIV positive patients with respiratory disease caused by *P. carinii* pneumonia. The most common abnormality seen in patients with pneumocystis pneumonia is bilateral perihilar haze which may be very subtle and easy to miss. More severely unwell patients may have more diffuse interstitial shadowing which may progress to severe consolidation with "white out" throughout both lung fields, with sparing of the apices and costophrenic angles. These radiographic appearances are non-specific and may also be seen in pyogenic bacterial, mycobacterial and fungal infection, and also in Kaposi's sarcoma and lymphoid interstitial pneumonitis. Between 5% and 10% of patients with pneumocystis pneumonia have atypical chest radiographs showing cystic changes, upper lobe infiltrates mimicking tuberculosis, hilar or mediastinal lymphadenopathy or focal consolidation. The chest radiograph in pneumocystis pneumonia may deteriorate very rapidly from being normal to showing severe abnormality in just a few days. By contrast, radiographic recovery can be slow. Nodular shadowing, adenopathy and pleural effusions on the chest radiograph suggest *Mycobacterium tuberculosis*, Kaposi's sarcoma or lymphoma.

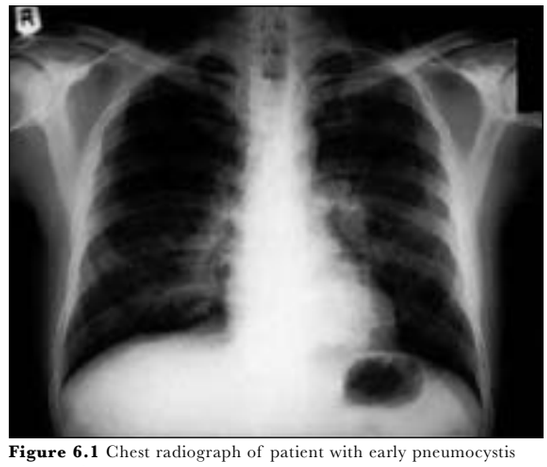


Figure 6.1 Chest radiograph of patient with early pneumocystis pneumonia

Arterial blood gases and oximetry
Hypoxaemia and a widened alveola–arterial oxygen gradient are very sensitive for the diagnosis of pneumocystis pneumonia but may also occur in other conditions. Exercise-induced arterial desaturation detected by oximetry is also sensitive for the diagnosis of pneumocystis pneumonia; desaturation may persist for several months following recovery from *P. carinii* pneumonia and occur also rarely in cytomegalovirus pneumonitis but is unusual in other respiratory conditions.

Pulmonary function tests
The single breath carbon monoxide transfer factor (TLCO), transfer coefficient (KCO), total lung capacity (TLC) and vital capacity (VC) may all be reduced in patients with pneumocystis pneumonia. Reductions in TLCO to 70% of predicted normal occur in HIV positive patients with pneumocystis and other respiratory disease, including Kaposi's sarcoma and bacterial infections, so this finding is not specific.

Invasive tests
These allow an aetiological diagnosis to be made.

Sputum induced by hypertonic saline
This procedure must be carried out away from other patients and staff in a separate room, ideally with "negative pressure" facilities in order to reduce the risk of nosocomial transmission of infection including tuberculosis. The patient inhales 20–30 ml of 2.7% (3 N) saline through an ultrasonic nebuliser. Saline deposits in the peripheral airways and alveoli, causing irritation and inducing bronchial secretion. Fluid is also drawn into the airways from the interstitium, loosening inflammatory exudate and casts from alveoli. These are mobilised by the mucociliary escalator and move centrally where they are coughed out by the patient. Careful preparation of the patient is needed, including starving for several hours before the procedure and rigorous cleansing of the mouth to remove oral debris so that the sputum sample is not contaminated (food debris and squames take up stain and make analysis difficult). Purulent samples of sputum suggest a bacterial cause. *P. carinii* infection is usually found in clear "saliva-like" samples that become viscid on cooling to room temperature. Fungal infection and mycobacterial infection may also be diagnosed by this technique. Many centres do not carry out sputum induction because of the need for special equipment and the low yield when the technique is compared with fibreoptic bronchoscopy, both for the diagnosis of pneumocystis pneumonia and other pathogens. Some patients find sputum induction unpleasant and become nauseated or dyspnoeic. Arterial desaturation may also occur during the procedure.

Fibreoptic bronchoscopy
Bronchoscopy allows inspection of the bronchi to be carried out and lesions of Kaposi's sarcoma may be identified. Bronchoalveolar lavage is routinely carried out from the middle lobe or from the area of maximum abnormality seen on the chest radiograph. Transbronchial biopsies are now rarely done as they add little to the diagnostic yield for *P. carinii* and other diagnoses, and the technique is associated with adverse effects including haemorrhage and pneumothorax. If transbronchial biopsy is not performed a diagnosis of non-specific or lymphocytic interstitial pneumonitis might be missed.

Open lung biopsy
It is rarely necessary to carry out open lung biopsy because of the high yield from bronchoalveolar lavage. This investigation may be necessary if fibreoptic bronchoscopy and lavage fail to

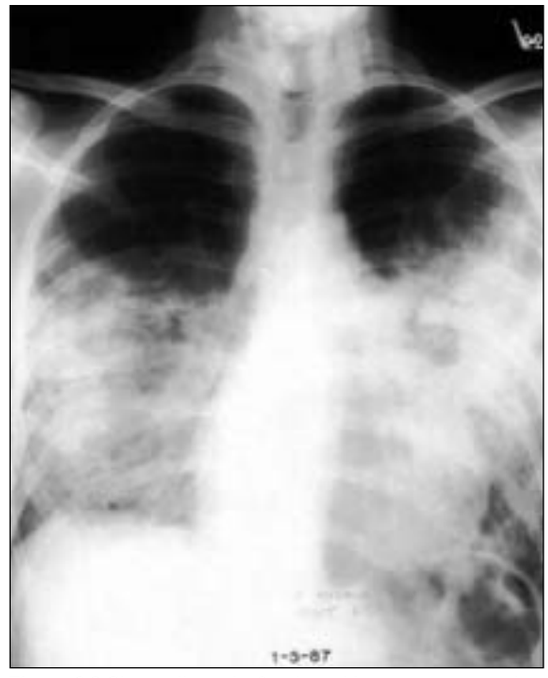


Figure 6.2 Chest radiograph of patient with severe pneumocystis pneumonia

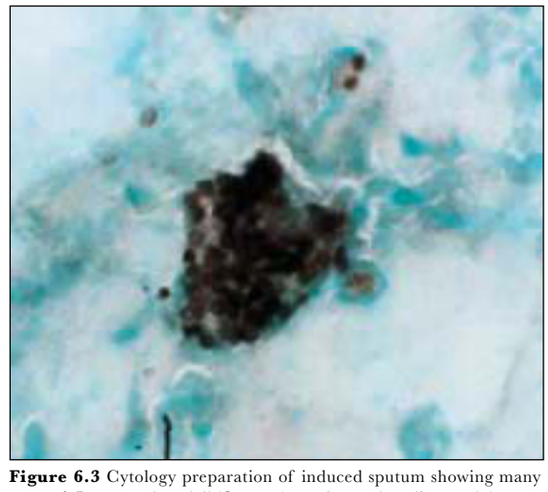


Figure 6.3 Cytology preparation of induced sputum showing many cysts of *Pneumocystis carinii* (Grocott's methenamine silver stain)

Box 6.3 Open lung biopsy
If fibreoptic bronchoscopy and lavage fail to identify diagnosis
or
where patient with bronchoscopic diagnosis deteriorates despite specific treatment

identify a diagnosis or in cases where a patient with a bronchoscopic diagnosis, deteriorates despite specific treatment.

The presenting clinical features and treatment of the common pulmonary manifestations of HIV disease are described below.

Pneumocystis carinii pneumonia

Despite widespread use of anti-pneumocystis prophy!axis and HAART, *P. carinii* pneumonia remains a common AIDS-defining diagnosis in patients who at presentation with pneumonia are unaware of their HIV serostatus or who, despite knowing they have HIV infection, are non-compliant with or intolerant of their prophylaxis and/or HAART.

Patients complain of a non-productive cough and increasing dyspnoea (over two to three weeks or more); they may also have fever and sweats. The chest radiograph may be normal or show interstitial infiltrates: in severe pneumonia there may be widespread alveolar consolidation.

Table 6.1 Grading of severity of *P. carinii* pneumonia

	Mild	Moderate	Severe
Symptoms and signs	Increasing exertional dyspnoea, with or without cough and sweats	Dyspnoea on minimal exertion, occasional dyspnoea at rest, fever with or without sweats	Dyspnoea at rest tachypnoea at rest, persistent fever, cough
Blood gas tensions (room air)	PaO_2 >11.0 kPa	PaO_2 8.0–11·0 kPa	PaO_2 <8.0 kPa
SaO_2 (at rest)	>96%	91–96%	<91%
Chest radiograph	Normal or minor perihilar interstitial infiltrates	Diffuse shadowing	Extensive interstitial shadowing with or without diffuse alveolar shadowing

PaO_2 = partial pressure of oxygen; SaO_2 = arterial oxygen saturation, measured with a transcutaneous pulse oximeter.

Treatment

It is important to assess the severity of the pneumonia in order to choose appropriate treatment, as some drugs are ineffective in severe disease. High-dose co-trimoxazole remains the "gold standard" treatment. Treatment is for 21 days, given intravenously for the first 10–14 days, diluted in 1 in 25 of 0.9% saline, subsequently, orally. Patients with mild disease may be treated with oral co-trimoxazole from the outset. The principal side-effects are nausea and vomiting, leucopenia and rash. Routine use of folic or folinic acid does not prevent leucopenia and may be associated with increased therapeutic failure. HAART is usually stopped while co-trimoxazole is being given to avoid profound myelosuppression. Conventionally used doses of co-trimoxazole (20 mg/kg day of the trimethoprim component) may be excessive: dose reduction to 75% of this dose (to maintain serum trimethoprim concentrations at 5–8 µg/ml) has equivalent efficacy and reduced toxicity.

Table 6.2 Treatment of *P. carinii* pneumonia

Choice	Mild	Moderate	Severe
First	Co-trimoxazole	Co-trimoxazole	Co-trimoxazole
Second	Clindamycin and primaquine or Dapsone and trimethoprim or Atovaquone	Clindamycin and primaquine or Dapsone and trimethoprim or Atovaquone	Clindamycin and primaquine or Trimetrexate and folinic acid or Pentamidine iv
Third	Pentamidine iv	Pentamidine iv or Trimetrexate and folinic acid	
Glucocorticoids	Unproven benefit	Of benefit	Of benefit

Alternative treatment regimens include:

Clindamycin–primaquine combination (clindamycin 600 mg ×4/day iv or orally and primaquine 15 mg/day orally) has been used in patients intolerant of, or failing to respond to, co-trimoxazole. Principal side-effects are rash, nausea and vomiting, and leuco(neutro)penia.

Dapsone–trimethoprim (100 mg/day dapsone and 20 mg/kg/day trimethoprim) given orally for 21 days is as effective as oral co-trimoxazole in mild to moderate disease and is better tolerated by patients. Side-effects include methaemoglobinaemia and hyperkalaemia, nausea and rash.

Atovaquone suspension (750 mg × 2/day) given orally for 21 days is less effective (and less toxic) than either co-trimoxazole or pentamidine for mild to moderate disease. Absorption from the gut is variable but may be increased if taken with food.

Pentamidine is not often used because of significant toxicity and because other regimens have similar efficacy and less toxicity. It is given at a dose of 4 mg/kg/day (of the isethionate salt) given diluted in 250 mg 5% dextrose by slow intravenous infusion (over 2 hours); it should not be given by intramuscular injection. The major side-effects are hypotension and hypoglycaemia; nephrotoxicity with increases in creatinine and urea concentrations may occur. Dose reduction to 3 mg/kg/day is associated with reduced toxicity but may be less effective. Blood pressure and blood glucose concentrations should be closely monitored. Response to pentamidine (defervescence of fever, reduction in dyspnoea and improvement in blood gases) may take longer (4–7 days) than intravenous co-trimoxazole.

Box 6.4 Adjuvant glucocorticoids in moderate/severe pneumonia

- Reduce risk of respiratory failure (by 50%)
- Reduce risk of death (by 33%)
- Should be started at same time as specific anti-pneumocystis treatment

Nebulised pentamidine is now no longer used to treat *P. carinii* pneumonia as there are several other more effective therapies and because this form of treatment does not suppress the development of extrapulmonary pneumocystosis.

Adjuvant glucocorticoids for patients with moderate or severe pneumocystis pneumonia reduces the risk of respiratory failure (by up to 50%) and the risk of death (by up to 33%). Glucocorticoids should be started together with specific anti-pneumocystis treatment in any patient presenting with a PaO_2 of 9.3 kPa breathing air. In some patients this will be on the basis of a presumptive diagnosis; clearly there will be a need to confirm the diagnosis rapidly. Treatment is with intravenous methylprednisolone 1 g/day for three days, followed by 0.5 g for two days, followed by oral prednisolone 40 mg daily tailing off over 10 days. Alternatively, prednisolone 40 mg orally twice daily is given for 5 days and then gradually reduced over 21 days (or intravenous methylprednisolone is given at 75% of these doses).

Figure 6.4 Transfer to the ICU may be necessary if respiratory failure occurs

Intensive care

Over 90% of patients respond to treatment and survive their first episode of pneumocystis pneumonia. In those who fail to respond and who develop respiratory failure, mortality is 50%. Transfer to the intensive care unit for mask CPAP ventilation or intubation and mechanical ventilation should be considered in this situation. When considering the appropriateness of intensive care, assess the patient's wishes and those of their partner and relatives as well as the patient's previous and expected quality of life in relation to their HIV disease.

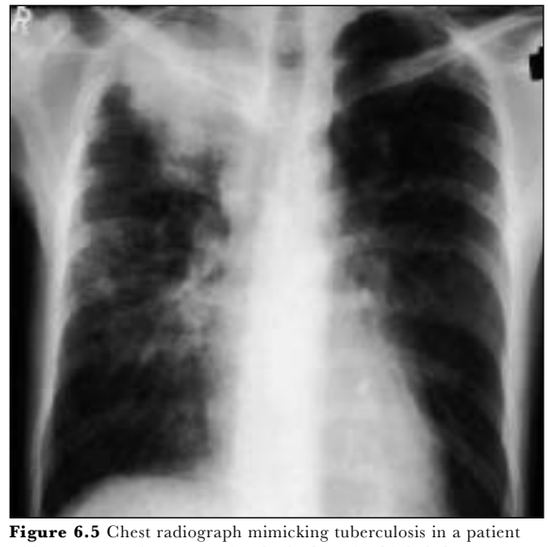


Figure 6.5 Chest radiograph mimicking tuberculosis in a patient with pneumocystis pneumonia who had received inhaled pentamidine prophylaxis

Prophylaxis

HIV positive patients, including those receiving HAART should receive primary prophylaxis against *P. carinii* pneumonia if they have a CD4 count < 200 cells/μl or a history of oral/pharyngeal candidiasis or if they have a CD4 lymphocyte count <14% of total lymphocyte count, or if they have other AIDS-defining diagnoses, for example Kaposi's sarcoma, regardless of CD4 count. If close monitoring of CD4 counts (at least every three months) is not feasible then prophylaxis should be considered for patients with CD4 counts between 200 and 250 cells/μl. Secondary prophylaxis is given to all HIV-infected patients after an episode of *P. carinii* pneumonia, regardless of CD4 count.

The prophylaxis regimen of choice is co-trimoxazole 960 mg once daily. A dose of 480 mg once daily or 960 mg three time a week are also effective and may be better tolerated by the patient. Co-trimoxazole also protects against bacterial infection and reactivation of cerebral toxoplasmosis. In patients who develop mild to moderate adverse reactions to co-trimoxazole, desensitisation may be attempted before changing to alternative therapy. Second-line prophylaxis (for those intolerant of, or unwilling to take, co-trimoxazole) include dapsone, with or without pyrimethamine, atovaquone or monthly nebulised pentamidine (300 mg given using a Respirgard II or similar nebuliser). There is a higher relapse rate of pneumocystis pneumonia with this regimen compared with that using co-trimoxazole. Some patients who relapse while receiving nebulised pentamidine have atypical chest radiographs with upper zone infiltrates which mimic tuberculosis. Atovaquone is as effective as dapsone or nebulised pentamidine but is much more expensive.

Stopping prophylaxis

Primary *P. carinii* prophylaxis can be discontinued in HIV-infected patients responding to HAART with an increase in CD4 count from below 200 cells/μl to above 200 cells/μl and a reduction in HIV-1 viral load, both sustained for 3–6 months. If

Box 6.5 Prophylaxis of *P. carinii* pneumonia

Primary prophylaxis
Any HIV positive patient with a CD4 count of <200 cells/μl
Or a history of oral/pharyngeal candidiasis
Or a CD4 count <14% of total lymphocyte count
Or another AIDS-defining diagnosis, for example Kaposi's sarcoma

Secondary prophylaxis
Any HIV positive patient *after* an episode of *P. carinii* pneumonia

despite HAART, CD4 counts again fall below 200 cells/μl and HIV-1 viral load rises, then the criteria for starting primary prophylaxis should be used. There are insufficient data to support the discontinuation of secondary prophylaxis.

Bacterial infections

Upper respiratory tract infections and pyogenic bacterial infection (sinusitis, bronchitis and pneumonia) occur more often in HIV-infected individuals than in the general population. Bacterial infections are particularly common in HIV positive intravenous drug users. The most commonly isolated organisms are *Streptococcus pneumoniae* and *Haemophilus influenzae*. Severe pneumonia due to *Staphylococcus aureus* or Gram negative bacteria such as *Pseudomonas aeruginosa* also occurs, especially in the later stages of AIDS. Respiratory infection may occur with rapid onset, the patient complaining of a cough with or without sputum and fever with chills; patients are frequently bacteraemic. There is a high rate of complications including intrapulmonary abscess formation and empyema. A rapid response usually occurs to treatment with appropriate antibiotics but relapse may occur. Some groups recommend that all HIV positive patients should be immunised with polyvalent pneumococcal polysaccharide vaccine although not all studies have demonstrated effective antibody responses to this agent, particularly in patients with CD4 counts <200 cells/μl.

Kaposi's sarcoma

Pulmonary Kaposi's sarcoma is the commonest non-infectious pulmonary manifestation of AIDS. Almost all patients with pulmonary Kaposi's sarcoma have mucocutaneous or lymph node Kaposi's sarcoma. Palatal Kaposi's sarcoma (with or without mucocutaneous Kaposi's sarcoma) strongly predicts for the presence of pulmonary Kaposi's sarcoma. Pulmonary Kaposi's sarcoma can affect the pulmonary parenchyma, bonchi, pleura and hilar/mediastinal lymph nodes. Chest radiographs most frequently show non-specific features, with bilateral interstitial (often nodular) or alveolar infiltrates; more than 40% of patients have pleural effusion and 25% have mediastinal lymph node enlargement. Routine respiratory function tests show decreased lung, volumes (FEV_1 and FVC) and decreased TLCO; airflow obstruction may occur with extensive Kaposi's sarcoma in the airways.

At fibreoptic bronchoscopy 45% of patients with pulmonary Kaposi's sarcoma have visible endotracheal and endobronchial lesions consisting of multiple, red or purple, flat or raised lesions. Biopsy is not routinely done as most patients have the diagnosis made by the presence of cutaneous Kaposi's sarcoma, because of the risk of haemorrhage (up to 30% will have a significant bleed) and the low diagnostic yield (<20%) which occurs because of submucous distribution of the tracheobronchial tumour.

Transbronchial biopsy also has a low yield of less than 20% due to the patchy nature of parenchymal disease. Histological diagnosis is difficult to make at bronchial or transbronchial biopsy as crush artefact and reactive fibrous tissue have similar appearances. Open lung biopsy has a diagnostic yield of >75% but this procedure is very invasive and should probably be avoided as patients with pulmonary Kaposi's sarcoma have a poor prognosis.

Treatment

Chemotherapy most often consists of bleomycin 10 000 units/m^2 and vincristine 2 mg once every three weeks. Liposomal formulations of daunorubicin and doxorubicin may

Box 6.6 Bacterial infections

- Increased incidence of sinusitis, bronchitis and pneumonia in HIV infected persons, compared to general population
- Bacterial infection especially common in HIV infected IVDU

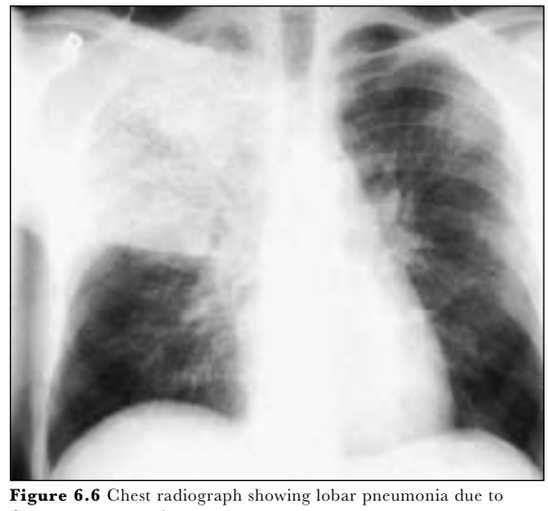


Figure 6.6 Chest radiograph showing lobar pneumonia due to *Streptococcus pneumoniae*

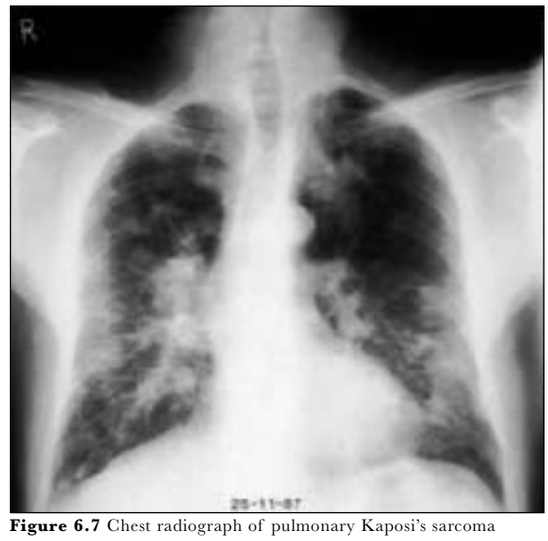


Figure 6.7 Chest radiograph of pulmonary Kaposi's sarcoma showing multiple pulonary nodules

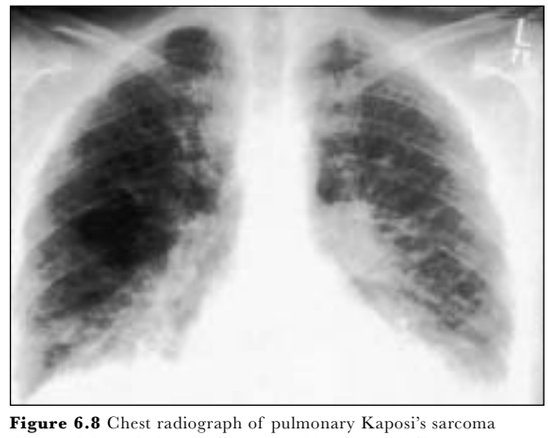


Figure 6.8 Chest radiograph of pulmonary Kaposi's sarcoma showing bilateral pleural effusions and interstitial infiltrates

also be used as single-agent chemotherapy. Treatment of pleural effusions (which occur secondary to Kaposi's sarcoma on the visceral pleura or to mediastinal glands) is problematical. Chemical pleuradesis is rarely successful and radiotherapy has not been shown to be of value.

Tuberculosis

Tuberculosis may speed up the natural history of HIV disease

Unlike opportunistic infections in AIDS tuberculosis is also infectious for healthy individuals. Tuberculosis is a potent stimulator of cell-mediated immunity and so may speed up the natural history of HIV disease. The incidence of tuberculosis is currently increasing in the USA; this is directly attributable to the effects of HIV in certain populations. No increase has occurred yet in Britain but the unpredictable features of the HIV epidemic in heterosexuals, migrants and injecting drug users means careful vigilance is required. Tuberculosis can precede the development of AIDS, be diagnosed at the same time or occur at any time during established AIDS. Tuberculosis in HIV positive patients is AIDS defining and in the USA, the UK and most other European countries is a statutorily notifiable disease.

Over two thirds of cases of tuberculosis in HIV-infected patients present with pulmonary disease. Clinical presentation varies according to the stage of HIV disease. Early on, with relatively well preserved cell-mediated immunity, pulmonary tuberculosis resembles classic adult post-primary disease with upper lobe infiltrates and cavitation; the tuberculin test is usually positive and acid and alcohol fast bacteria (AAFB) are frequently seen when sputum is examined by microscopy. With advanced HIV disease and destroyed cell immunity, presentation is non-specific with fever, weight loss and fatigue, with or without cough. Patients with low CD4 counts $<150 \times 10^6/1$ may also have extrapulmonary disease affecting bone marrow, lymph node, central nervous system or liver. In the chest, the clinical pattern is one of primary infection with hilar and mediastinal adenopathy, diffuse or miliary shadowing; pleural effusions are common. Cavitation occurs rarely and up to 10% of chest radiographs are normal. The tuberculin test is usually negative, sputum (and bronchoalveolar lavage) are often smear negative and culture may also be negative.

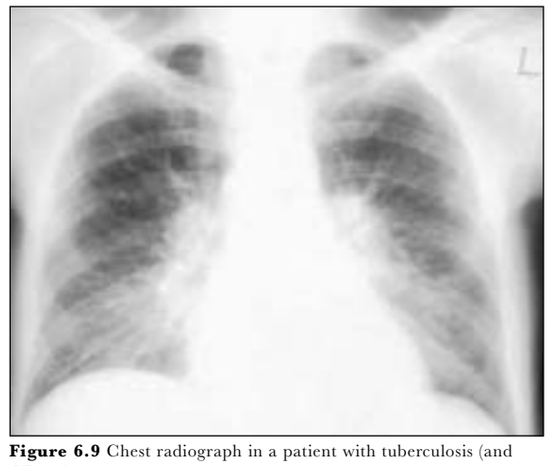


Figure 6.9 Chest radiograph in a patient with tuberculosis (and CD4 of $100 \times 10^6/1$) showing hilar lymphadenopathy

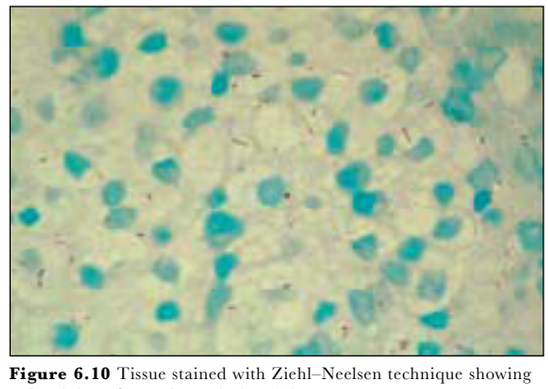


Figure 6.10 Tissue stained with Ziehl–Neelsen technique showing red staining of mycobacteria (× 400)

As culture and species identification may take up to six weeks, *M. tuberculosis* infection should be assumed if AAFB are found in respiratory sample, an aspirate or biopsy site, or blood, and conventional antituberculous therapy should be started. Treatment can be modified if culture subsequently reveals an atypical mycobacterium and not *M. tuberculosis*.

Treatment

Clinical response to conventional treatment with four-drug regimens is good, but compared with the non-HIV-infected general population, survival is poor. The incidence of adverse reactions to antituberculous drugs, including isoniazid, rifampicin and thiacetazone, is higher in HIV-infected patients than in the general population.

Many of the drugs used to treat tuberculosis share routes of metabolism and elimination or have overlapping toxicities with other medication taken by HIV-infected patients, so there exists the potential for drug–drug interactions. For example, rifampicin renders dapsone, as prophylaxis of *P. carinii* pneumonia, ineffective — by inducing its hepatic metabolism. Clinically important drug–drug interactions occur between rifampicin/rifabutin and antiretroviral therapy particularly protease inhibitors such as ritonavir and non-nucleoside reverse transcriptase inhibitors such as delavirdine.

Compliance with therapy is a problem in some groups and directly observed therapy (DOTS) may be needed.

Box 6.7 Treatment of tuberculosis

- Conventional four-drug regimens are associated with good response
- Adverse reactions to anti-tuberculous therapy occur more frequently in HIV infected patients
- Important drug-drug interactions occur between drugs used to treat tuberculous and drugs used to treat HIV infection

Multi-drug resistant (MDR) tuberculosis
This is tuberculosis resistant to rifampicin and isoniazid with or without resistance to other drugs. Outbreaks of MDR tuberculosis have occurred in the USA, the UK and elsewhere in Europe. Most MDR tuberculosis arises because of inadequate treatment or poor compliance with therapy. Some cases occur in HIV-infected patients who are exogenously re-infected whilst receiving treatment for drug-sensitive disease. Despite treatment, MDR tuberculosis has a poor prognosis in HIV-infected and non-infected patients and healthcare workers who acquire the infection.

Chemoprophylaxis
Some expert groups, for example WHO and International Union Against Tuberculosis and Lung Disease (IUATLD), recommend that HIV-infected patients co-infected with *M. tuberculosis* (but without disease) should receive chemoprophylaxis. There are few data to support this policy. The preferred policy should be close clinical monitoring rather than chemoprophylaxis, because of increasing rates of drug resistance and MDR tuberculosis, difficulties in distinguishing between infection and disease, and concerns that single-drug prophylaxis is associated with the development of resistance.

Once clinical disease is excluded, HIV-infected patients who have had recent contact with a smear-positive index case should receive chemoprophylaxis and life-time follow up should be instituted. Once HIV-infected patients have successfully completed a course of treatment for tuberculosis close clinical monitoring is recommended: such patients do not need to take life-long secondary prophylaxis.

Box 6.8 Chemoprophylaxis of tuberculosis

- Once active tuberculosis excluded, close clinical monitoring, rather than chemoprophylaxis is preferred policy in patients co-infected with HIV and tuberculosis
- HIV-infected patients in close contact with a smear positive index case should receive chemoprophylaxis and careful follow-up.

Fungal pneumonia

Infection with *Cryptococcus neoformans*, *Histoplasma capsulatum*, *Aspergillus fumigatus* and other fungi is well recognised in HIV positive patients in the USA and Africa. Infection with these organisms is relatively uncommon in the UK. Cryptococcal pneumonia often occurs as part of a disseminated infection with fungaemia and meningoencephalitis; respiratory symptoms of cough and dyspnoea are non-specific. The chest radiograph may be normal or show diffuse shadowing which may be nodular. Diagnosis is made by culture of bronchoalveolar lavage or transbronchial biopsy specimen (or blood, bone marrow, or cerebrospinal fluid in disseminated infection). Treatment of cryptococcus infection is with fluconazole 400–600 mg/day or intravenous amphotericin B or itraconazole 400 mg twice a day. Aspergillus pulmonary infection has a very poor prognosis despite treatment with amphotericin. It occurs almost exclusively in patients with advanced HIV disease who are either neutropenic or who have received broad-spectrum antibiotics.

Box 6.9

- Cryptococcal pneumonia often part of disseminated infection
- Respiratory symptoms of cough and dyspnoea non-specific
- Chest radiograph may be normal or show diffuse shadowing

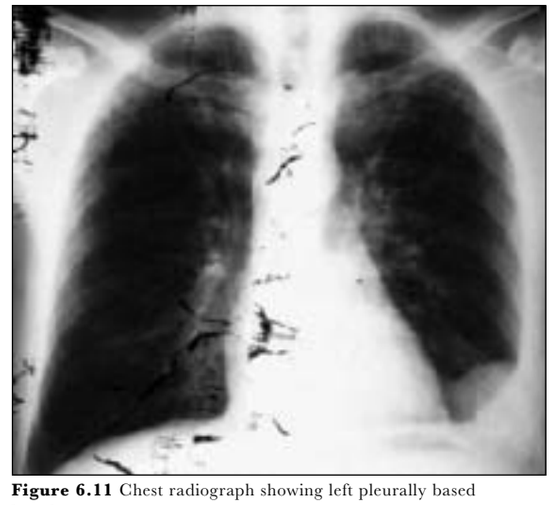


Figure 6.11 Chest radiograph showing left pleurally based lymphoma

Lymphoma

Lymphoma occurs more often in HIV positive patients, particularly in those with advanced HIV disease. Most lymphomas are B cell in origin and are of high grade. Intrathoracic disease most frequently occurs in the context of disseminated disease. Symptoms are non-specific. The chest radiograph may show mediastinal lymphadenopathy, pleural lesions or focal parenchymal abnormalities. The prognosis is poor and there is a high relapse rate after treatment. Median survival is <1 year, reflecting the advanced stage of HIV disease.

Lymphocyte interstitial pneumonitis

This condition occurs more commonly in children; it is unusual in HIV-infected adults. Parotid enlargement and lymphocytic infiltration of the liver and bone marrow may accompany pulmonary involvement. Patients often present with slowly progressive dyspnoea and cough, symptoms that cannot be distinguished from infection. Examination of the chest may be normal or reveal fine end inspiratory crackles. The chest radiograph usually shows bilateral reticulonodular infiltrates but may show diffuse shadowing and thus mimic *P. carinii* pneumonia. Diagnosis is made by transbronchial biopsy or open lung biopsy. Some patients have been shown to respond to HAART and others to treatment with prednisolone 60 mg once a day.

Box 6.10 Lymphocytic interstitial pneumonitis

- Commoner in children
- Parotid enlargement and lymphocytic infiltration of liver/bone marrow may also occur
- Presents with slowly progressive dyspnoea/cough
- Treatment is with HAART or prednisolone

Non-specific pneumonitis

This condition is important as patients present with symptoms and chest radiographic appearances similar to those of *P. carinii* pneumonia. It may also occur when the CD4 count is still normal. The diagnosis can only be made by biopsy. Episodes are usually self limiting but prednisolone may be of benefit.

Cytomegalovirus

Cytomegalovirus (CMV) infection in HIV positive patients with advanced disease and low CD4 counts (<100 cells/μl) is common and is a well-documented cause of retinitis, colitis, adrenalitis and radiculopathy. In patients with renal allografts and bone marrow transplants, CMV may cause pneumonitis on an immunopathogenic basis and this is frequently fatal.

CMV was originally thought to be an important cause of pneumonitis in patients with AIDS but it is now known that CMV pulmonary infection occurs only rarely in the absence of other pathogens and its presence does not adversely affect outcome and survival. Treatment with specific anti-CMV treatment such as foscarnet (phosphonoformate) does not seem to improve outcome (as would be expected if CMV was causing the pneumonitis).

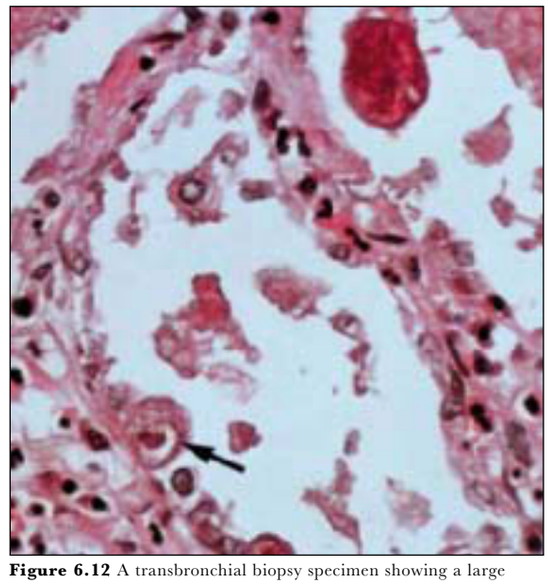


Figure 6.12 A transbronchial biopsy specimen showing a large eosinophilic nuclear inclusion (arrowed) in a pneumocyte infected with cytomegalovirus (haematoxylin and eosin stain)

Lung cancer

Lung cancer in HIV-infected patients presents an earlier age than in the general population and appears to have a poorer outcome, as it is more aggressive. Smoking is strongly associated with the development of lung cancer.

Figure 6.3 is reproduced courtesy of Dr Gabrijela Kocjan; Figures 6.10 and 6.12 are reproduced courtesy of Dr Meryl Griffiths.

7 Gastrointestinal and hepatic manifestations

Ian McGowan, Ian VD Weller

Gastrointestinal symptoms are a common manifestation of HIV infection. Significant clinical problems tend to occur in patients with advanced immunosuppression. The differential diagnosis of gastrointestinal disease is broad and includes opportunistic infection, malignancy, and the effects of medication. Antiviral drugs and antibiotics have gastrointestinal side effects such as nausea, vomiting, and diarrhoea. HIV can be readily detected in mucosal tissue but the direct role of mucosal HIV infection in the cause of clinical disease remains controversial.

This chapter will focus on the differential diagnosis and management of common gastroenterological syndromes associated with HIV infection. Clinical investigation may not always be appropriate in advanced disease. It is important to counsel patients about the risks and benefits of invasive procedures as many "specific" diagnoses may not be treatable.

Box 7.1 Differential diagnosis of HIV-associated gastrointestinal disease

- Infection
- Malignancy
- Medication
- HIV infection

Oral and oesophageal disease

Oral cavity pain or discomfort are caused by candidiasis, herpetic or aphthous ulceration, periodontal disease, and tumours. Often the diagnosis can be made by simple inspection and appropriate treatment initiated without further investigation. Systemic oral therapy of herpes simplex ulceration and candidiasis is preferred for reasons of efficacy and ease of use. Recurrence is common and if frequent, maintenance therapy may be required rather than the short treatment of each occurrence. Maintenance therapy may be more likely to induce resistance.

About one third of patients develop oesophageal disease. The likelihood of candidiasis is so high that a therapeutic trial with a systemic antifungal agent is indicated before considering further investigation. If symptoms fail to respond, or recur despite adequate maintenance therapy, endoscopy is performed to exclude herpes simplex, cytomegalovirus and other causes of oesophageal ulceration including malignant lesions.

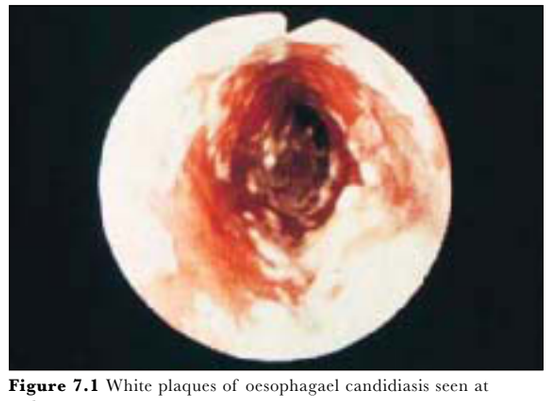


Figure 7.1 White plaques of oesophagael candidiasis seen at endoscopy

Diarrhoea

Patients with diarrhoea lasting more than two weeks should be investigated. The diagnostic yield is likely to be highest in patients with CD4 counts $<200 \times 10^6$/l. Careful microbiological and parasitological examination of multiple stool specimens is the most cost-effective initial investigation. Endoscopy with collection of tissue from the distal duodenum, ascending and descending colon should be performed to exclude cytomegalovirus and occult parasitic infection.

Bacterial infection with *Campylobacter*, *Salmonella* or *Shigella* spp. may present with severe diarrhoeal symptoms and/or bacteraemia. It is important to exclude toxic megacolon with plain abdominal radiography. Organisms are usually sensitive to conventional therapy but drugs may need to be given parenterally. Evidence of atypical mycobacterial infection is found in 60% of patients with advanced HIV disease at necropsy. Gastrointestinal infection may be associated with fever, weight loss, diarrhoea, and malabsorption. Diagnosis can be made by acid fast staining of the stool or biopsy material and by culture. Positive stool culture alone indicates colonisation only. *Mycobacterium tuberculosis* infection of the bowel does occur but is less common. Antibiotic-associated diarrhoea,

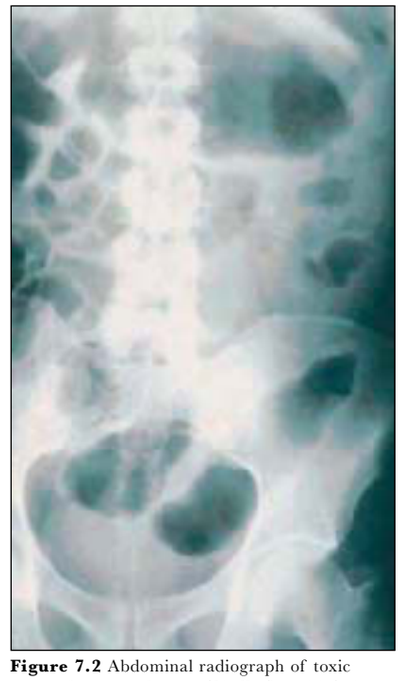


Figure 7.2 Abdominal radiograph of toxic megacolon secondary to *Shigella flexneri infection*

including pseudomembranous colitis due to *Clostridium difficile*, is occasionally seen in patients with HIV infection and is treated with oral metronidazole or vancomycin.

Cryptosporidium spp. is one of the most common pathogens isolated from HIV-infected patients with diarrhoea. The degree of immunosuppression influences patient prognosis and patients with a CD4 count $>200 \times 10^6$/l may recover spontaneously. Treatment is supportive as no agent has shown convincing efficacy. The organism is heat sensitive and immunosuppressed patients are advised to boil water for drinking purposes.

Microsporidia are also an important cause of diarrhoea as well as being associated with hepatitis, peritonitis, sclerosing cholangitis, sinusitis, and renal failure. Diagnosis is difficult as the spores are only 1–5μm in diameter. A number of centres have reported successful identification of spores in stool using trichrome and fluorescent stains but morphology is best determined using electron microscopy. Albendazole has shown promise in AIDS patients with microsporidiosis but may only be active against *Encephalitozoon intestinalis* and not *Enterocytozoon bienusi*.

Isospora belli is an infrequent cause of diarrhoea in AIDS patients in the USA and Europe but accounts for up to 25% cases of chronic diarrhoea in patients in tropical and subtropical countries. Response to trimethoprim–sulphamethoxazole has been described.

Cyclospora sp. is the most recent protozoan to be associated with diarrhoea in AIDS. It appears to be more common in the developing world and in returning travellers and like *Isospora belli* appears to be sensitive to trimethoprim–sulphamethoxazole.

Other protozoa including *Entamoeba histolytica* are frequently identified in stools from HIV-infected homosexual men but appear not to be pathogenic.

Cytomegalovirus colitis occurs in less than 5% of patients with AIDS. Symptoms include bloody diarrhoea, abdominal pain, and fever. Sigmoidoscopy may show diffuse erythema and mucosal ulceration. Diagnosis is histopathological and is made on the basis of characteristic intranuclear "owl's-eye" inclusion bodies or detection of CMV antigen with monoclonal antibodies. Treatment is with ganciclovir or foscarnet.

Adenoviruses have been identified by culture and electron microscopy in HIV-infected homosexual men with diarrhoea. No specific treatment is available.

Weight loss and anorexia

Weight loss is a major problem in AIDS and directly influences survival. The causes of weight loss are complex and several factors may coexist in individual patients. Anorexia may occur secondary to drug therapy, opportunistic infection, taste disturbance, or oral discomfort, resulting in inadequate food intake. Malabsorption of fat, lactose, vitamin B12, and bile salts has been demonstrated.

Simple dietary measures such as encouraging smaller, more frequent, meals may be helpful and a wide variety of nutritional supplements are available. Appetite stimulants such as megestrol acetate may be beneficial but weight gain is usually modest. Recombinant human growth hormone, although expensive, may partially reverse HIV-associated weight loss. In patients unable to tolerate oral feeding, enteral and parenteral feeding are alternative forms of nutrition but their efficacy and place in management are still being evaluated. Enteral nutrition offers a safer and cheaper alternative to total parenteral nutrition which is perhaps most useful in patients with severe diarrhoea, nausea, and vomiting, in whom fluid balance and control of symptoms has been difficult.

Box 7.2 Infective causes of diarrhoea

- Bacteria
  - campylobacter, salmonella, shigella
  - atypical mycobacteria
  - *Clostridium difficile*
- Protozoa
  - cryptosporidium
  - microsporidia
  - *Isospora belli* and cyclospora
- Viruses
  - cytomegalovirus
  - adenovirus
  - (HIV)

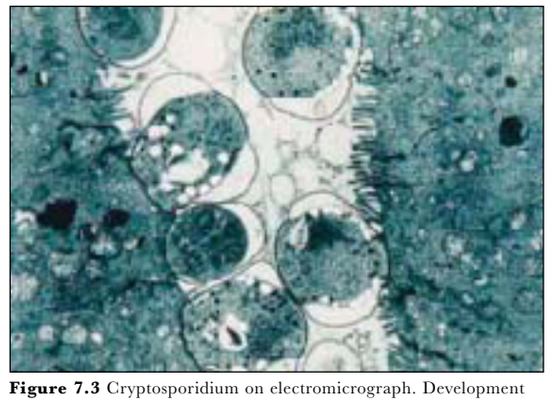


Figure 7.3 Cryptosporidium on electromicrograph. Development stages of cryptosporidium on the surface of enterocytes (note microvilli). The cryptosporidia are surrounded by a parasitophorous vacuole, the outer layers of which are derived from host cell outer membranes

Box 7.3 Treatment of HIV-associated diarrhoea

- Specific
  - antibiotics
  - antivirals
- Fluid replacement
- Antidiarrhoeal agents
  - loperamide
  - diphenoxylate
  - codeine
- Slow release morphine
- Subcutaneous diamorphine

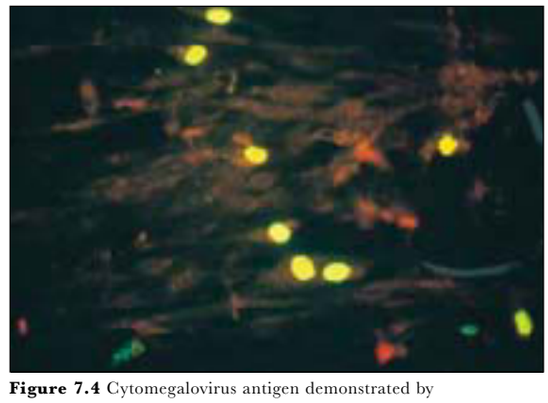


Figure 7.4 Cytomegalovirus antigen demonstrated by immunofluorescence microscopy after culturing human fibroblasts with homogenised intestinal tissue

Hepatitis and cholestasis

Abnormal liver biochemistry and/or hepatomegaly are common clinical problems although frank jaundice is uncommon. With the multiple therapies being used in treatment and prophylaxis, a drug-induced hepatitis must always be considered in a patient with AIDS and abnormal liver function tests. The differential diagnosis is wide and may involve the use of serology, abdominal ultrasound, ERCP, and liver biopsy. These latter two diagnostic procedures are clearly invasive and would not be indicated unless treatment of opportunistic infection, malignancy or biliary strictures was contemplated. In the absence of dilatated bile ducts on ultrasound, liver biopsy usually shows a granulomatous hepatitis caused by atypical mycobacteria.

AIDS sclerosing cholangitis presents with right upper quadrant pain, accompanied by a raised alkaline phosphatase. Abdominal ultrasound is abnormal in the majority of patients with biliary tract dilatation. ERCP may demonstrate papillary stenosis, dilatation of the common bile duct and dilatations and strictures with "beading" of the intrahepatic ducts. The disease is commonly associated with cryptosporidiosis, microsporidiosis, or cytomegalovirus infection. Endoscopic sphincterotomy may give pain relief in a proportion of patients with papillary stenosis. Liver function tests do not usually improve, and as it is a late-stage manifestation, the prognosis is poor, with most patients dying from some other HIV-related complication within six months of diagnosis.

HIV infection may alter the natural history of hepatitis B infection in a number of ways. The response rate to hepatitis B vaccination is lower in HIV-infected recipients. Immunodeficiency may favour the establishment of chronic infection following acute infection and HBV replication is increased with a reduction in the rate of spontaneous loss of HBe antigen. Interferon therapy would appear to be less effective in chronic HBV/HIV dual infection. The immune restoration following the initiation of antiretroviral therapy may lead to a hepatitis "flare" in chronic HBV carriers.

Hepatitis C virus infection is found primarily in intravenous drug users, although it may also be sexually transmitted. HIV can modify the natural history of HCV infection and patients with HIV/HCV dual infection tend to have more aggressive liver disease.

Figure 7.5 ERCP of AIDS sclerosing cholangitis with intrahepatic biliary tract distortion and dilatation of the common bile duct

Box 7.4 Differential diagnosis of liver disease

- Hepatitis or cholestasis
  - *M. avium-intracellulare complex*
  - Drug-induced
  - Viral hepatitis
  - Cytomegalovirus
  - *Mycobacterium tuberculosis*
  - *Cryptococcus*
  - Microsporidia
  - Lymphoma
  - Kaposi's sarcoma
- Biliary disease
  - *Cryptosporidium*
  - Cytomegalovirus
  - Microsporidia
  - Lymphoma
  - Kaposi's sarcoma

Anorectal disease

Perianal discomfort is often caused by recurrent herpes simplex infection. The diagnosis should be confirmed by viral culture. Patient-initiated intermittent aciclovir can give adequate symptom control in some cases but many patients will require long-term maintenance therapy. Resistance to both aciclovir and ganciclovir has been reported. Foscarnet is then the treatment of choice.

Anal warts are common but rarely cause much in the way of symptoms and should be treated on merit given the absence of any effective antiviral therapy. Anal intraepithelial neoplasia has been described in association with human papillomavirus infection but reports of invasive malignancy are still infrequent.

Patients may present with a mucopurulent proctitis, possible causes of which include recently acquired or long-standing *Neisseria gonorrhoeae* or *Chlamydia trachomatis* infection.

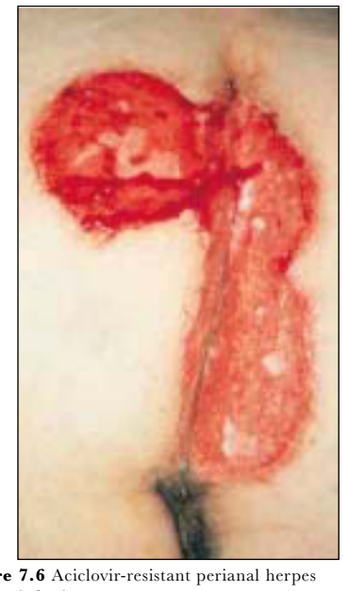


Figure 7.6 Aciclovir-resistant perianal herpes simplex infection

Neoplasia

Kaposi's sarcoma (KS) is commonly seen in the gastrointestinal tract and occurs in homosexual men more frequently than in patients from other risk groups. A new human herpes virus (HHV8) or Kaposi's sarcoma-associated herpes virus (KSHV) has been recently identified as a likely aetiological agent. KS lesions in the gut have the range seen in the skin, from small telangiectatic lesions, not well shown on contrast studies and only seen at endoscopy, to larger nodular or polypoid lesions. Complications from gastrointestinal disease are unusual, but include ulceration, obstruction, haemorrhage, and diarrhoea.

Lymphoma is much less common than KS however, although the incidence of KS has decreased along with the incidence of life-threatening opportunistic infections in association with the introduction of highly active antiretroviral therapy. The incidence of lymphoma has not been affected. HIV-associated lymphomas are usually high grade non-Hodgkin's type, of B-cell origin. Extranodal involvement is typical and the gut is one of the commonest sites involved.

We thank Dr Wilfred Weinstein, UCLA Medical School, Los Angeles for providing the photograph of oesophageal candidiasis and Dr David Casemore, PHLS Glan Clwyd, North Wales for the electronmicrograph of cryptosporidium.

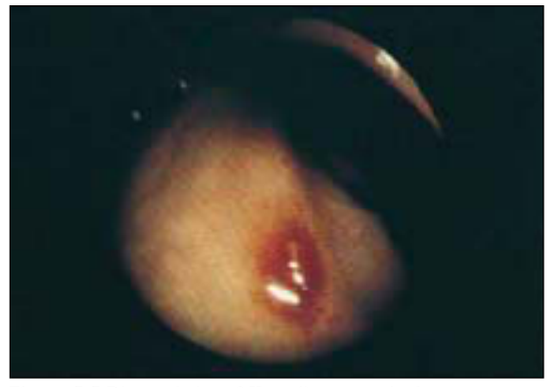


Figure 7.7 Discrete lesion of Kaposi's sarcoma in the rectum

8 Neurological manifestations

Hadi Manji

In patients infected with HIV, the whole neuraxis is vulnerable to damage. Up to 10% of patients may present with a neurological disorder at seroconversion (Box 8.1). The aseptic meningoencephalitis, which is usually self limiting, presents with headache, meningism, cranial nerve palsies and seizures. An acute demyelinating polyradiculoneuropathy (Guillain–Barré syndrome) is identical to that found in non-HIV-infected individuals, clinically and in the response to treatment with intravenous immunoglobulin or plasmapharesis. However, the cerebrospinal fluid shows a pleocytosis of over 20 cells/mm^3 which is unusual in non HIV cases. A high index of suspicion is required and HIV should be considered in all such cases.

During the asymptomatic phase of the illness, which may be of variable duration, headache and cranial nerve palsies (especially VIIth nerve – Bell's palsy) may be the only manifestation of a low-grade chronic meningitis.

The opportunistic infections and tumours as well as the complications ascribed to HIV itself usually develops when the CD4 count drops below 200/mm^3 (Box 8.2). Since the introduction of HAART, there has been a significant reduction in the incidence of infections such as toxoplasmosis and CMV.

Box 8.1 Seroconversion neurological presentations

- Encephalitis
- Aseptic meningitis
- Myelitis
- Cauda equina syndrome
- Acute demyelinating neuropathy (Guillain–Barré syndrome)
- Myositis

Box 8.2 Neurological complications in HIV infection

Opportunistic infections
- *Toxoplasma gondii* – abcesses and encephalitis
- *Cryptococcus neoformans* – meningitis
- JC virus – leucoencephalopathy (PML)
- CMV – retinitis, encephalitis, cauda equina syndrome, mononeuritis multiplex

Tumours
- Primary CNS lymphoma

HIV-related disorders
- HIV-associated dementia complex
- Vacuolar myelopathy
- Peripheral neuropathy (distal sensory polyneuropathy)
- Polymyositis

Clinical approach

The CD4 count is a useful guide to the aetiology of a neurological presentation – toxoplasmosis and cryptococcal meningitis occur at CD4 counts below 200/mm^3 whereas CMV complications occur below 50/mm^3. Since HIV infection itself results in CSF abnormalities such as a raised white cell count and an elevated protein level, more specific tests are required to diagnose encephalitic and meningitic illnesses. These include the measurement of cryptococcal antigen levels in cases of meningitis due to *C. neoformans* and CSF–VDRL and TPHA if syphilis is a differential. The inflammatory response is impaired and patients with meningitis may present with only mild symptoms of headache and no neck stiffness or photophobia. The threshold for investigating with CT/MRI and lumbar puncture is necessarily low. The measurement of serum antibodies to diagnose, for example toxoplasmosis, is unhelpful since the usual rise in levels of IgM does not occur. Infection with more than one organism occurs not infrequently, for example *Cryptococcus neoformans* and *Mycobacterium tuberculosis* and needs to be considered in cases of non-response or deterioration.

Box 8.3 Clinical guidelines

- CD4 useful guide to aetiology
- Persistent CSF abnormalities due to HIV
- Reduced inflammatory response
- Impaired antibody response
- Multiple simultaneous infections
- Maintenance treatment required

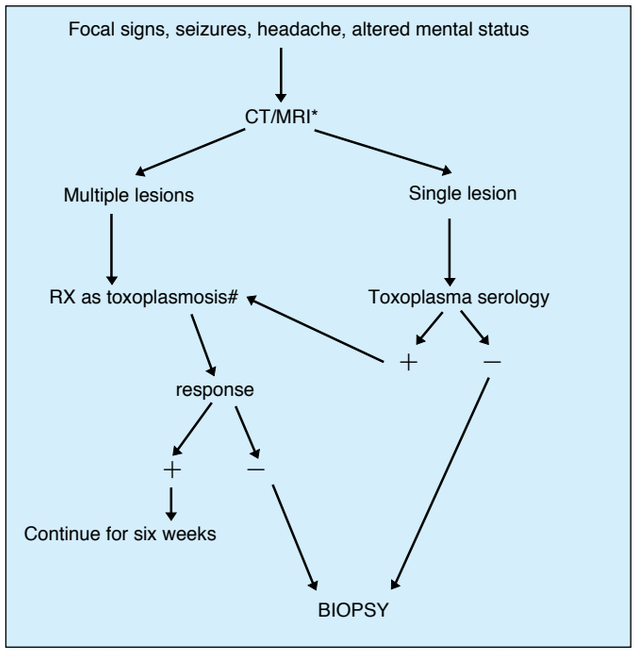


Figure 8.1 Management of mass lesions in HIV infection. * MRI is preferred mode of imaging. # If significant mass effect treat with reducing course of dexamethasone in addition to toxoplasma therapy

Opportunistic infections

Toxoplasma gondii

Toxoplasmsosis in HIV infection is usually a reactivation of latent infection in individuals who have been exposed previously to the organism. The clinical presentation is with headache with rapidly evolving focal neurological deficits over one to two weeks which include hemiparesis, dysphasia, visual field deficits, movement disorders (chorea/athetosis, parkinsonism) and seizures. Rarely, toxoplasmosis may affect the spinal cord and present with a myelopathy or a cauda equina syndrome. Blood serology for *T. gondii* is only helpful if negative since this makes the diagnosis less likely. Patients should have their toxoplasma serology documented at the first diagnosis of HIV infection. The risk of developing toxoplasma encephalitis in IgG seropositive patients is between 12% and 30%. These patients should be offered primary prophylaxis with co-trimoxazole at CD4 counts below 200.

CT/MRI shows multiple enhancing lesions with mass effect in the region of the basal ganglia and at the grey/white interface. A response to treatment is seen in 85% by day 7 and in over 90% by day 14. Repeat imaging should be performed after two weeks even if there is clinical improvement in cases of mixed pathology.

In patients with significant mass effect and cerebral oedema who are in danger of coning, additional treatment with dexamethasone will be necessary. A deterioration after this has been tailed off makes it necessary to consider a biopsy.

Cryptococcus neoformans

C. neoformans is a ubiquitous organism acquired by inhalation. Patients with meningitis may present acutely or insidiously over days or weeks with a headache, general malaise, confusion or seizures. The classical signs of meningism – neck stiffness, photophobia and Kernig's sign – are frequently absent.

Brain imaging is usually normal but MRI may reveal small abcesses – cryptococomas. The CSF cell count and protein may be normal and the diagnosis is confirmed by the presence of cryptococcal antigen in the CSF in 95% of cases. India ink staining is positive in 75%. 85% of cases are culture positive – the gold standard. Measurement of the serum cryptococccal antigen is a useful screening tool in patients presenting with headache or fever but should not be considered definitive.

Intracranial hypertension in the absence of mass lesions or hydrocephalus is an important cause of mortality and visual failure in approximately 20%. This is managed by repeated lumbar punctures or by the insertion of a lumbar or ventricular drain.

JC virus

Progressive multifocal leucoencephalopathy (PML) results from reactivation of the JC virus in immunosuppressed individuals. 80% of the general population will have been exposed to this virus as a banal childhood upper respiratory infection and have positive serology.

The presentation is with slowly evolving focal neurological deficits such as a hemiparesis, visual field and language problems and incoordination due to cerebellar involvement. Occasionally patients develop a dementia in association with these focal abnormalities. Symptoms and signs of raised intracranial pressure are absent although headache may be a feature.

Blood serological testing is unhelpful. Cranial CT shows non-enhancing areas of low attenuation in the white matter. MRI shows characteristic scalloping abnormalities at the grey/white interface with no mass effect or enhancement. The diagnosis may be confirmed by isolating JC virus by polymerase

Box 8.4 Focal lesions in AIDS

- Toxoplasmosis
- Primary CNS lymphoma
- Tuberculoma
- PML

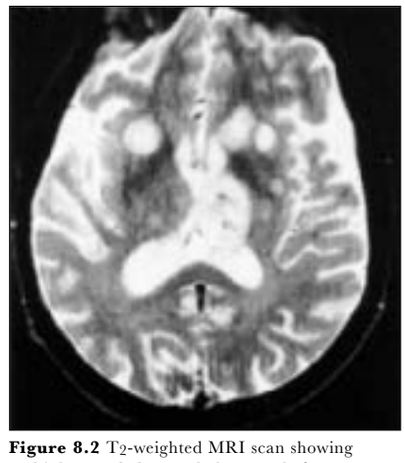


Figure 8.2 T_2-weighted MRI scan showing multiple rounded or oval abscesses before treatment in cerebral toxoplasmosis

Box 8.5 Meningitis in HIV infection

Fungal
- Cryptococcus neoformans

Bacterial
- Mycobacterium tuberculosis
- Listeria monocytogenes
- Streptococcus pneumoniae
- Treponema pallidum

Viral
- HIV
- Herpes simplex, herpes varicella zoster

Box 8.6 Poor prognostic features of AIDS-related cryptococcal meningitis

- Relapse episode
- CSF cryptococcal Ag titre > 1:10 000
- Positive India ink preparation
- Hyponatraemia
- Culture of extrameningeal cryptococcus

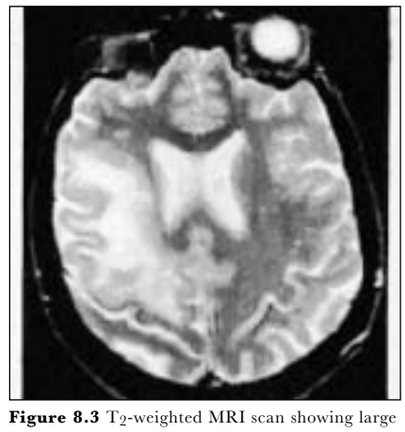


Figure 8.3 T_2-weighted MRI scan showing large area of high signal in one hemispheric white matter with no mass effect. Biopsy proved progressive multifocal leukoencephalopathy

chain reaction (PCR) techniques in the CSF in 75% of cases. If this is negative, a brain biopsy may need to be performed. This typically shows areas of focal demyelination, bizzare enlarged astrocytes and abnormal oligodendrocytes with inclusions which stain for JC viral antigens.

There is at present no specific treatment for PML. Cytosine arabinoside has been shown to be ineffective but trials are underway looking at the efficacy of drugs such as cidofovir (an anti CMV drug) and alpha interferon. Improvement in immune function with HAART has resulted in significantly better survival times.

Box 8.7 Clinical signs and symtoms in PML

- Motor function abnormalities (including hemiparesis)
- Mental status charges
- VIIth cranial nerve palsy
- Cerebellar syndrome
- Language disorders (dysphasia)
- Visual problems (for example hemianopia)
- Seizures

Cytomegalovirus

Over 90% of HIV-infected individuals have serological evidence of CMV infection. The neurological complications, which occur at CD4 counts below 50/mm^3, include retinitis, a cauda equina syndrome, an encephalitis and a mononeuritis multiplex. Apart from retinitis, the other complications occur infrequently.

CMV retinitis

The initial presentation of CMV retinitis depends upon the location – patients may be asymptomatic, complain of floaters, lose peripheral vision or if the lesions are centred around the macula, have poor visual acuity. Patients will often have evidence of CMV disease elsewhere such as colitis and such patients need to be screened for retinitis regularly.

On fundoscopy, there is a perivascular yellow-white infiltrate with retinal haemorrages. The differential diagnosis includes retinal complications of toxoplasmosis, lymphoma, syphilis, herpes zoster and herpes simplex.

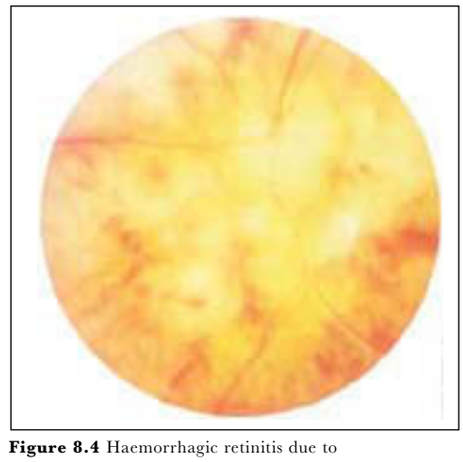


Figure 8.4 Haemorrhagic retinitis due to cytomegalovirus

CMV polyradiculopathy

This well-recognised syndrome presents over a period of days with back pain followed by the development of a progressive flaccid weakness of the legs with sensory loss and sphincter disturbance. Imaging studies which are essential to exclude compressive lesions due to, for example, lymphoma are normal or may show thickened nerve roots. The CSF shows a characteristic neutrophil pleocytosis which is unusual in a viral infection. Without treatment there is a progression of the neurological deficits, with death in 2 or 3 months.

CMV encephalitis

Although evidence of CMV infection is often found in the brains of patients dying from AIDS, the clinical correlates are unclear. A CMV encephalitis needs to be considered in patients presenting with a rapidly progressive encephalitis with cranial nerve palsies and seizures. CMV may be isolated from the CSF using PCR.

Primary CNS lymphoma (PCNSL)

PCNSL is the most common cause of mass lesions in children and the second most common in adults after toxoplasmosis. Histologically, this is a high-grade B cell lymphoma. The Ebstein–Barr virus can be isolated from tissue specimens and is believed to have a causal role in the development of the lymphoma.

The clinical presentation is similar to that of toxoplasmosis with focal neurological deficits such as hemiparesis and seizures. There are usually signs and symptoms of raised intracranial pressure with increasing headache, vomiting and papilloedema.

Although the isolation of EBV by PCR in the CSF is specific, most patients present with mass lesions and raised intracranial pressure. Lumbar puncture is therefore contraindicated. The CT and MRI findings may be

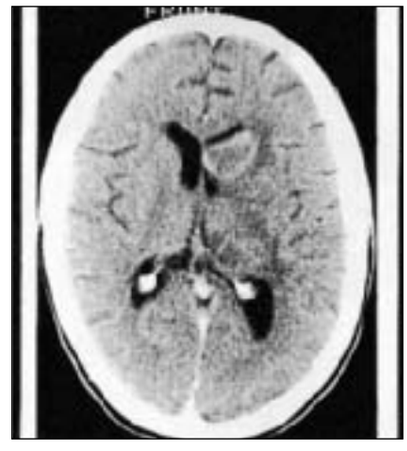


Figure 8.5 Enhancing right frontal mass lesion due to lymphoma

indistinguishable from those due to toxoplasmosis with multiple enhancing lesions with associated cerebral oedema and mass effect. However, a single lesion on MRI especially if the toxoplasma serology is negative, is more likely to be lymphoma, as are lesions which closely adhere to the ventricular walls.

The diagnosis of PCNSL is usually made by biopsy.This may be performed after failure of treatment with antitoxoplasma therapy for at least two weeks. However, since prognosis is poor even with whole brain radiotherapy, it is reasonable not to proceed with a biopsy unless there is a suspicion that other more treatable pathogies may be identified.

HIV-associated dementia complex (AIDS dementia complex) or HIV dementia

This complication occurs in 15% of AIDS patients usually in patients with a CD4 count below 200/mm^3. The early reports of evidence of cognitive abnormalities in HIV positive asymptomatic individuals have been discounted by large cohort studies using clinical, neuropsychological, MRI and neurophysiological methods of assessment. The clinical picture is of a variably progressive dementia with psychomotor slowing and impairment of memory.

The diagnosis is one of exclusion of other infective or neoplastic aetiologies by brain imaging and CSF examination. Although there is a correlation between the CSF HIV RNA viral load and the severity of dementia, there is too much overlap for use as a diagnostic test. A neuropsychological assessment is also helpful. MRI shows cortical atrophy and diffuse or patchy white matter high signal on T2-weighted images.

The underlying pathophysiological mechanisms are unclear but HIV is usually isolated from the microglial cells and astrocytes rather than neuronal cells. Productive infection of the macrophages and microglia with the release of cytokines such as TNF results in neuronal damage.

Since the introduction of zidovudine and subsequently HAART the incidence of HIV-associated dementia has progressively declined. However, more recently there is concern that the CNS may become a sanctuary for HIV, since most of the newer drugs penetrate the CNS poorly.

Peripheral nerve disorders in HIV infection

DSPN is the commonest neurological complication encountered in HIV patients, with 30% of AIDS individuals experiencing symptoms. It is unusual in the asymptomatic stages of HIV infection. Pathologically, this is a length-dependent axonal neuropathy usually sparing the hands. The symptoms and signs are typical of a small fibre neuropathy. Treatment is symptomatic using antidepressant and anticonvulsant drugs.

The neuropathy due to the nucleoside analogue drugs (ddl, ddC and D4T) is similar and therefore difficult to differentiate from DSPN. These drug related neuropathies are dose dependent and reversible. However, patients may continue to deteriorate for 6–8 weeks after stopping the drug – "coasting".

Box 8.10 Symptoms and signs of DSPN

- Numb, burning feet
- Pins and needles
- Contact hypersensitivity
- Little or no weakness
- Impaired pain and temperature sensation
- Depressed or absent ankle jerks

Box 8.8 Symptoms and signs of HIV dementia

Early
- Poor concentration
- Forgetfulness
- Clumsiness
- Unsteady gait
- Apathy
- Impaired eye movements
- Brisk reflexes
- Slowed fine finger movements

Late
- Global dementia
- Incontinent of urine and faeces
- Seizures
- Spastic paraparesis (due to vacuolar myelopathy)
- Myoclonus

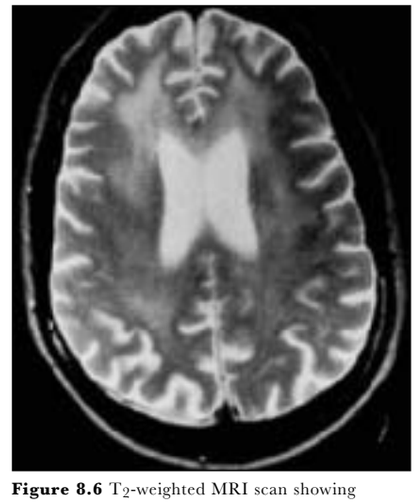


Figure 8.6 T_2-weighted MRI scan showing "milky" hyperintensity of the hemispheric white matter due to HIV dementia

Box 8.9 Peripheral nerve disorders in HIV infection

HIV related
- Axonal neuropathy (distal sensory peripheral neuropathy, DSPN)
- Demyelinating neuropathy – acute (Guillain-Barré syndrome), chronic (CIDP)
- Vasculitic neuropathy (mononeuritis multiplex)
- Diffuse infiltrative lymphocytic syndrome (DILS)

CMV related
- Vasculitis (mononeuritis multiplex)
- Lumbosacral polyradiculopathy

Toxic
- ddl, ddC, D4T
- isoniazid
- thalidomide
- dapsone

Box 8.11 Investigations in HIV neuropathy

- Neurotoxic drugs, including excess vitamin B_6
- Excess alcohol
- Blood tests: vitamin B_{12}, glucose, VDRL, vitamin E (if severe diahorrea)
- Nerve conduction tests – only if marked weakness or unusual presentation
- Nerve biopsy may be indicated to exclude an inflammatory neuropathy (vasculitis or demyelination)

9 Treatment of infections and antiviral therapy

Ian VD Weller, IG Williams

The treatment of HIV infection can be largely divided into: (i) specific antiviral agents that inhibit viral replication, (ii) measures that either treat or prevent (prophylaxis) its complications – namely opportunistic infections and tumours. Major advances in the treatment of HIV infection have occurred in the last few years. This has resulted in marked falls in the reported number of new AIDS cases and deaths in the developed world since 1996. Effective antiretroviral therapy regimens which substantially inhibit HIV replication and allow sustained improvements in the immune system are the main reason for this. There are currently three classes of antiretroviral agents: the nucleoside and non-nucleoside reverse transcriptase inhibitors and the protease inhibitors. Improved formulations and new drugs are continuously being evaluated and there is increasing interest in the possible role of immunotherapy combined with antiretroviral therapy to improve specific immune responses.

However, in those who are severely immunosuppressed the treatment and prophylaxis of opportunistic infections remains important. Though it cannot be overemphasised that the most effective way to prevent first episodes or recurrences of opportunistic infections is treatment with antiretroviral drugs. This chapter will cover both antiretroviral therapy and the treatments of the infections previously described in other parts of this book, in an attempt to bring all of these together in a comprehensive manner.

Box 9.1 Treatment strategies in HIV disease

- Antiretroviral therapy: suppresses viral replication
  results in immune reconstitution
- Prophylaxis of opportunistic infections
- Prevent exposure to opportunistic pathogens

Protozoal infections

***Pneumocystis carinii* pneumonia (PCP)**

Although recently recognised as being more like a fungus, *P. carinii* is considered under protozoa here. Nowadays PCP most commonly occurs in those at risk who fail to take adequate prophylaxis or who are newly diagnosed with HIV infection in advanced disease where it is frequently the presenting illness.

Clinical suspicion is aroused early in patients who are under regular medical supervision, leading to earlier diagnosis. Later diagnosis is asssociated with more severe disease and poorer treatment outcome. Techniques of diagnosis include sputum induction with nebulised saline; this obviates the need for bronchoscopy but the diagnostic sensitivity is lower. The use of lavage alone at bronchoscopy avoids transbronchial biopsy with its complications of haemorrhage and pneumothorax. Exercise oximetry and alternative imaging techniques with radiolabelled compounds are also being used in diagnosis. Monoclonal antibodies to pneumocystis proteins and sensitive DNA probes have been developed but have yet to reach the bedside. In the absence of a confirmatory test, a presumptive diagnosis may be made based on the clinical presentation and chest *x* ray appearances in a patient severely immunosuppressed and at risk.

High-dose intravenous co-trimoxazole for two to three weeks remains a standard first-choice regimen for severe PCP, but once fevers and symptoms have settled and blood gas values have improved the drug can be given by mouth. Side-effects are common, typically after 7–10 days. If co-trimoxazole treatment is not tolerated, alternative treatment regimens include either intravenous pentamidine or a combination of clindamycin and primaquine. Pentamidine is as effective as co-trimoxazole but has side-effects that can be life threatening and should be given

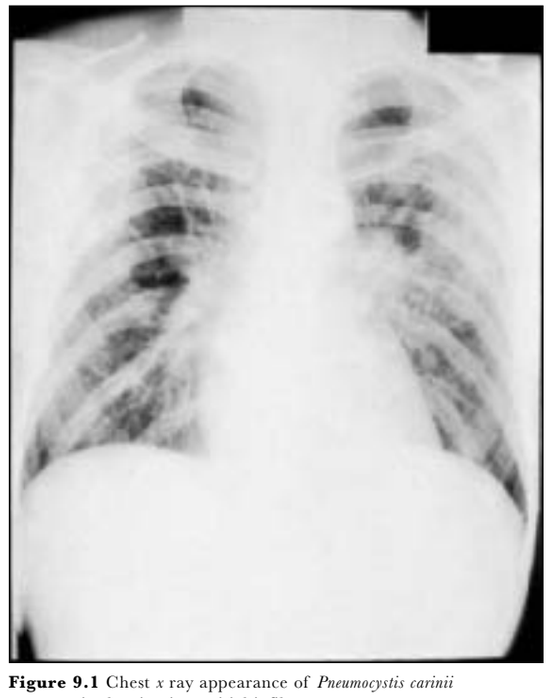


Figure 9.1 Chest *x* ray appearance of *Pneumocystis carinii* pneumonia showing interstitial infiltrates

by slow intravenous infusion with careful monitoring. In patients with moderate or mild PCP a combination of clindamycin and primaquine has proven clinical efficacy and is an alternative first choice for those patients who have a previous history of severe co-trimoxazole hypersensitivity. Side-effects of rash and diarrhoea are frequent.

In patients presenting with severe hypoxaemia high-dose adjunctive corticosteroid therapy is indicated and has been shown in clinical studies to reduce both mortality and morbidity

Alternative second-line therapies include dapsone with trimethoprim, trimetrexate with folinic acid or Atovaquone, a hydroxy-naphthoquinone. The efficacy of atovaquone has only been established in mild to moderate *P. carinii* infection. Like trimetrexate it is probably less effective than co-trimoxazole but it is less toxic. New formulations have improved atovaquone's bioavailability but it still should not be given to patients with malabsorbtion conditions, previous severe diarrhoea or those not taking oral nutrition. Due to acquired resistance, where possible atovaquone should not be given as single-agent therapy. It is commonly combined with intravenous pentamidine as an effective second-line treatment.

Prophylaxis for PCP pneumonia is essential after a first attack (secondary prophylaxis) but is also recommended for all patients once their CD4 cell counts falls below 200×10^6/l (primary prophylaxis). The risk of a first episode PCP below this CD4 count level in patients not on antiretroviral therapy is estimated to be 18% at 12 months for those who are asymptomatic, rising to 44% for those who have early symptomatic disease (for example, oral candida, fever). Co-trimoxazole 960 mg given by mouth daily or three times per week is the most effective agent. In patients who are intolerant, alternative regimens include oral dapsone 100 mg with pyrimethamine 25 mg daily or three times per week, atovaquone 1500 mg daily or nebulised pentamidine. Dose of the latter depends on the nebuliser system: with a Respirgard II nebuliser the recommended regimen is 300 mg every four weeks. In patients with more advanced disease and CD4 counts less than 100×10^6/l, 300 mg given every two weeks should be considered in view of the high failure rate of the monthly regimen.

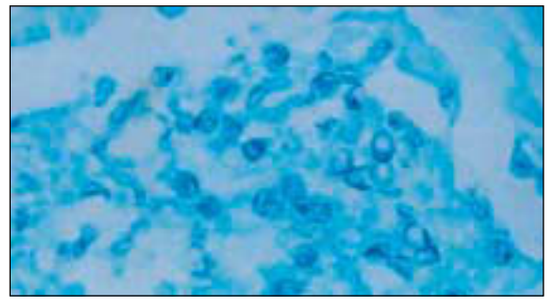


Figure 9.2 Cysts of *Pneumocystis carinii* in broncho lavage specimen

Table 9.1 *Pneumocystis carinii* pneumonia treatment

Drug	Duration	Side-effects	Comments
First choice:			
Co-trimoxazole (trimethoprim component 15–20 mg/kg per day p.o./i.v. in divided doses).	21 days	Nausea, vomiting, fever, rash, marrow suppression, raised transaminases	Intolerance common (25–50% of treated patients)
Alternative regimens:			
1. Severe disease: Pentamidine isethionate 4 mg/kg per day as slow intravenous infusion	21 days	Hypotension, hyper- and hypoglycaemia, renal failure, marrow suppression, nausea, vomiting, cardiac arrest	80% of patients will respond to treatment
Trimetrexate 45 mg/m^2 i.v. and folinic acid 80 mg/m^2	21 days	Marrow suppression, raised transaminases rash, anaphylaxis	Should only be used as third or fourth line treatment
2. Mild to moderate disease: Clindamycin 600 mg 6 hourly p.o./i.v. and primaquine 15 mg daily p.o.	21 days	Diarrhoea, rash, nausea, vomiting, marrow suppression, methaemoglobinaemia, haemolysis	*Clostridium difficile* toxin associated diarrhoea is a frequent complication of clindamycin therapy
Trimethoprim 20 mg per kg/day p.o./i.v. in 2–3 divided doses and dapsone 100 mg daily p.o.	21 days	Rash, nausea, methaemoglobinaemia, marrow suppression	Alternative regimens should be used in patients with G6PD deficiency
Atovaquone suspension 750 mg twice daily	21 days	Rash, raised transaminases and neutropenia	Must be taken with food. Consider combination with i.v. pantamidine as resistance reported with monotherapy
Adjuvant high-dose steroids (for example, prednisolone 40–60 mg daily p.o.)	5 days tapering over 14–21 days		Indicated in severe disease. Optimal dose not determined

Although clinical trials have shown greater efficacy for co-trimoxazole compared to other regimens, there is a high rate of discontinuation due to side-effects. Desensitisation regimens are used with the aim of reducing the rate of intolerance but there is uncertainty about their efficacy and which regimen is best.

In patients responding to antiretroviral therapy, primary or secondary prophylaxis can be safely discontinued once the CD4 count has increased to levels persistently above 200×10^6/l.

Toxoplasmosis

Cerebral toxoplasmosis is the commonest manifestation of toxoplasma infection. As toxoplasmosis is the most common cause of ring-enhancing lesions on contrast CT brain scans a presumptive diagnosis is usually made and treatment started. The condition responds well if treatment is started early, and a combination of sulphadiazine 4–6 g/day and pyrimethamine 50–100 mg a day (both by mouth in divided doses with folinic acid 15 mg daily) is the treatment of choice. Side-effects may prevent continued use of sulphadiazine, and clindamycin 600–1200 mg four times a day has been shown to be an effective alternative in controlled studies.

Corticosteroids are sometimes used in addition to first-line treatment to reduce symptomatic cerebral oedema, but a clinical and radiological response seen after two weeks of treatment may be due solely to the corticosteroid effect rather than the anti-toxoplasma treatment. A presumptive diagnosis of toxoplasma may therefore be made, although the underlying lesion may be due to something else, such as lymphoma or another infection. Relapse is common after treatment is stopped, and maintenance treatment is therefore necessary. In patients responding to antiretroviral therapy with sustained increases in CD4 count, discontinuation of prophylaxis is safe but there is limited current data to make definite recommendations.

Atovaquone 750 mg four times a day with or without pyrimethamine may be considered an alternative and the new macrolides clarithromycin 2 g daily and azithromycin, both given with pyrimethamine 75 mg/day, have also been effective in small uncontrolled studies. The most appropriate regimen for secondary prophylaxis has not been determined but treatment doses of either sulphadiazine and pyrimethamine or clindamycin and pyrimethamine are usually halved.

Of patients with positive toxoplasma serology and a CD4 count of less than 100×10^6/l, approximately 1 in 3 will develop cerebral toxoplasmosis within 12 month without prophylaxis. Primary prophylaxis in patients with positive serology with a CD4 count of less than 100×10^6/l is therefore recommended. Co-trimoxazole or dapsone with pyrimethamine have been shown to reduce the incidence of toxoplasmosis compared to patients taking nebulised pentamidine for prophylaxis against PCP. Atovaquone with or without pryrimethamine may also be considered but this is based on more limited data. The macrolides clarithromycin and azithromycin might be anticipated to provide broad-spectrum prophylaxis for toxoplasmosis, atypical mycobacterial and bacterial infections, but bacterial resistance might limit their use in this situation.

Patients who are toxoplasma serology negative should be given advice to prevent exposure in primary infection with toxoplasmosis. They should be advised not to eat raw or undercooked meat and avoid directly handling cats' faeces.

Cryptosporidiosis and other protozoa

In patients with less advanced HIV disease (CD4 counts >200 $\times 10^6$/l) cryptosporridial infection usually causes a self-limiting gastrointestinal illness and symptomatic treatment with

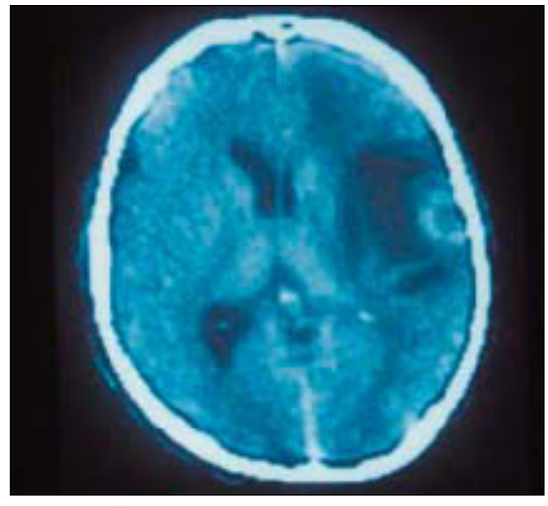


Figure 9.3 CT scan showing ring-enhancing lesions of cerebral toxoplasmosis surrounded by cerebral oedema (dark area)

Box 9.2 Treatment of toxoplasmosis

First line

Sulphadiazine 4–6 g per day or clindamycin 600–1200 mg × 4 per day

+

Pyrimethamine 50–100 mg per day

+

Folinic acid 15 mg per day

Alternatives

- Clarithromycin 2 g per day or
- Atovaquone 750 mgs 4 × per day p.o.
  +
  Pyrimethamine 50–100 mg per day p.o.

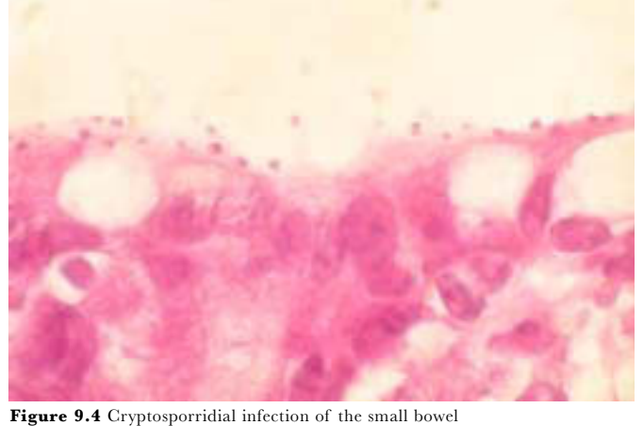


Figure 9.4 Cryptosporridial infection of the small bowel

anti-diarrhoeal agents is all that maybe needed. In those with more severe immunosuppression and persistent symptoms treatment is more difficult and reported successes with a variety of agents are still anecdotal. Symptoms and excretion of cysts may be intermittent. Responses have been described after treatment with a variety of agents, including spiramycin, erythromycin, diclazuril, letrazuril, hyperimmune bovine colostrum, paromamycin, azithromycin and subcutaneous somatostatin.

Symptomatic treatment with antidiarrhoeal and antiemetic agents together with fluid, electrolyte and nutritional support should be provided. Case reports suggest that immune reconstitution is likely to result in improvement and resolution of both symptoms and infection. Thus in the absence of an effective specific treatment against cryptosporidium, infected patients should be started on antiretroviral therapy to increase the CD4 count.

Patients at risk of infection should be advised to avoid possible exposure in water supplies particularly at times of documented outbreaks. Although unproven, measures that may be considered for patients with CD4 counts less than 200×10^6/l include using bottled water, point of use filters or boiling water for more than one minute.

For microsporidiosis there have been anecdotal reports of symptomatic improvement with albendazole 400 mg twice a day or metronidazole 500 mg three times a day.

Isosporiasis is less common and appears to respond to co-trimoxazole 960 mg four times a day, but relapses occur in half of all cases.

Diarrhoea often occurs in the absence of recognised pathogens in the stool, and metronidazole has relieved symptoms in some cases.

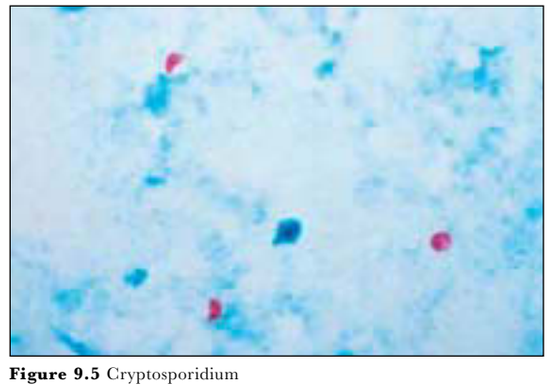


Figure 9.5 Cryptosporidium

Table 9.2 Viral opportunistic infections

Infection	Drug	Duration	Side-effects	Comments
Herpes simplex				
Treatment	Aciclovir 200 mg 5 × a day orally or 10 mg/kg 8 hourly i.v.	5–7 days		Duration may be extended in severe infections
Prophylaxis	Aciclovir 200 mg 4 × a day or 400 mg 2 x day	Indefinite		
Cytomegalovirus				
Treatment	Ganciclovir 5 mg/kg twice a day i.v.	14-21 days	Neutropenia, anaemia	GCSF support may be required
	Cidofovir 5 mg/kg i.v. once a week	2 weeks	Nephrotoxicity: impaired creatine clearance, proteinuria, hypophosphataemia Neutropenia Ocular toxicity	Co-administer with probenecid and adequate hydration to reduce risk of nephrotoxicity
	Foscarnet 180 mg/kg daily i.v.	14–21 days	Nephrotoxicity, hypomagnesaemia, hyper- and hypocalcaemia, hyper- and hypophosphataemia, hypokalaemia, nausea vomiting, genital ulceration	Dose must be adjusted according to renal function
Maintenance ganciclovir	Ganciclovir 3 gr daily orally	Until CD4 count > 100×10^6/l on HAART	As above	May be combined with intraoccular implants Avoid in patients with diarrhoea Increases levels of didanosine
	Cidofovir 5 mg/kg once every 2 weeks	Until CD4 count > 100×10^6/l on HAART	As above	As above
Alternative secondary prophylaxis regimens include daily intravenous foscarnet or ganciclovir, intravitreal injections of ganciclovir or foscarnet and intraoccular ganciclovir implants				

Viral infections

Severe mucocutaneous and systemic infections with herpes simplex virus are best treated with aciclovir. Prophylaxis is used after severe infection and in patients with increasing severity and frequency of recurrences. These recurrences can be a prelude to the chronic persistent mucocutaneous ulceration characteristic of AIDS.

Varicella zoster virus infections are usually treated with high-dose aciclovir given by mouth. However, dissemination of infection from dermatomal zoster is unusual even without treatment.

Valaciclovir is a pro-drug of aciclovir which is used in the treatment of herpes zoster and herpes simplex infections of skin and mucous membranes. Valaciclovir is a L-valine ester of aciclovir that is rapidly converted to aciclovir after oral administration. The antiviral spectrum and mode of action is therefore the same as aciclovir. Aciclovir has, however, a low oral bioavailability (about 15–20%). Valaciclovir has three or four times the oral bioavailability of aciclovir.

Famciclovir is a diacetyl ester of 6-deoxy penciclovir which has been used in the treatment of herpes zoster and genital herpes infections. Famciclovir is metabolised to penciclovir in the intestinal wall and liver. Penciclovir and aciclovir have similar antiviral spectrum.

Aciclovir resistant herpes simplex infections can occur, particularly in patients with advanced disease and severe immunosuppression. Alternative agents to treat resistant infections include foscarnet and cidofovir.

Reactivation of cytomegalovirus with viraemia and end-organ disease tends to occur when CD4 cell counts are persistently below 50×10^6/l. Ganciclovir (an acyclic analogue of deoxyguanosine), foscarnet phosphonoformate (a pyrophosphate analogue, which inhibits polymerase enzymes) and cidofovir (a nucleoside analogue with potent *in vitro* activity against viruses) are used for the treatment of cytomegalovirus retinopathy, gastrointestinal and neurological disease. Treatment arrests progression retinitis in most patients, and maintenance therapy is required in those patients who continue to be severely immunosuppressed to delay the time to further relapse. There is little comparative data to guide initial choice of treatment. A study comparing ganciclovir with foscarnet for treatment of CMV retinitis found no difference between the drugs in their ability to delay progression of disease, but there was a survival advantage in those patients treated with foscarnet. However, foscarnet is not as well tolerated as ganciclovir, as it produces reversible renal failure and electrolyte disturbances. Careful and frequent monitoring is required which complicates outpatient management. The major side-effect of ganciclovir is bone marrow suppression, particularly neuropenia. Support therapy with granulocyte colony simulating factor (GCSF) maybe required.

The detection of mutations in the CMV UL97 gene is associated with an increase in CMV DNA levels in blood and clinical progression of CMV retinitis during ganciclovir therapy. High-level resistance to ganciclovir results in cross-resistance to cidofovir. Resistance to foscarnet can occur but the mechanism is different.

Cidofovir has been shown to be effective however in delaying progression and time to relapse in patients who have experienced therapy failure on ganciclovir and foscarnet. The dosing schedule of cidovofir is convenient and more suitable to outpatient care than with either intravenous ganciclovir or foscarnet. It is given once weekly (5 mg/kg) for two weeks as induction therapy and then at the same dose every two weeks thereafter as maintenance therapy. The main side-effect is

Box 9.3 Management of CMV disease: key points

- Population at highest risk of clinical disease:
  CD4 < 50×10^6/l
  positive CMV viraemia
- Diagnostic criteria: combination or clinical presentation +/− histopathology +/− virus isolation (culture or antigen detection)
- Choice of first line therapy dependent upon renal function, haematological indices and risk of toxicity
- Where possible all patients should be started on an effective HAART regimen to increase the CD4 count to above 100×10^6/l
- Secondary prophylaxis maybe discontinued once the CD4 count has risen and remains above 100×10^6/l

Table 9.3 Opportunistic infections: recommendations for initiation of primary prophylaxis

Opportunistic infection	Recommendations
Pneumocystitis carinii pneumonia	CD4 count <200 × 10^6/l
Cerebral toxoplasmosis	CD4 count <100 × 10^6/l and positive lg G toxoplasma serology
Mycobacterium avium complex	CD4 count <50 × 10^6/l
CMV disease	under evaluation: may consider if CD4 <50 × 10^6/l and positive CMV viraemia
Tuberculosis	If recent close contact of smear positive index patient and no evidence of active clinical disease National Guidelines for use of Tuberculin skin testing for screening varies.

Figure 9.6 Penile ulceration caused by intravenous foscarnet therapy.

nephrotoxicity. The dose needs to be adjusted or treatment delayed or discontinued if there is evidence of renal tubular dysfunction, for example proteinuria, hypophosphataemia and impaired creatinine clearance.

The choice of initial treatment is therefore dependent on the preferred dosing schedule, the risk of drug-associated toxicity and previous anti-CMV treatment history. Alternative treatment strategies include combination regimens of foscarnet and ganciclovir, intravitreal injections of ganciclovir or foscarnet and intraoccular implants of ganciclovir. The latter effectively prevents relapse in the treated eye for up to three months but there is an increased risk of early retinal detachment. There is a risk of CMV disease occurring in the contralateral eye or elsewhere, and thus concomitant oral ganciclovir is indicated.

Following induction therapy, secondary prophylaxis is required but can be safely discontinued without risk of relapse of retinopathy in patients who have responded to highly active antiretroviral therapy (HAART). Improved cytotoxic T-lymphocyte responses to CMV and suppression of CMV viraemia is seen in those patients with advanced disease who sustain a rise in CD4 count on HAART. Effective antiretroviral therapy has resulted in dramatic falls in the incidence of new episodes of CMV disease and of relapse. However, in patients who remain severely immunosuppressed and at risk of CMV disease and relapse, secondary prophylaxis is required. Daily intravenous foscarnet or ganciclovir regimens require an in-dwelling intravenous catheter which is inconvenient and complicated by the risk of bacterial infections. Either daily oral ganciclovir or two-weekly intravenous cidofovir are preferable. Although ganciclovir is poorly absorbed, the oral preparation at a daily dose of 3 g has similar efficacy to intravenous regimens in preventing progression of retinitis. Combinations of ganciclovir with greater oral bioavailability are under evaluation.

Primary prophylaxis against CMV retinitis with oral ganciclovir has been investigated, but the results of two large clinical trials are conflicting, and in view of the high cost has not gained acceptance in routine clinical practice. Immune preservation or reconstitution as a result of HAART is the best prophylaxis (both primary and secondary) against CMV end-organ disease and other major opportunistic infections.

Table 9.4 Fungal opportunistic infections

Infection	Drug	Duration	Side-effects	Comments
Candidiasis				
Local treatment	Nystatin oral suspension or pastilles, miconazole oral gel or amphotercin lozenges all 4–6 times a day	As required		Systemic therapy is commonly required
Systemic treatment	Ketoconazole 200 mg a day (p.o.)	1–2 weeks	Nausea (less if taken with food), abnormal liver function tests, hepatitis thrombocytopenia, rash	In patients who remain severely immunosuppressed, relapse is common and maintenance therapy is required
	Fluconazole 50–200mg a day (p.o.)	1–2 weeks	Nausea, abnormal liver function tests	As above
	Itraconazole capsules or solution 200 mg/day (p.o.)	1–2 weeks	Nausea, abnormal liver function tests	
Cryptococcosis				
Treatment	Amphotericin B 0.7–1.0 mg/kg/day (i.v.) ± flucytosine 75–100mg/kg/day in 3–4 divided doses		Nausea, vomiting, rash, bone marrow suppression, renal impairment, hypocalcaemia	In patients who remain severely immunosuppressed, relapse is common and maintenance therapy is required Liposomal preparations of amphotericin reduces risk of nephrotoxicity
	or Fluconazole 800 mg daily 1–3 days 600 mg daily thereafter (p.o. or i.v.)		As above	

Fungal infections

Dermatophytic fungal infections respond well to imidazole creams. Oral candida is often asymptomatic in its early stages and may not require treatment. In more severe infections local treatment with frequent nystatin suspension, or pastilles, or amphotericin lozenges can be used. Systemic treatment with oral ketoconazole or fluconazole daily is required for more severe oropharyngeal and oesophageal candidiasis. Long-term maintenance treatment may be required to prevent recurrences, and liver function tests should be monitored. Clinical resistance to treatment can occur and in the case of fluconazole may be related to emerging candida species that are less sensitive to fluconazole or to *Candida albicans*-resistant strains. Intermittent therapy rather than maintenance may be a more appropriate strategy to reduce this risk but has yet to be assessed in a large controlled trial. Itraconazole solution has been found to be useful in cases of clinical resistance and this may be related to its topical action, better absorption and greater spectrum of activity.

Vulvovaginal candidiasis can be a recurrent problem in women and should be treated either with topical agents (clotrimazole or miconazole pessaries and cream) or single high dose fluconazole.

Cryptococcal meningitis is treated with either fluconazole or amphotericin B with or without flucytosine. A large comparative study has shown that the overall mortality was similar in both treatment groups. However, there were more early deaths in the fluconazole group, and amphotericin sterilised the cerebrospinal fluid more rapidly but fluconazole was better tolerated. There was a 20% mortality and the factors predictive of death were an abnormal mental state, a cryptococcal antigen titre above 1 024 and a white cell count below 0.02×10^9/l in the cerebrospinal fluid. Physicians will probably therefore prefer to treat patients with these poor prognostic markers with amphotericin rather than fluconazole. With a 20% mortality irrespective of what treatment is used it is clear that improvements in treatment are required.

Maintenance treatment is required in those who remain severely immunosuppressed, as replase is common. Fluconazole (200 mg/day) was more effective than amphotericin B (1 mg/kg/week) in a large randomised study. The comparative efficacy of higher doses of amphotericin maintenance treatment is unknown. Liposomal preparations of amphotericin B may be useful, particularly in patients at risk of renal toxicity. Controlled studies of high doses of fluconazole suggest greater efficacy. As with other severe opportunistic infections, immune reconstitution following HAART will allow safe discontinuation of secondary prophylaxis regimens.

Amphotericin B is still the mainstay of treatment of other systemic fungal infections. Itraconazole has shown to be effective in induction and maintenance treatment of disseminated histoplasmosis.

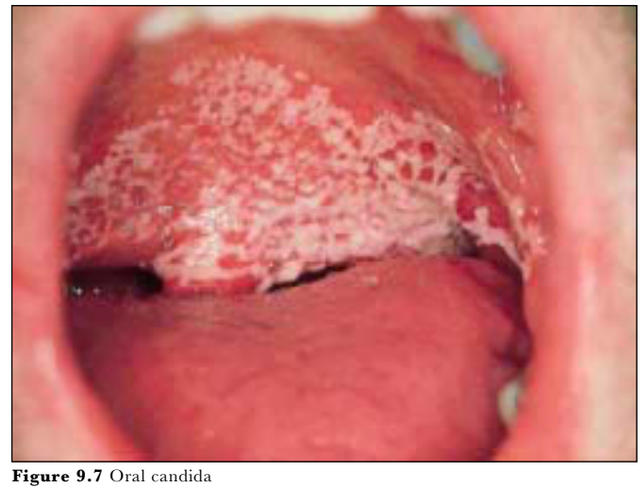


Figure 9.7 Oral candida

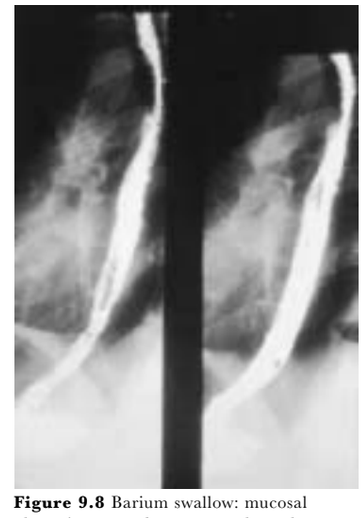


Figure 9.8 Barium swallow: mucosal ulceration secondary to oesophageal candida infection

Bacterial infections

Tuberculosis in HIV infection is treated in the standard way with isoniazid and rifampicin plus either pyrazinamide or ethambutol. Rifampicin is a potent enzyme inducer and increases the metabolism of drugs such as oral contraceptives, dapsone, fluconazole, ketoconazole and anticonvulsants. Clinicians should also be aware of drug interactions between rifamycins (rifampicin and rifabutin) and antiretroviral drugs, particularly the protease inhibitors (PIs) and the non-nucleoside reverse transcriptase inhibitors (NNRTIs). Certain combinations of each are contraindicated or require dose adjustment to

Box 9.4 Treatment of MAC

Clarithromycin 1 g–2 g daily in divided doses
+
Ethambutol 15 mg/kg/day daily
+
Either Rifabutin 450–600 mg daily
Rifampicin 450–600 mg daily
Ciprofloxacin 500 mg twice daily
Clofazimine 100 mg daily

3 or 4 drug regimens are recommended

maintain therapeutic levels. Knowledge of these potential interactions is essential to avoid loss of clinical efficacy or increased risk of drug toxicity.

Although extrapulmonary disease is more common in HIV seropositive patients than in uninfected controls, the responses to treatment appear similar in the developed world if patients are compliant. Over the last few years there have been several outbreaks of tuberculosis with multiple drug-resistance (MDR) in the USA and Europe including the UK. Transmission of drug-resistant strains has occurred between patients and from patients to family members, healthcare workers and prison guards. Mortality from drug-resistant tuberculosis in this setting is high, around 70–90%. To reduce the risk of MDR TB it is essential to ensure adherence to antituberculosis therapy by patients and for healthcare facilities to have in place procedures and facilities to reduce the risk of nosocomial transmission.

Disseminated infection with *Mycobacterium avium* complex (MAC) causes considerable morbidity and mortality in the later stages of HIV infection (when CD4 counts are persistently below 50×10^6/l). Various combinations of drugs have been shown to decrease mycobacteraemia and improve symptoms in uncontrolled studies. Four, three and two drug regimens have and are being assessed in clinical trials. A commonly used regimen in clinical practice is rifampicin or rifabutin (450–600 mg/day), ethambutol (15 mg/kg, max 1 g/day) and clarithromycin (500 mg twice a day). Other drugs that have been studied and may be considered include: clofazimine (100 mg/day), ciprofloxacillin (500–75 mg twice a day), parental amikacin (7.5–15 mg daily for 2–4 weeks) and another macrolide azithromycin.

Primary prophylaxis has been shown to significantly reduce the incidence of M.avium complex bacteriaemia and should be considered in patients whose CD4 counts are less than 75×10^6/l. A variety of agents have been shown to be effective including rifabutin 300 mg daily, clarithromycin 500 mg twice daily or azithromycin 1200 mg once weekly. Resistant strains on clarithromycin and azithromycin prophylaxis can occur in those who develop breakthrough bacteriaemia, and there is cross-resistance. A combination of once weekly azithromycin and once daily rifabutin is probably the most effective prophylaxis regimen and may also provide additional prophylaxis against PCP.

Salmonella infections are treated with either co-trimoxazole or ciprofloxacin and campylobacter with ciprofloxacin. In salmonella infections relapses of enteritis or bacteraemia are common.

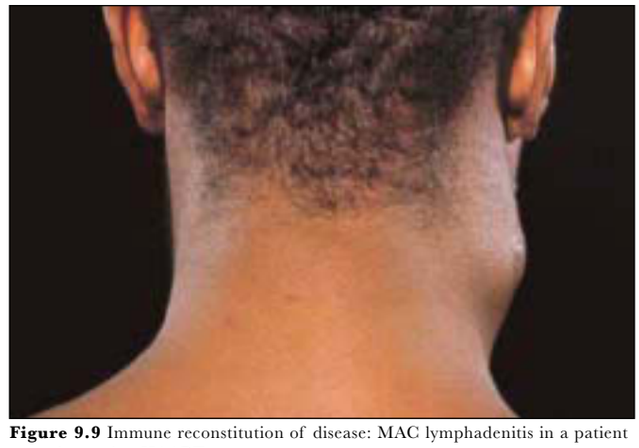


Figure 9.9 Immune reconstitution of disease: MAC lymphadenitis in a patient recently starting HAART

Table 9.5 Targets for antiretroviral therapy

Target	Treatment
Virus receptor and entry	Fusion inhibitors, chemokine receptor blockers
Reverse transcriptase	Inhibitor/DNA chain terminators
RNAase	Inhibitors
Integration	Viral integrase inhibitors
Viral gene expression	Inhibitors of HIV regulatory genes and their products
Viral proteins synthesis	Enzyme inhibitors, for example, protease inhibitors
Viral budding	Interferons (also act at other sites of replication cycle) antibodies and ligands

Antiretroviral drugs

The clinical effectiveness of antiretroviral therapy has improved markedly over the last few years. Since 1996 in the developed world there have been dramatic falls in the incidence of new AIDS cases and AIDS-associated deaths. Published data in the late 1990s estimated the mortality rate in patients with CD4 counts of less than 100×10^6/l had fallen by nearly two-thirds to <8 per patient years. Although the long-term clinical efficacy of the current antiretroviral treatment regimens remains uncertain, the biological rationale for maintaining a clinical response has been established. Sustained inhibition of viral replication results in partial reconstitution of the immune system in most patients, substantially reducing the risk of clinical disease progression and death. Reservoirs of HIV in latently infected resting T-lymphocytes and other long-lived cell populations makes it unlikely that HIV can be eradicated by antiretroviral therapy alone. Strategies to sustain suppression of viral replication in the long-term will be necessary.

Box 9.5 Antiretroviral regimens

1. 2 NRTIs: eg. Zidovudine or Stavudine + Lamivudine or Didanosine
   plus either
   1 NRTI: Nevirapine or Efavirenz
   or
   1 PI: Nelfinavir, Saquinavir soft gel or a low dose Ritonavir boosted PI
   or
   2 PIs: eg. Saquinavir + Ritonavir
2. 3 NRTIs: Zidovudine, Lamivudine + Abacavir

Antiretroviral regimens for the initial treatment of chronic infection in adults (2001). Choice would depend upon efficacy, tolerability, adherence and resistance profile of the regimen. Treatment guidelines are constantly reviewed and updated.

There are several potential targets for antiretroviral drugs in the viral replication cycle. Three classes of antiretroviral drugs are currently used in combination for the treatment of HIV infection, which target the activity of two viral enzymes. New therapeutic agents are constantly being evaluated.

Reverse transcriptase inhibitors

The first drugs made available for clinical use were inhibitors of the HIV reverse transcriptase enzyme. Before the virus can be integrated into the host cell genome DNA, a copy of the viral RNA has to be formed (pro-viral DNA). This is regulated by the specific HIV DNA polymerase: reverse transcriptase (RT). If a DNA copy is not formed, the viral RNA genome becomes susceptable to destruction by cellular enzymes.

The nucleoside reverse transcriptase inhibitors (NRTIs) are both competitive inhibitors of RT and DNA chain terminators. The normal 2' deoxynucleosides which are substrates for DNA synthesis link to form a chain by phosphodiester linkages bridging the 5' and 3' positions on the five carbon sugar molecule. The 2', 3'-dideoxynucleosides analogues are formed by the replacement of the 3'-hydroxy group by an azido (zidovudine), hydrogen or other group. These nucleoside analogues as substrates will bind to the active site of the HIV RT enzyme and will be added to the growing HIV proviral DNA chain. However, once inserted, the normal 5' to 3' links will not occur resulting in HIV proviral DNA chain termination.

Genotypic mutations at various codons in the RT gene result in decreased susceptability of HIV to inhibition by the NRTIs. Several NRTIs are currently licensed for the treatment of HIV infection in combination regimens and newer agents with better tolerability and resistance profiles are under evaluation.

The non-nucleoside reverse transcriptase inhibitors (NNRTIs) are a group of structually diverse agents which bind to RT at a site distant to the active site resulting in confirmational changes at the active site and inhibition of enzyme activity. These agents show high antiviral activity *in vitro* and have relatively low toxicity. They are also highly specific, inhibiting the reverse transcriptase of HIV-1 but not HIV-2. As monotherapy, rapid emergence of resistant strains associated with single point mutations of the RT gene, high-level phenotype resistance and loss of antiviral effect occurs. The drugs therefore need to be combined with other antiretroviral agents, usually two NRTIs, to achieve and maintain an effective long-term treatment response.

Protease inhibitors

The protease inhibitors bind competitively to the substrate site of the viral protease enzyme. This enzyme is responsible for the post-translational processing and cleavage of a large structural core protein during budding from the infected cell. Inhibition results in the production of immature virus particles. Their potent anti-HIV activity and introduction to clinical use from 1996 was one of the main reasons for the observed substantial falls in morbidity and mortality associated with HIV infection in the developed world. However, tolerability, relatively high pill burden and poor adherence were frequent problems with the initial protease inhibitor containing regimens. Specific genotypic mutations in the protease gene can result in high levels of phenotype resistance to individual protease inhibitors and cross-resistance. New protease inhibitors are under evaluation.

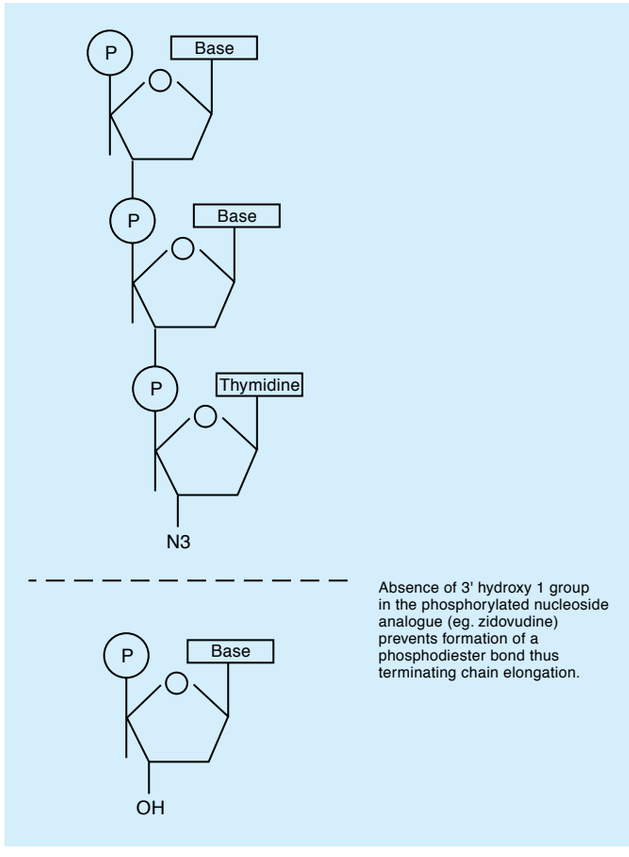


Figure 9.10 Mechanism of action of nucleoside reverse transcriptase inhibitors

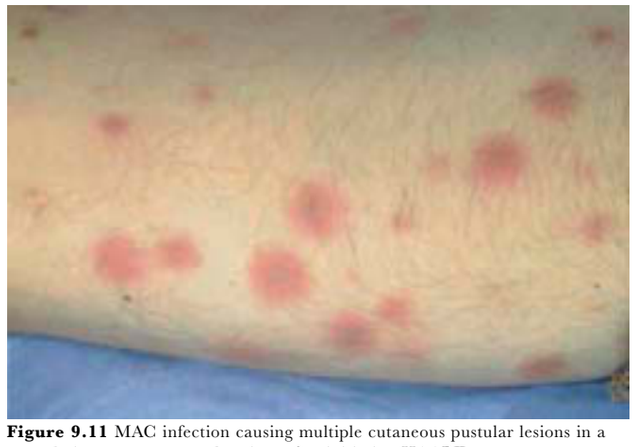


Figure 9.11 MAC infection causing multiple cutaneous pustular lesions in a severely immunosuppressed patient after initiating HAART

Treatment of chronic adult infection

In the mid 1990s, several large clinical endpoint studies demonstrated a strong association between falls in plasma HIV RNA levels (plasma viral load) in the first few weeks on therapy and clinical outcome at one year. It is now accepted that falls in plasma viral load combined with increases in CD4 count are predictive of the clinical treatment response on different combination regimens at 1–2 years, although changes in the markers probably do not fully predict the observed clinical effect.

Studies have also shown an association between the plasma viral load nadir on therapy and both the risk of subsequent viral load rebound, and the emergence of viral genotypic mutations associated with reduced drug susceptibility. Where possible an objective of antiretroviral therapy is to reduce and sustain plasma viral load levels to below the level of detectability of the current ultra-sensistive viral load assays (< 50 copies/ml). If patients are adherent to therapy, the likelihood of a viral load rebound and drug resistance is minimal. Despite inhibition of viral replication in plasma, lymph nodes and at other sites, reservoirs of HIV infection in latently infected resting T-lymphocytes remain. Continued activation of these cells will theoretically result in the reduction of this reservoir, however new cells probably continue to be infected either as a result of localised small bursts of viral replication or loss of the antiretroviral effect of the treatment regimen. Even in patients who have sustained, undetectable levels of plasma viral load (< 50 copies/ml) for three years or more, discontinuation of antiretroviral therapy results in rapid rebound of plasma viral load to pretreatment levels.

Sustained inhibition of viral replication does however result in substantial immune reconstitution, even in those patients with advanced disease who start antiretroviral therapy at very low CD4 counts. Reduction in immune activation markers, increases in both memory and naïve CD4 and CD8 T cells and development of improved lymphoproliferative responses to antigens such as CMV and mycobacteria occur in patients on HAART. Immune responses to HIV are generally not regained, and it remains uncertain what levels of immune reconstitution can be achieved over time. This may depend on any residual thymic function or the ability of extrathymic pathways to facilitate immune reconstitution.

To achieve sustained falls in plasma viral load it is standard of care in patients starting antiretroviral therapy for the first time to use a triple drug regimen containing two NRTIs in combination with either one NNRTI or one or two protease inhibitors. In clinical trials, a combination of two NRTIs and a protease inhibitor has been shown to reduce the risk of progression to AIDS or death compared to treatment with two NRTIs alone. There is no similar clinical endpoint data for NNRTI-containing combinations, however randomised trials have shown that treatment with a combination of two NRTIs and one NNRTI results in similar falls in plasma viral load and increases in CD4 count after one year to treatment with two NRTIs and a protease inhibitor. On the basis of these results, it is recommended to inititate therapy with either a PI or an NNRTI containing triple combination. Large randomised trials are under way to evaluate which starting regimen is better in the long term.

The efficacy of antiretroviral therapy has improved over the last few years, however only approximately 50–70% of patients will have sustained plasma viral loads to <50 copies/ml at one year. An important factor associated with treatment success is adherence. Patients who are able to tolerate and adhere to their treatment regimen are more likely to achieve and sustain

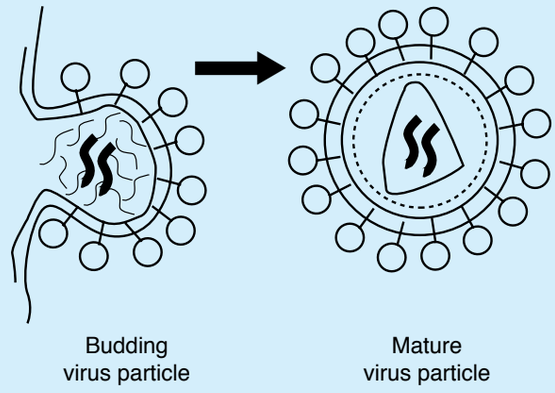


Figure 9.12 Site of Action of Protease Inhibitors

Table 9.6 Recommendations for starting antiretroviral therapy in adults

Disease stage	BHIVA (1)	USDHHS (2)
Symptomatic	Treat	Treat
Asymptomatic: CD4 $<200\times10^6$/l	Treat	Treat
CD4 count $200–350\times10^6$/l	Consider treatment depending upon VL, rate of CD4 count decline, symptoms and patient wishes	Treatment should generally be offered
CD4 $>350\times10^6$/l	Defer	Defer or consider treatment if high VL

(1) BHIVA: British HIV Association Guidelines (March 2001)
(2) USDHHS: United States Department of Health and Human Services (February 2001)

Box 9.6 Factors associated with virological treatment failure

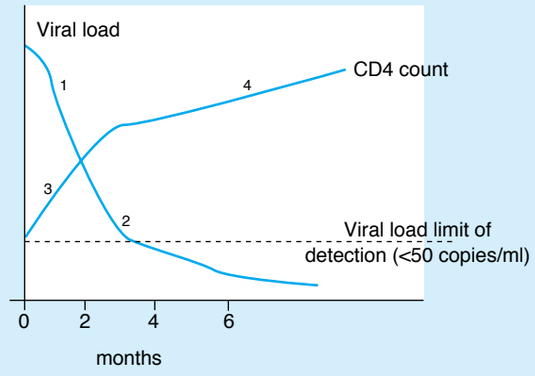
- poor adherence
- drug intolerance and toxicity
- drug–drug interactions resulting in sub-optimal drug levels
- development of genotypic mutations associated with reduced drug susceptibility
- pharmacological resistance resulting in decreased intracellular drug levels

suppression of plasma viral load than those who do not. Few patients experience virological treatment failure as a result of poor antiviral potency. The ability of a patient to adhere to a treatment regimen is important in determining the choice of treatment regimen. The dosing schedule, pill burden, the requirement or not for dietary restrictions, risk of side-effects and patient motivation are important in determining adherence. Other factors which contribute to the initial treatment choice are baseline viral load, resistance profile of the drug, future options for treatment, known efficacy of the treatment regimen, the potential for drug to drug interactions and the presence of drug resistance at baseline.

The optimal time to initiate therapy with the current antiretroviral drugs has not been established in clinical studies. CD4 count and plasma viral load are predictors of the estimated risk of progression to AIDS which is a factor in determining when to start treatment. The motivation of a patient to start and adhere to therapy and the known effectiveness of current regimens are also important. Clinical practice across Europe and North America varies, but most clinicans would consider initiating therapy at some point between a CD4 count of 200–500 $\times$ 10^6/l and in all patients who are symptomatic. Even in patients who initiate therapy with CD4 counts of <100 $\times$ 10^6/l, substantial increases in CD4 count and clinical benefit can be achieved. Patients on therapy should have CD4 count and plasma viral load levels monitored at regular intervals. On effective therapy, plasma viral load falls rapidly as viral replication is inhibited. By four weeks a fall of greater than 1 log and by 3–6 months a fall to < 50 copies/ml should be expected.

Apart from drug intolerance, indications to change therapy have not yet been fully defined and evaluated. Physicians, however will use evidence of clinical progression, a fall in CD4 count and a rise in plasma viral load as markers of therapy failure, and consider changing therapy. When to switch therapy will also depend on available treatment options, patient adherence and the emergence of drug resistance and the potential for cross-resistance to other drugs.

In the antiretroviral experienced patient, the objective of therapy remains similar to that in patients starting treatment for the first time. Therapeutic options, however, are more limited because of previous drug toxicity and the presence of genotypic mutations conferring drug resistance.



1. Initial rapid fall in viral load: as a result of inhibition of viral replication in productively infected lymphocytes
2. Slower second phase decay in viral load, becoming undetectable (<50 copies/ml) at 4–6 months or treatment
3. Initial increase in CD4 count due to redistribution and expansion of CD4 memory cells
4. Subsequent rise due to production and expansion of new naive and memory cells

Figure 9.13 Changes in CD4 count and plasma HIV RNA levels (viral load) in patients treated with HAART

Box 9.7 Factors determining when to start and choice of therapy

- Risk of clinical disease progression (CD4 count, viral load)
- Willingness of patient to start therapy
- Clinical effectiveness of combination regimen
- Ability and motivation of patient to adhere to therapy
- Drug toxicity profile
- Pill burden and dosing schedule
- Transmitted drug resistance
- Future therapy options
- Likelihood of drug resistance
- Drug–drug interactions

Treatment of primary infection

It remains uncertain whether patients who present with acute primary HIV infection should be started on antiretroviral combination therapy. It is likely that the quality and breadth of cytotoxic T lymphocyte-cell (CTL) responses to different HIV antigens at the time of primary infection determines how well the immune system controls HIV replication over time. Studies have also shown that specific T-helper cell responses to HIV antigens are rapidly lost during primary infection, disabling the immune response to HIV. In patients who start antiretroviral therapy within the first few weeks of initial infection, specific T-helper cell responses to HIV are preserved, and CD4 cell counts are higher compared to the levels in patients who do not start therapy. The relevance of this immunological benefit is uncertain, or whether there is any long-term clinical advantage compared to patients who initiate effective antiretroviral therapy at some point during established chronic infection. Even after 2–3 years of sustained reduction in plasma viral load, discontinuation of therapy results in rapid viral rebound. Whether specific immune responses to HIV are more likely to be preserved and the level of viral load on discontinuation of

Box 9.8 Treatment of primary HIV infection

Pros:

- Preserve immune function including specific T helper cell responses to HIV
- Decrease magnitude of virus dissemination and establishment of viral reservoirs
- Potentially alters viral set-point in favour of slower disease progression on discontinuation of HAART

Cons:

- Longterm toxicity associated with drug treatment, potential for therapy failure and emergence of drug resistance
- Uncertainty of improved longterm clinical benefit compared to initiating treatment during established chronic HIV infection
- Probably need to treat within a few weeks of exposure to HIV infection to gain immunological benefit.

therapy is lower than might have occurred in patients who were not treated during primary infection is not known. Antiretroviral therapy should be considered in patients presenting with acute primary HIV infection, however the immunological arguments need to be balanced against the unknown long-term efficacy of such a strategy, the risk of drug toxicity over time and the development of drug resistance.

Drug resistance

Soon after the introduction of zidovudine into clinical practice it was recognised that viral isolates taken from patients six months after therapy were less susceptible to zidovudine than at baseline. The emergence of genotypic mutations in the reverse transcriptase gene was associated with reduced susceptibility. Genotypic and phenotypic resistance can develop against all currently antiretroviral drugs and is a major factor contributing to therapy failure. Multiple mutations in the RT and the protease genes have now been identified to be associated with reduced drug susceptibility. The pattern and number of mutations which emerge and whether they confer cross-resistance within the class differs between each drug and regimen. For certain drugs, for example, lamivudine, nevirapine or efavirenz, the emergence of a single point mutation within the RT gene confers a very high fold decrease in susceptibility. For other drugs the fold decrease in susceptibility is much lower and multiple mutations may be needed to confer high-level drug resistance. Cross-resistance within a class can occur particularly with the NNRTIs and the protease inhibitors. For the NNRTIs this requires single genotypic mutation only, while for the protease inhibitors this usually requires a primary mutation plus four or five other secondary mutations. The emergence of resistance to all drugs does not always occur with a combination in a patient who experiences virological rebound on therapy. Some patients do not develop any genotypic mutations on treatment failure and this may reflect poor adherence and low drug selection pressure. Patients who achieve sustained falls in plasma viral load to less than 400 copies per ml are less likely to develop genotypic mutations associated with drug resistance than those who do not. Drug-resistant viruses can be transmitted and various recent studies have shown that 10–15% of patients presenting with primary HIV infection have genotypic mutations associated with drug resistance particularly in the RT gene. Drug-associated genotypic mutations usually fade on withdrawal of drug therapy but frequently rapidly re-emerge if the same drugs are taken again in combination.

The presence of drug resistance may affect the choice and efficacy of therapy in patients who have previously failed one, two or more combination regimens. Genotypic and phenotypic resistance assays are available in clinical practice and early randomised studies suggest that their utility in helping with the choice of therapy may result in greater falls in viral load in the short term. There are larger randomised studies ongoing and the exact role of these assays in clinical practice is yet to be established. The usefulness of these assays may depend upon the availability of alternative effective antiretroviral agents in a treatment experienced patient.

Drug toxicities

The tolerability and side-effects of a combination regimen is very important in determining the antiviral response. In clinical practice 40–50% of patients will not have sustained falls in plasma viral load by one year of therapy and a major factor contributing to this is poor tolerability. Drug-specific side-effects are listed in Table 9.6.

Table 9.7 Common Primary Genotypic Mutations associated with reduced drug susceptibility in-vivo

NRTIs	**Codon**
Zidovudine	K70N, T215Y
Abacavir	K65R, L74V, M184V
Didanosine	L74V
Lamivudine	M184V
Zalcitabine	K65R, T69D, L74V
Stavudine	T215Y + other TAMS[1]
Multi nucleoside resistance	Q151M, T69S-SS
NNRTIs	
Nevirapine	K103N, V106A, Y181C, G190A, Y188C/L/H
Efavirenz	K103N, G190A, Y188C/L/H
PIs	
Saquinavir	G48V, L90M
Ritonavir	V82A/T/F
Indinavir	M46I/L, V82A/T/F
Nelfinavir	D30N, L90M
Amprenavir	I50V, I54L/M, I84V

1. TAMS: Thymidine analogue mutations
2. Resistance profile of lopinavir/r in-vivo is uncertain

Clinically relevant phenotypic resistance may require only a single primary mutation or 2 or more primary and secondary mutations. Interpretation of a genotypic test is complex and requires expert advice.

Box 9.9 Role of resistance testing

- detection of transmitted resistance in primary infection
- detection of resistance prior to starting treatment for the first time
- guide choice of new treatment regimen in patients experiencing virological treatment failure on first or subsequent regimens
- guide treatment choice in pregnant mothers for prevention of vertical transmission

The utility of genotypic and phenotypic resistance tests are continuing to be evaluated.

In the last two to three years abnormalities of fat redistribution have been observed in patients on combination regimen. Observational cohort studies suggest that lipodystrophy may occur in up to 50–60% of patients after one to two years on therapy. Patients either present with peripheral fat wasting affecting the buttocks, limbs and face or fat accumulation round internal viscera in the abdomen resulting in a distended abdomen and bloating. The exact pathogenosis of these fat distribution syndromes is unknown but age of patient, antiretroviral drug therapy and time on therapy may all be implicated. They have been reported in both protease inhibitor and NRTI-containing combination regimens, and it is likely that it is a mixed syndrome with a multifactorial cause. The occurrence of lipodystrophy can affect the psychological well being of the patient but as yet we do not know how it is best managed.

It has recently been suggested that mitochondrial toxicity may account for some of the toxcities associated with the NRTIs as a result of inhibition of mitochrondirial gamma DNA polymerase. Severe lacticacidosis is a rare complication of NRTI therapy.

Future agents

For the reasons of poor tolerability, suboptimal antiviral potency and long-term drug toxicity, it is important that new antiretroviral agents and therapeutic strategies are developed and evaluated. New formulations of current drugs which improve tolerability and reduce pill burden will help to improve adherence in patients. New protease and reverse transcriptase inhibitors are currently undergoing clinical trials which *in vitro* appear to be effective against viral isolates which are resistant to different drugs. Whether these agents will prove to be clinically effective will be important in treating those patients who have previously failed combination therapies.

New classes of drugs are also being developed. Fusion inhibitors which block the activity of the GP41 viral transmembrane protein are in Phase III clinical trials and are likely to be the first new class of drug to reach the bedside.

As well as specific drugs that inhibit targets in the viral replication cycle, immunotherapeutic approaches are so being assessed. Treatment with cycles of the cytokine interleukin 2 results in substantial increase in CD4 counts but has little effect on plasma viral load levels. Interleukin 2 may also improve immune responses to HIV and a large randomised international trial is underway to assess its efficacy in combination with effective antiretroviral combination regimens. Therapeutic vaccines are also under evaluation which might improve specific immune responses and assist immunological control of HIV replication. Their clinical effectiveness remains uncertain.

Few areas of medicine have seen such dramatic changes in treatment with a resulting reduction in morbidity and mortality as there has been in patients with HIV infection. It is very likely that therapeutic options will continue to improve, although the long-term efficacy of treatment over many years still remains uncertain.

Table 9.8 Drug toxicities

Drug	Toxicity
NRTIs	
Class associated	Lactic acidosis
	Hepatitic steatosis
	Lipodystrophy (peripheral fat wasting)
Drug specific	
Zidovudine	Bone marrow suppression, nausea, vomiting, myopathy
Stavudine	Peripheral neuropathy, hepatitis
Zalcitabine	Peripheral neuropathy, mouth ulcers
Didanosine	Pancreatitis, dry mouth, peripheral neuropathy
Lamivudine	Few side-effects
Abacavir	Hypersensitivity reaction, nausea
NNRTIs	
Nevirapine	Rash, hepatitis, Steven–Johnson syndrome
Efavirenz	Rash, dysphoria, mood changes, vivid dreams, hypercholesterolaemia, hepatitis
PIs	
Class specific	Lipodystrophy (fat wasting or accumulation)
	Hyperlipidaemia, diabetes mellitus
Drug specific	
Nelfinavir	Diarrhoea, rash
Saquinavir	Few side-effects
Indinavir	Hyperbilirubinaemia, nephrolithiasis, nail changes, dry skin
Ritonavir	Perioral dysathesia, flushing, hepatitis, diarrhoea, nausea, vomiting
Amprenavir	Rash, nausea, diarrhoea
Lopinavir	Diarrhoea

10 HIV infection and AIDS in the developing world

Alison D Grant, Kevin M De Cock

Epidemiology of HIV-1 and HIV-2 infections in developing countries

The epidemiology and burden of HIV in the developing world are discussed earlier (see chapter 1). Two distinct viruses, HIV types 1 and 2 (HIV-1/HIV-2), cause AIDS. HIV-1 is responsible for the great majority of infections globally, HIV-2 being very rare outside of West Africa. Individual cases of HIV-2 infection have been described in other parts of Africa, Europe, the Americas and Asia (India), but most people with HIV-2 infection have some epidemiological link to West Africa.

The routes of transmission of HIV-1 and HIV-2 (as described in chapter 1) are the same worldwide, but the relative importance of different modes of transmission differs according to region. In most developing countries, heterosexual transmission is the dominant mode of spread, and mother-to-child transmission of HIV is much more common than in industrialised countries. Homosexual transmission is rare in Africa, but is more common in south-east Asia and central and south America. Transmission associated with injecting drug use is particularly frequent in parts of south and south-east Asia and central and south America. Acquisition of infection from contaminated blood remains a problem, especially in parts of sub-Saharan Africa and south Asia; in some countries commercial blood donation acts to amplify the spread of transfusion-transmitted HIV infection, both to the recipients of blood as well as to donors who may become infected through exposure to unsterile equipment. Women and children are at especially high risk for transfusion-transmitted HIV infection, the former because of the high incidence of anaemia and haemorrhage associated with pregnancy and childbirth, and the latter because of malarial anaemia.

The transmission of HIV-2 infection is less efficient than that of HIV-1; this applies particularly to mother-to-child transmission, with only about 1% of HIV-2 infected mothers passing the infection on to their offspring. By comparison, up to 42% of HIV-1 infected mothers pass the infection to their children by all routes (intrauterine, puerperal and breast milk). Sexual transmission of HIV-2 is also less efficient, especially before the development of end-stage immune deficiency. Postnatal transmission of HIV-1 by breast milk is more important than previously believed and approximately doubles the risk of mother-to-child transmission.

Natural history of HIV-1 and HIV-2 infections in developing countries

There is still relatively little information available about the natural history of HIV-1 and HIV-2 infections in the developing world. Prospective studies from industrialised countries before highly active antiretroviral therapy (HAART) was widely used suggest that after 10 years of infection with HIV-1, approximately 50% of people will have developed AIDS. There has been a widespread belief that HIV-1 disease progresses more rapidly in developing countries, but more recent evidence suggests that the rate of progression from infection to severe immunosuppression may be little different from that documented in industrialised countries in the pre-HAART era.

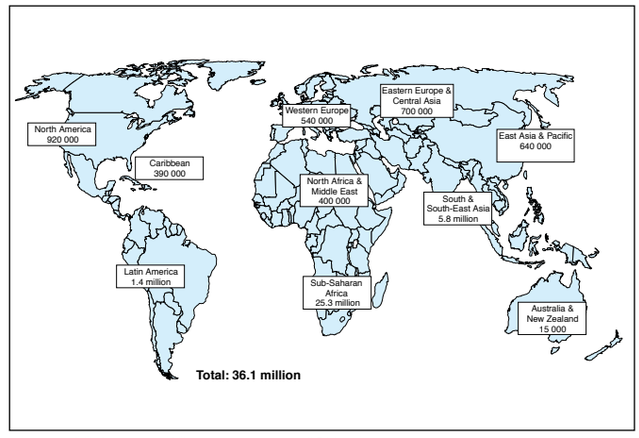


Figure 10.1 UNAIDS estimates that 95% of people living with HIV/AIDS are in developing countries

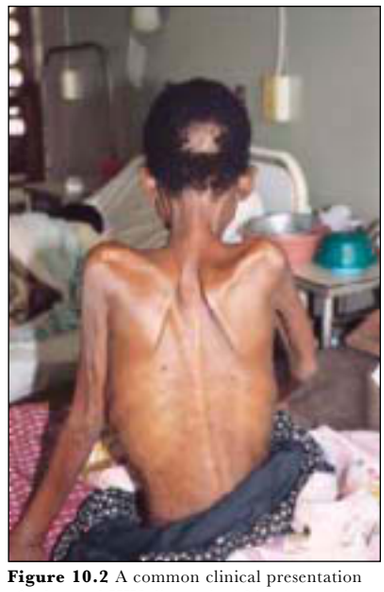


Figure 10.2 A common clinical presentation of advanced HIV disease in African countries is with marked wasting, known in Uganda as "slim" disease (courtesy of Professor Sebastian Lucas)

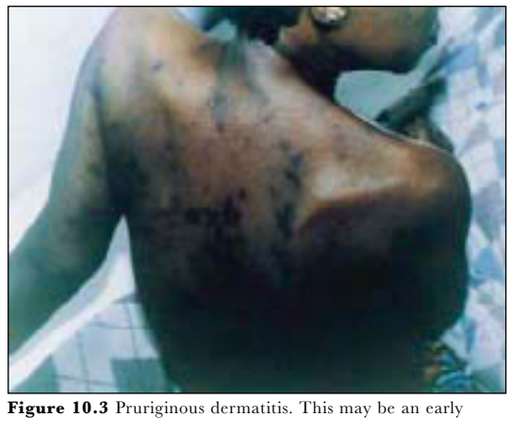


Figure 10.3 Pruriginous dermatitis. This may be an early manifestation of HIV infection

Because bacterial diseases such as pneumococcal disease and tuberculosis are prevalent in developing countries and may occur at relatively high CD4+ lymphocyte (CD4) counts, some HIV-infected persons may appear to become symptomatic earlier. In addition, outcome may be worse in developing countries than in the industrialised world because of lack of access to care: this is almost certainly the main explanation for the reduced survival following the development of an AIDS-defining illness in developing countries, generally around six to nine months.

Progression from infection to disease is substantially slower for HIV-2 infection compared with HIV-1. Evidence for this includes individual reports of long survival with HIV-2 infection, higher levels of CD4 counts in HIV-2- than in HIV-1-infected people in cross-sectional studies, and a lower incidence of CD4 decline and AIDS in cohort studies comparing HIV-1- and HIV-2-infected people. Nonetheless, HIV-2 may eventually cause severe immunosuppression, accompanied by disease which is clinically indistinguishable from that caused by HIV-1.

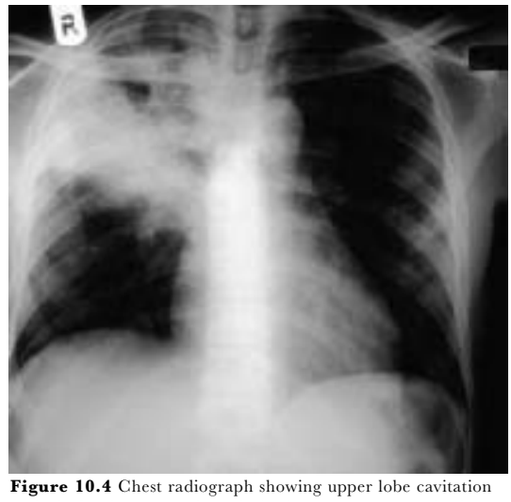


Figure 10.4 Chest radiograph showing upper lobe cavitation typical of pulmonary tuberculosis. Appearances may also be atypical (see chapter 6)

Clinical aspects of HIV disease in developing countries

As in industrialised countries, the spectrum of clinical manifestations associated with HIV infection is wide, ranging, as the CD4 count falls, from an asymptomatic state, through symptomatic disease, to fatal illness characterised by opportunistic infections, malignancies, neurological disease and wasting. Initial acquisition of HIV infection ("acute HIV infection" or "seroconversion illness") may be complicated by a syndrome resembling infectious mononucleosis, or a wide range of other manifestations as described in chapter 4. However, since this syndrome is not specific, it is rarely recognised even when clinically apparent.

Early manifestations of HIV disease

Common early symptoms and signs are weight loss, fever, night sweats and diarrhoea. Skin disorders are frequent early manifestations, especially varicella zoster, fungal infections and pruriginous dermatitis, an itchy rash consisting initially of papules which become shallow ulcers due to scratching, and finally heal, leaving pigmented macules.

Tuberculosis

Tuberculosis is unquestionably the most important opportunistic infection complicating HIV infection in developing countries, and may present at any stage in the course of immunodeficiency. In early HIV disease, pulmonary tuberculosis is similar to that found in HIV-negative people. In advanced immunodeficiency, tuberculosis is often disseminated and multibacillary in nature. Nocardiosis, while much less common, is a differential diagnosis in some areas.

Bacterial septicaemia

An inadequately recognised manifestation of HIV disease in developing countries has been bacterial septicaemia. Gram-negative organisms are the most common pathogens identified, especially non-typhoid *Salmonella* spp. Invasive pneumococcal disease is also frequent, and may occur earlier than Gram-negative infections. In some patients with advanced HIV disease, mycobacteraemia is detectable; this is much more frequently due to *Mycobacterium tuberculosis* than *Mycobacterium avium intracellulare* complex.

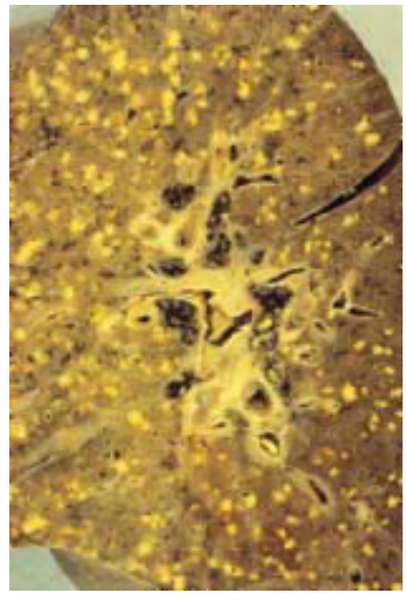


Figure 10.5 Post-mortem lung showing miliary tuberculosis (courtesy of Professor Sebastian Lucas)

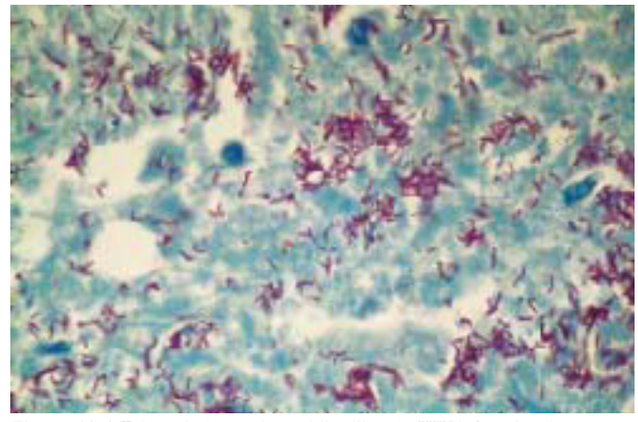


Figure 10.6 Tuberculosis may be multibacillary in HIV-infected patients: non-reactive tuberculous infection of the pericardium, showing abundant acid-fast bacilli (courtesy of Professor Sebastian Lucas)

Diarrhoeal disease and HIV wasting syndrome
The best known clinical picture of AIDS in Africa is "slim", the term given by people in rural Uganda to the HIV wasting syndrome. Profound wasting, chronic diarrhoea and fever are the typical features. About half the time no specific aetiology can be found for the diarrhoea: among identified causes, the most common are cryptosporidiosis, microsporidiosis, isosporiasis and bacterial infections. The commonest autopsy finding in African patients with HIV wasting syndrome is disseminated tuberculosis, and undue emphasis may have been put on searching for a primary gastrointestinal cause of this whole syndrome. As with all medical causes of wasting, an important contributing factor to the HIV wasting syndrome is reduced food intake.

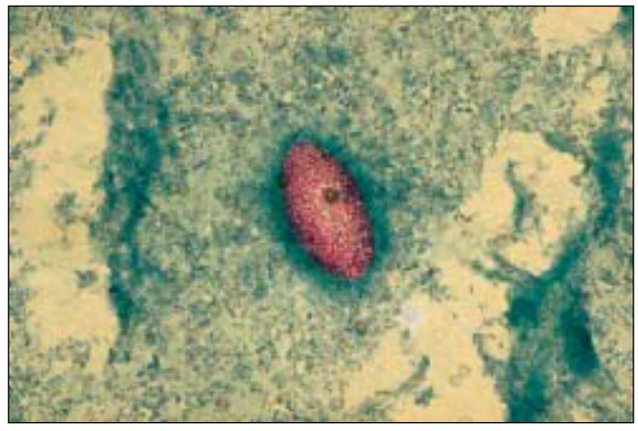


Figure 10.7 *Isospora belli*, a treatable course of diarrhoea in HIV-infected people

Neurological disease
Cerebral toxoplasmosis and cryptococcal meningitis are probably more frequent causes of severe HIV-related disease in developing than industrialised countries, and their prevalence may vary by geographical region. Cerebral toxoplasmosis most often presents as a space-occupying lesion of the brain, and cryptococcosis as a chronic meningitis.

Regional variation in disease spectrum
Tuberculosis and bacterial infections, particularly pneumococcal disease, are common HIV-related diseases in developing countries worldwide. Other HIV-related diseases show regional variation. Pneumocystosis, cytomegalovirus disease and disease due to atypical mycobacteria such as *Mycobacterium avium intracellulare*, common in industrialised countries, are unusual in adults in many African countries (although *Pneumocystis carinii* pneumonia is common in HIV-infected African infants). The reasons for this are uncertain, but may include development of diseases such as tuberculosis at higher levels of CD4 counts, and shorter survival once the stage of profound immunodeficiency has been reached. Endemic Kaposi's sarcoma is more common in Central and East than in West Africa, and this is probably also true for the AIDS-associated form.

Some HIV-related diseases are limited to specific geographic areas, such as disease due to the fungus *Penicillium marneffei*, which is confined to south-east Asia. Penicilliosis causes disseminated disease in patients with advanced immune deficiency, with nodular skin lesions as the most obvious manifestation. Tuberculosis, salmonellosis and cryptococcosis are other frequent AIDS-defining conditions in south and south-east Asia. Tuberculosis is frequent in Latin America, where the spectrum of disease is otherwise similar to that in the industrialised world.

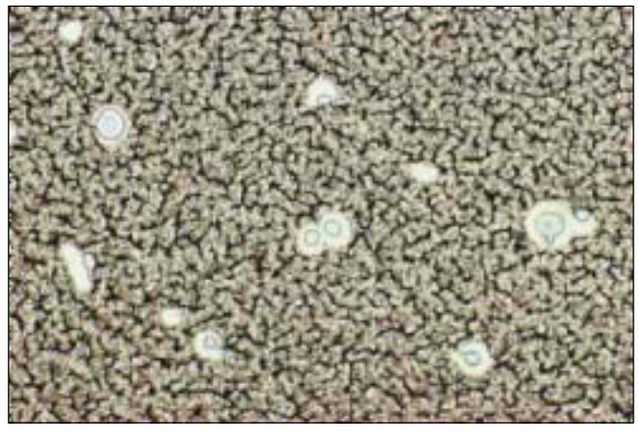


Figure 10.8 India ink stain of cerebrospinal fluid showing *Cryptococcus neoformans*, a common cause of meningitis (courtesy of Professor Sebastian Lucas)

Association with endemic tropical diseases
The association between endemic tropical diseases and HIV infection has only been studied to a limited degree. Theoretically, HIV infection could increase the incidence of tropical diseases, and alter their natural history, clinical expression, or response to treatment. Malaria is indirectly linked to HIV infection by causing anaemia in children, who may then be at risk for HIV infection transmitted through blood transfusion. HIV-infected pregnant women experience greater frequency and severity of malarial parasitaemia, and increased frequency of placental malaria compared with HIV negative women. HIV-infected people with *Schistosoma mansoni* excrete fewer eggs than those who are HIV negative, but it is not known whether the severity of schistosomiasis is affected by HIV infection, and response to treatment seems to be unaffected by HIV status. Amoebiasis and strongyloidiasis might be expected to be more frequent in HIV disease, but are not; on the basis of

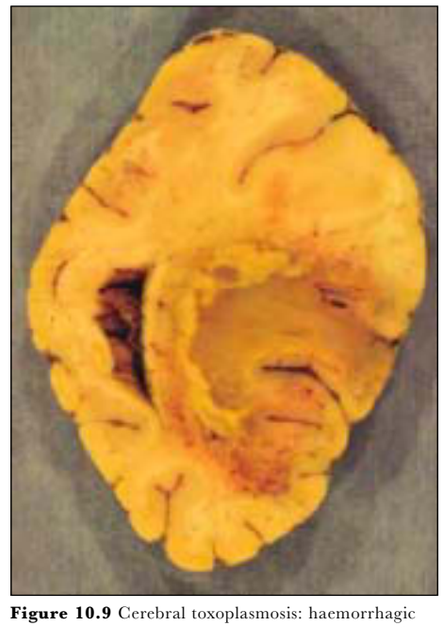


Figure 10.9 Cerebral toxoplasmosis: haemorrhagic and necrotic mass in the occipital lobe (courtesy of Professor Sebastian Lucas)

limited data, the same seems to be true of trypanosomiasis and leprosy. Little information is available concerning the influence of HIV infection on filariasis. Visceral leishmaniasis, often disseminated, appears to be increased in incidence in HIV-infected persons, although most reports have been from southern Europe rather than sub-Saharan Africa or South America. HIV-infected persons with leishmaniasis require maintenance treatment as relapse is otherwise likely.

AIDS case definitions and staging of HIV disease

Case definitions of AIDS for epidemiological surveillance

For epidemiological surveillance, a practical case definition of severe HIV-related disease is needed. The Centers for Disease Control (CDC) AIDS surveillance case definition is used in many industrialised countries (see chapter 1), but cannot be used in most developing countries because it requires access to sophisticated laboratory investigations. For this reason, the World Health Organisation (WHO) introduced a clinical case definition that could be used in settings where laboratory facilities are inaccessible (Box 10.1). In 1994, this definition was expanded to incorporate HIV serology (thus increasing specificity) and to take account of revisions of the CDC case definition (Box 10.2). If serological testing is unavailable or inaccessible, the clinical case definition should be used; if serological testing is available, the expanded case definition should be used.

Diagnosis and clinical staging of HIV disease in resource poor settings

Although advanced HIV disease may be easy to diagnose clinically, it is desirable to have HIV serology on patients with suspected HIV disease, particularly since HIV negative tuberculosis may be clinically indistinguishable from advanced HIV disease.

The case definitions in Boxes 10.1 and 10.2 were developed for epidemiological surveillance, and are not intended to be used for clinical staging of patients, for which they are neither sensitive nor specific. In order to estimate prognosis in individual patients, a clinical staging system is more useful than a case definition. Box 10.3 overleaf outlines the WHO proposed staging system for HIV infection and disease, using clinical and laboratory data, which can be used in developing countries. This system categorises patients into four stages based on clinical features of prognostic significance. The stages are interpreted as:

Stage 1: asymptomatic infection.
Stage 2: early (mild) disease.
Stage 3: intermediate (moderate) disease.
Stage 4: late (severe) disease.

The system can be refined using a laboratory axis: the CD4 count is the most useful laboratory marker for clinical staging, but is rarely available in developing countries. The total lymphocyte count can be used as a surrogate, although this is not ideal. Manifestations of HIV disease are rare at CD4 counts above 500×10^6/l and severe illness and death are rare in patients with counts above 200×10^6/l. Tuberculosis and pneumococcal disease may occur at higher as well as lower CD4 counts. Once patients in developing countries have developed advanced HIV disease, they die with higher CD4 levels than in industrialised countries because of lack of access to high quality medical care; nonetheless, most patients die at the stage of advanced immunodeficiency.

Box 10.1 WHO AIDS case definition for AIDS surveillance

For the purposes of AIDS surveillance an adult or adolescent (>12 years of age) is considered to have AIDS if at least two of the following major signs are present in combination with at least one of the minor signs listed below, and if these signs are not known to be due to a condition unrelated to HIV infection.

Major signs
- Weight loss 10% of body weight
- Chronic diarrhoea for > 1 month
- Prolonged fever for > 1 month (intermittent or constant)

Minor signs
- Persistent cough for > 1 month*
- Generalised pruritic dermatitis
- History of herpes zoster
- Oropharyngeal candidiasis
- Chronic progressive or disseminated herpes simplex infection
- Generalised lymphadenopathy

The presence of either generalised Kaposi's sarcoma or cryptococcal meningitis is sufficient for the diagnosis of AIDS for surveillance purposes.

*For patients with tuberculosis, persistent cough for >1 month should not be considered as a minor sign.

Box 10.2 Expanded WHO case definition for AIDS surveillance

For the purposes of AIDS surveillance an adult or adolescent (>12 years of age) is considered to have AIDS if a test for HIV antibody gives a positive result, and one or more of the following conditions are present:
- 10% body weight loss or cachexia, with diarrhoea or fever, or both, intermittent or constant, for at least 1 month, not known to be due to a condition unrelated to HIV infection
- Cryptococcal meningitis
- Pulmonary or extrapulmonary tuberculosis
- Kaposi's sarcoma
- Neurological impairment that is sufficient to prevent independent daily activities, not known to be due to a condition unrelated to HIV infection (for example, trauma or cerebrovascular accident)
- Candidiasis of the oesophagus (which may be presumptively diagnosed based on the presence of oral candidiasis accompanied by dysphagia)
- Clinically diagnosed life-threatening or recurrent episodes of pneumonia, with or without aetiological confirmation
- Invasive cervical cancer

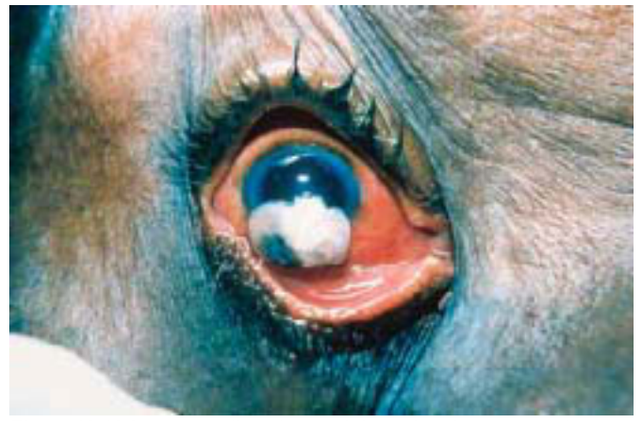


Figure 10.10 Squamous cell carcinoma of the conjunctiva: an unusual cancer, strongly associated with HIV infection. Its incidence has increased markedly in Uganda and Rwanda (courtesy of Dr Keith Waddell)

Box 10.3 Proposed WHO staging system for HIV infection and disease

Clinical staging

Patients with HIV infection who are aged 13 years are clinically staged on the basis of the presence of the clinical condition, or performance score, belonging to the highest level

- Clinical stage 1: asymptomatic or persistent generalised lymphadenopathy; performance scale 1 (asymptomatic, normal activity)
- Clinical stage 2: weight loss <10% body weight; minor mucocutaneous manifestations, varicella zoster within the last five years, recurrent upper respiratory tract infections (bacterial sinusitis); performance scale 2 (symptomatic but normal activity)
- Clinical stage 3: weight loss > 10% body weight, unexplained chronic diarrhoea > 1 month, unexplained chronic fever > 1 month, oral candidiasis, oral hairy leukoplakia, pulmonary tuberculosis within the past year, severe bacterial infections; performance scale 3 (bedridden < 50% of day during the last month)
- Clinical stage 4: most other CDC AIDS-defining diseases (but not pulmonary tuberculosis); performance scale 4 (bedridden > 50% of day during the last month)

Clinical/laboratory classification

Laboratory axis			Clinical axis			
	Lymphocytes or ($\times 10^6$/l)	CD4 ($\times 10^6$/l)	1 Asymptomatic	2 Early	3 Intermediate	4 Late
A	>2000	>500	1A	2A	3A	4A
B	1000–2000	200–500	1B	2B	3B	4B
C	<1000	<200	1C	2C	3C	4C

Treatment of HIV-infected people in developing countries

General approach

The general approach to treatment in developing countries should ideally be no different from that in the industrialised world, but is hampered by lack of infrastructure and resources for diagnosis and treatment. As for other diseases in resource-poor countries, treatment must often be decided on the basis of very limited information. Patients should be counselled about HIV infection and prevention of its transmission.

Treatment of opportunistic infections

Specific opportunistic infections should be treated as recommended (see chapter 9). Patients with some common diseases, such as tuberculosis and pneumococcal infection, usually respond well to standard treatment. Toxoplasmosis also responds well to treatment if diagnosed early, but as with many HIV-related diseases, is likely to relapse unless maintenance treatment is taken. In situations where precise diagnoses cannot be confirmed, a syndromic approach may be more practical. Individual symptoms such as diarrhoea or prurigo should be treated symptomatically if no treatable cause can be identified.

Prevention of opportunistic infections

Clinical trials have demonstrated the efficacy of preventive regimens against tuberculosis (for example, using isoniazid) among HIV-infected people in developing countries, and co-trimoxazole has been shown to reduce morbidity and mortality among HIV-infected individuals in Côte d'Ivoire. Both these interventions are now recommended by UNAIDS but have yet to be introduced on a large scale, particularly in the poorest countries, and will need to be evaluated under operational conditions. Pneumococcal polysaccharide vaccine was recently found to be ineffective in preventing invasive pneumococcal disease among HIV-infected people in Uganda; conjugate pneumococcal vaccines may be more effective, and are currently being investigated in children.

Antiretroviral therapy

Antiretroviral therapy is currently available to only a very small minority of HIV-infected people in developing countries. As

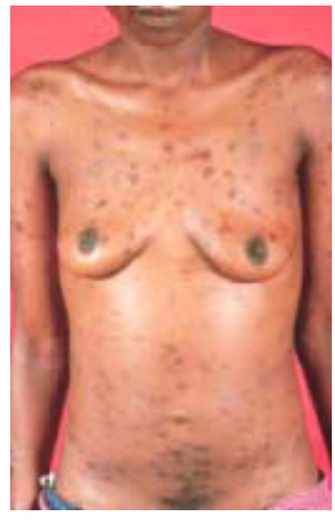


Figure 10.11 Kaposi's sarcoma: multiple skin nodules and plaques (courtesy of Dr AC Bayley)

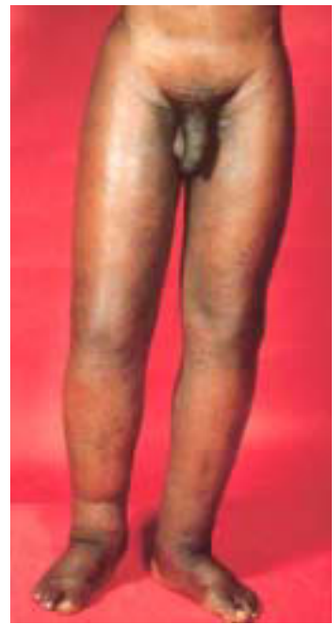


Figure 10.12 Kaposi's sarcoma: bilateral leg oedema and inguinal lymph node enlargement, without any evident skin lesions (courtesy of Dr AC Bayley)

antiretrovirals become less expensive, they will inevitably become more widely available and used. However, widespread implementation poses huge challenges in resource-poor countries, including identifying HIV-infected people before the stage of terminal disease; monitoring the response to therapy; continuity of drug supply; adherence, especially to complex regimens; and managing treatment failures. A priority will be to minimise the development of antiretroviral drug resistance by using rational and effective regimens, and maximising continuity of treatment and adherence. Resistance is probably inevitable unless triple drug regimens are used. Some populations will be easier to reach through existing infrastructures, such as occupational health schemes; tuberculosis programmes could also potentially be built upon if more resources were available.

Public health priorities

In response to the global epidemic of HIV/AIDS, WHO established the Joint United Nations Programme on HIV/AIDS (UNAIDS) in 1994. HIV/AIDS is the only disease for which a joint programme between several United Nations agencies has been established.

Essential components of public health programmes for HIV/AIDS

The key elements of public health programmes for HIV/AIDS are listed in Box 10.4. The global response to the HIV pandemic is based on fundamental principles, but HIV/AIDS prevention and control is implemented, successfully or unsuccessfully, at the local level. Involvement of heavily affected communities and non-governmental organizations has been crucial to a successful response. Key requirements of programmes are prevention of new infections, as listed in Box 10.4: youth (both in- and out-of-school), especially young women, are an important target group for HIV awareness and life-skills training. Preventing mother-to-child transmission is also important: antiretroviral drugs are most cost effective when used for this purpose, and effective and safe strategies for the reduction of transmission via breast feeding are also needed. Other important areas include prevention of discrimination and assurance of confidentiality for HIV-infected people, integration of sexually transmitted diseases control into HIV/AIDS prevention activities, and provision of services for HIV/AIDS care. Because of the close interrelationship between HIV/AIDS and tuberculosis, some countries have integrated tuberculosis and HIV/AIDS control activities.

Box 10.4 Essential components of HIV/AIDS programmes

Prevention of new infections
- Reduce sexual transmission
  - Awareness and life-skills education, especially youth
  - Condom promotion
  - STD control, including for commercial sex workers
  - Partner notification
- Blood safety
  - HIV testing of transfused blood
  - Avoid non-essential blood transfusion
  - Recruitment of safe donor pool
- Interventions to reduce transmission among injecting drug users (where necessary)
- Reduce mother–child transmission
  - antiretroviral therapy
  - avoidance of breast feeding (where safe): consider replacement feeding, or early weaning

Surveillance for HIV infections and AIDS

Voluntary counselling and testing

Mitigation of HIV-related disease
- Rational approach to care for HIV-related disease, especially tuberculosis
- Appropriate preventive therapies

Mitigating social impact
- Minimising stigma: respect for confidentiality, protection against discrimination
- Care for AIDS orphans

11 Injection drug use-related HIV infection

RP Brettle

Introduction

A variety of important medical problems, both infective and non-infective in nature, are associated with injection drug use (IDU) including the blood-borne viruses such as HIV, hepatitis B (HBV) and hepatitis C (HCV), all of which may be transmitted via the sharing of injection equipment. Consequently the medical care of patients using drugs requires a knowledge of both drug- and infection-associated conditions. The use of recreational drugs either occasionally or continually should not be a bar to or be used as a means of discriminating against access to health care in the UK as has been alleged recently. This right of access was explicitly addressed in the updated "*Guidelines on Clinical Management, Drug Misuse and Dependence*" published by HMSO in 1999. This report stated that:

- Drug misusers have *the same entitlement* as other patients to the services provided by the NHS.
- It is the responsibility of *all doctors* to provide care..., *whether or not the patient is ready to withdraw from drugs*.
- This should include *the provision of evidence-based interventions*, such as hepatitis B vaccinations, and providing harm minimisation advice.

Although combination therapy for HIV is not specifically mentioned in this document, because it is an evidenced based intervention the same principles apply.

Medical care systems

The difficulties of engaging drug users for medical care should not be underestimated. There are some particular characteristics of IDU that it may be helpful to be aware of, and the details will vary with geographical location (Box 11.1).

Drug users usually require a substantial supply of money to fund their addiction 'habit', which in itself results in other problems.

Not surprisingly the problems and illegality associated with the use of recreational drug use is associated with a number of difficulties for any health service in delivering medical care for drug users. For the health service these numerous crises, whether social, financial, legal, etc., lead to the impression of a chaotic lifestyle; in reality hospital appointments usually have a fairly low priority because of the enormity of their problems.

The social effects of HIV infection are similar for all risk groups – the infection effectively impoverishes the patient; however in the case of drug users these effects may be a little more dramatic.

More importantly the inability to fund a drug habit can have important consequences for a health service which are often not appreciated:

- A need to find additional sources of income – benefits fraud, drug dealing, hospitalisation (save money on food, etc.) – all of which increase the pressure on the NHS to prescribe addictive drugs (which may be greater than actual habit in order to provide additional funds).
- The physical weakness and mental slowing leads to peer victimisation.
- Practical problems such as problems with visitors, unexplained absences from ward, frequent self-discharge,

Box 11.1 General characteristics of IDU in UK

- Mainly an illegal activity
- Male dominated (10–30% females)
- Usually involves the young and initially healthy – 2 years before any contact with NHS
- Do not seem to be particularly health conscious
- Have often spent time in prison (up to 70%)
- Tend to have a crisis lifestyle
- Are often associated with violent or unpredictable behaviour, which in part is related to an excess of or withdrawal from recreational drugs and is more often than not related to problems with their peers

Box 11.2 Specific problems relating to drug addiction

- Expensive to maintain (£100–£200/day) and may be funded by a variety of means such as:
  - theft – car crime, burglary
  - fraud – credit and cheque cards, DSS benefits
  - drug dealing
  - prostitution – male or female
- Are often short of money
- May need to avoid police "warrants"
- Live for today (shortened "future time perspective")
- May require 3–4 shots per day for opiates and every hour for cocaine

Box 11.3 Recreational drug use and health care

"Unreliable" individuals with chaotic type of lifestyle

- Irregular attendances – missed appointments, wrong day, frequent self-discharges from ward
- Suspect motives for many of symptoms
- Unexplained absences from wards
- Disruptive visitors
- Day/night reversal
- Self-medication and drug dealing
- Theft from other patients and staff
- A threat for patients and staff
- Attention seeking, demanding of time and often noisy
- Aggressive behaviour both verbal and physical
- Utilise a number of offensive weapons – knives, guns

Come with a variety of staff "attitudes"

- "Others more deserving" of care
- "Manipulative" of staff
- "Dangerous"
- "Frightening"
- "Upset other patients"
- "Not enough time"
- "Never change"

day/night reversal, theft of hospital property (related to a falling income), noise, manipulation of staff or other patients and attention-seeking behaviour.
- Increased frequency of verbal and physical abuse of both staff and other patients.

The net result may be an inability to cope in the community or the hospital, resulting in frequent precipitous admissions and discharges – "revolving door" type admissions with considerable frustration for patients, relatives and staff.

Without a *modified* healthcare system which understands and considers these problems, drug users have a tendency to record a high default rate in terms of attendance or frequent discharges from hospital units. The aim of an IDU service should be to initiate and maintain contact primarily in order to deliver health care and health education. The initiation and maintenance of that contact may require a variety of initiatives as described in box 11.5.

A system of providing both drug services as well as medical care from the same site by the same doctors seems to be an efficient model of care for drug users, whereas a system of delivering care via two distinct physical sites (one for drugs and one for physical care) is less efficient and seems to provide either a poor medical and/or a poor HIV service.

The dependency needs of IDU-related HIV are both physical and psychological. The physical care varies from mildly ill to high dependency, whilst on the psychological side it may vary from being entirely well to toxic confusional states, obsessive-compulsive states, anxiety and agitation as well as frank psychosis. The differential diagnosis is extensive and admission is commonly required to exclude the diagnosis of an organic psychosis. The time that patients may remain in a medical unit varies from a few days to over a month and this mixture of serious physical and mental ill health is rarely found in other areas of medicine. There is also the danger of fire from careless cigarettes, since the majority of patients smoke heavily and consume excessive amounts of sedative drugs. Because addiction to cigarettes seems to be greater or at least equal to opiates, it appears impossible to enforce a total no smoking policy for the inpatient areas if the policy of maintaining contact with patients is to be followed. In addition to the difficulties described above, there are also the problems of nursing individuals in some form of isolation. The requirement for cubicles is high as a consequence of an increased risk of infectious agents associated with HIV and IDU, such as tuberculosis. There is also an increased need for privacy because of mixed sexes (one third are female), mixed risk groups (homosexuals and drug users) and disturbed patients.

Management strategies for IDU-related HIV

There are a number of strategies which may be adopted in order to cope with IDU-related HIV admissions, including higher staffing levels, avoidance of high occupancy levels and continuity of care by both nursing and medical staff. Other issues are as listed overleaf:

Box 11.4 Social effects – physical and mental slowing

- More vulnerable to exploitation from peers
- More likely to get caught by the law
- Reduced income – "criminal unemployment"
- Increased demands on NHS to replace the missing funds
  - pressure for prescribed drugs
  - pressure for access to state benefits

Box 11.5 Initiatives for initiation and maintenance of contact with drug users

- Needle exchange
- Methadone prescribing
- Social provisions such as helping with housing
- Medical care

Box 11.6 Specific problems of IDU-related HIV for a health service

- Mixture of physical and psychological dependency
- Frequent security and fire incidents
- Increased need for cubicles
- Need for increased staffing levels

Box 11.7 Management strategies for IDU-related admissions

- Continuity of care from medical and nursing staff
- Increased numbers of nursing and medical staff
- Response to violence is to call police
- Tight control on drug prescribing
- A written smoking policy which is given to every patient on admission
- Coordination of drug prescribing between different agencies around admissions and discharges
- Increase contact with healthcare system gradually
- Clear guidelines and policies which all staff sign up to
- Such policies to be supportive and caring and based on health and safety principles
- Avoid situations leading to confrontation
- Avoid withdrawals in ward area but make it clear that no guarantee of increases on discharge
- Awareness of need for relief of pain and psychological distress

Box 11.8 Strategies for coping with IDU-related HIV admissions

- Violent activities, assaults, etc. are managed by calling the police. Patients need to be informed that recreational drug use is just as illegal in hospital as out, that other patients may complain to the police via a local drugs hot line if they observe illegal drug dealings or use on the wards and this could result in police raids.
- There is tight control of prescribed drugs early on in the disease process with accompanying harm reduction messages. The message concerning prescribing is that its function is to provide a safety net for the physical discomfort of addiction rather than to provide a free buzz or "stone".
- There is a gradually increasing level of contact between hospital and patient over time which allows the service to get used to the behaviour of patients and for them to get used to the hospital's routines. This is one form of re-socialisation for the individual with problem drug use.
- The aim is to provide a supportive and caring environment associated with firm discipline over misbehaviour and illegal activities. Wherever possible the rules are based on health and safety principles rather than moral or legal ones. Injecting in the hospital is forbidden because of the dangers to staff. Similarly being stoned is discouraged because of the increased risks of hypostatic pneumonia or fire hazards from concomitant smoking.
- A written smoking policy is provided to every patient on admission. It is based on health and safety principles and the need to reduce the danger of fires for everyone's sake.
- The regime for outpatient appointments is reasonably flexible (anytime on a set day) in order to allow for missed appointments. However the patients are made aware of the need for some structure in the system by making the patient aware of the hospital's limitations.
- The law relating to the prescribing of drugs such as methadone is explained in verbal and written instructions.
- Confrontation is generally avoided in situations that cannot be resolved. This means adapting the regime or removing the patient from the environment that they find difficult; this may require us either to allow the patient to self-discharge or if necessary to discharge the patient from the unit. The patients are always offered an outpatient appointment if they leave the hospital or are discharged.
- Treatment in the community is often arranged in order to avoid long spells in hospital. Increased drug taking in hospital or difficult behaviour is often a symptom of boredom and much can be done to avoid this, for instance by providing satellite TV or computer games.
- Coordination of substitute prescribing with other carers is very important to avoid double prescribing via hospital admissions. Careful prescribing on discharge is also required to avoid similar problems in the community.
- Illegal drug use in the ward requires careful discussion in order to arrive at a compromise over the amount of drugs prescribed and the amount of drugs used illegally. Generally this compromise is achieved by suggesting that the dose of prescribed drugs will be reduced until a satisfactory level of consciousness is achieved that reduces the fire risk and the necessity for increased nursing observation. Such reductions result in increased cost for the patient in terms of the need to purchase black market supplies of drugs.
- On occasions, in order to deliver inpatient care, illegal or extra drug use needs to be covered. In such situations the patients are warned that this does not imply any sort of contract or obligation for increased doses on discharge. If the admission is prolonged then an offer of detoxification to the doses prescribed would be made.
- Obvious withdrawal symptoms (alcohol or opiates) during a physical illness would be covered with extra doses of opiates or short courses of benzodiazepines (diazepam or chlodiazepoxide). Agitation from recent stimulant use would also be covered for inpatients. The prime aim would be to reduce the chance of agitation and disturbed behaviour in the wards.
- Fear of pain may be a major problem for drug-using patients. We have generally used either a subcutaneous infusion of opiate over and above maintenance drugs or the use of oral slow-release morphine preparations. Provided observation reveals that the patients are not excessively sedated from a health and safety point of view there is no upper limit on the doses employed to relieve pain.
- There may be concern amongst the staff over the level of prescribing of sedative drugs. The nursing staff need to have confidence in the medical management policy relating to sedative and pain control prescribing. A number of patients, particularly drug users, request high levels of sedation prior to death and this may cause concern amongst a number of staff, medical and nursing as well as relatives.

Management of IDU-related problems

Controlled Drug prescriptions

A working knowledge of the regulations surrounding Controlled Drug prescriptions is important when managing drug users. Attention to detail when a patient is admitted is imperative if centres are to avoid the problems of double prescribing. The front pages of the BNF give exact guidelines on how to prescribe Controlled Drugs legally and additional information is available via the updated *Guidelines on Clinical Management, Drug Misuse and Dependence* published by HMSO in 1999.

The medical effects of recreational drugs

Carers need to have a working knowledge of the effects of recreational drugs and equivalent doses of drugs (methadone or diazepam) if patients need to be temporarily covered for the effects of withdrawal. Tables of equivalence for opiates and benzodiazepines can be found in *Guidelines on Clinical Management, Drug Misuse and Dependence* (HMSO 1999). The differential diagnosis in a patient with IDU-related HIV is extensive and requires consideration of both infective and non-infective disorders. These are summarised in Table 11.1.

When patients are admitted with respiratory problems there is the dilemma of how to manage opiate prescribing.

- For those patients with mild respiratory depression a discussion over a temporary reduction in oral drugs by around 10–20% or splitting the daily dose into 3 or 4 doses may suffice.
- In those with more severe respiratory depression rapid improvement in pulmonary function is required. However if the opiate withdrawal is excessive as with intravenous bolus injections of naloxone, the patient may become disruptive with loss of venous access.
- The preferred solution is a naloxone infusion (2 mg in 500 ml perhaps starting at around l0 ml per hour) to achieve an acceptable improvement in respiratory rate (and therefore oxygenation) without too great an increase in physical arousal. The aim is to improve oxygenation rather than induce withdrawal from opiates. This improved oxygenation can be assessed by respiratory rate, oxygen desaturation or arterial blood gases. Such an infusion may be required for up to 48 hours in those on methadone because of its relatively long half-life compared to other opiates. In the event of a lack of venous access then regular small doses of intramuscular nalaoxone (0.2 mg i.m. every 1 hour initially) can also be employed to maintain oxygenation.

Excessive doses of benzodiazepines also produce drowsiness and/or coma which can usually be managed by simple supportive therapy with care over respiratory rate, etc. In extreme cases it is possible to utilise the antagonist flumazenil but there is a danger of inducing fits in those on chronic long-term doses. *It is therefore preferable to reverse the opiate element first (with an infusion of naloxone) before resorting to flumazenil.*

Opiate withdrawals should be considered in any agitated patient known to be on opiates, particularly those who have recently commenced drugs that induce liver enzymes such as rifampicin, rifabutin, phenytoin, etc.

Table 11.1 Medical (non-infection) problems of drug use

Problem	Medical complications
Drug effects	
Excess opiate	Narcosis, coma, small pupils, respiratory depression, aspiration pneumonia, and *rhabdomyolysis secondary to pressure*
Opiate withdrawal	Mild "URT" (sweating, coryza, lacrimation), pupillary dilatation, insomnia, nausea, vomiting, diarrhoea, lethargy, muscle weakness, myalgia, muscle twitching, tachycardia and hypertension
Excess cocaine	Apprehension, dizziness, syncope, blurred vision, dysphoric states, paranoia, confusion and aggressive behaviour, seizures, coma, hyperthermia, respiratory depression, apnoea, sudden death, spontaneous rhabdomyolysis
Excess amphetamine	Headaches, anorexia, nausea, tremors, dilated pupils, tachycardia and hypertension
Stimulant withdrawal	Sleepiness, lethargy, increased appetite, food binging, depression or even suicide
Trauma	
Frequent injecting	Track marks and skin scars, lack of veins and thrombophlebitis, deep venous thrombosis, persistent peripheral oedema, venous stasis and ulcers secondary to chronic venous obstruction
Misplaced injections	Arterial damage and insufficiency with secondary tissue damage, muscle compartment syndrome and traumatic rhabdomyolysis, false aneurysms and pulmonary emboli, traumatic neuropathy
Immunology	
IDU	Enlarged nodes, elevated IgM, false positive syphilis serology
Endocrinology	
Opiate use	Increase prolactin levels and gynaecomastia, amenorrhoea (may be secondary to weight loss)
Cannabis	Oligospermia, impotence and gynaecomastia
Neurology	
Stimulants	Psychosis, depression, cerebral infarcts and haemorrhages (CVAs)
Chronic use of benzodiazepines or barbiturates	Brain damage
Cardiology	
Cocaine	Cardiac arrhythmias such as sinus tachycardia, ventricular tachycardia and fibrillation as well as asystole, myocardial infarction, severe hypertension
Adulterants of illicit drugs, for example quinine	Cardiac arrhythmias and death
Tricyclic antidepressants	Cardiac arrhythmias and death
Cannabis	Sinus tachycardia and postural hypotension
Pulmonary	
Inhaled cocaine	Excessive use of Valsalva – spontaneous pneumomediastinum and pneumopericardium
Excess sedatives or stimulants	Respiratory depression, coma and pneumonia
Opiate withdrawals	Mild "URT"
Stimulant use, for example cocaine	Tachypnoea
Opiates or cocaine	Pulmonary oedema
Hepatitis B	Polyartertis nodosa
Foreign body emboli, (particles injected intravenously) for example talc granulomas	Pulmonary hypertension (and right heart failure) abnormal pulmonary function eg reduced DCO, restrictive defect due to interstitial lung disease

Medical problems of HIV-infected drug users

The extent of IDU-related conditions requires consideration of not only the clinical features of IDU but also the associated medical conditions such as HIV since confusion may arise as to the aetiology of specific symptoms. (See Box 11.9)

Pre-AIDS deaths

The phenomenon of pre-AIDS death amongst HIV-infected drug users was described soon after the onset of the AIDS epidemic in the USA. The IDU non-AIDS death rate was 2.5/100 person-years in Edinburgh (compared to 0.9/100 person-years in other risk groups), 3.8/100 person-years in Amsterdam and 2.6/100 person-years in New York. In Edinburgh 20% of pre-AIDS deaths were expected and related to conditions not ostensibly related to HIV. Liver disease was the single commonest cause of these deaths, accounting for 75% of expected pre-AIDS deaths or 25% of all pre-AIDS deaths, and is presumably related to the heavy co-infection with hepatitis B and C.

Respiratory infections

A review of pneumonia in all HIV positive patients suggested an increased annual incidence of bacterial pneumonia; 97–290 per 1000 compared to 21 per 1000 for HIV negative individuals. IDU-related HIV patients also have an overall higher incidence of bacterial infections; 12% (mortality of 2.2%) compared to 3% (mortality of 0%) in HIV negative drug users. The overall rate of bacterial sepsis in Edinburgh drug users was 7.0 per 100 person-years whilst in the Bronx cohort of drug users the rate was 8.0 per 100 person-years. In Spain 60% of the pneumonias in the HIV-infected patients occurred before a diagnosis of AIDS, in 55% of patients the problem was recurrent and the mortality was increased for HIV-infected patients (19% vs. 4%). *Streptococcus pneumonia* and *Haemophilus influenzae* were the commonest organisms involved in the pneumonias. Additional susceptibility factors for drug users may be the use of opiates themselves because they are known to depress the cough reflex as well as the immune system. Latterly it has been suggested that the inhalation of drugs as well as the injection of drugs may increase the risks of bacterial pneumonias. The odds of developing pneumonia were twice as great for those reporting smoking cocaine, crack cocaine and marihuana and over 20-fold increased for those also having prior PCP and a low CD4. The effects of tobacco were not examined since all patients utilised this drug. The incidence of tuberculosis is much higher in HIV-infected drug users than in other risk groups outside the tropics, or in HIV negative drug users. In the USA, most patients with AIDS and tuberculosis have been drug users. One study showed a prevalence of 15% in drug users with AIDS but only 4% in other risk groups within a New York hospital. In New York the rate of tuberculosis was 4% among HIV positive drug users compared to 0% in HIV negative drug users. The 36% increase in reported cases of tuberculosis between 1984 and 1986 has been largely ascribed to infection amongst HIV positive drug users.

Hepatitis

Drug users with a history of IDU are highly likely to be co-infected with hepatitis B and C viruses (anywhere from 40% to 100% depending on location).

Anti-HIV drugs have long been associated with hepatitis and it is uncertain at present whether modern combination therapy for HIV will have a deleterious effect on those

Table 11.1 *continued*

Smoking of tobacco, heroin, marijuana	Abnormal pulmonary function for example reduced DCO, COAD
"Snorting" stimulants	Chronic rhinitis, rhinorrhoea, anosmia, atrophy of the mucosal membranes, ulceration and perforation of the nasal septum
"Snorting" opiates	Recurrent sinusitis

Box 11.9 Associated medical problems of IDU

- Lymphadenopathy is associated with both HIV and the injection of foreign materials.
- Fatigue, lethargy and excessive sweating are features of HIV as well as mild withdrawal from opiates.
- Diarrhoea, a common presentation of early symptomatic HIV (CDC stage IVA), is also a common symptom of opiate withdrawal.
- Weight loss and fever are both key symptoms of the constitutional symptoms associated with HIV (CDC stage IVA), infection with mycobacteria as well as heavy opiate or stimulant (amphetamines or cocaine) use.
- Epileptic seizures require consideration of cerebral toxoplasmosis in HIV, the intermittent use of benzodiazepines or even hepatic encephalopathy.
- The excessive use of cannabis and benzodiazepines interferes with memory and other cognitive functions in a similar manner to HIV as does frequent head injuries. Thus early dementia is difficult to detect in current drug users especially since reducing drugs will also help the dementing patient to improve function in relation to activities of daily living.
- Syncopal attacks in HIV may be associated with an autonomic neuropathy or a failing adrenal cortex but it is also associated with the use of antidepressant tricyclic drugs such as amitriptylene.
- Jaundice may be a result of acute or chronic hepatitis B or C infection, excessive alcohol ingestion or a side effect of the treatment of mycobacterial infections in HIV.
- Lastly shortness of breath and a persistent cough are common early symptoms of *Pneumocystis carinii* pneumonia (PCP) but can occur with endocarditis, bacterial pneumonia, excessive smoking, recurrent bronchitis and obstructive airways disease.

Greater detail can be obtained from the web site of the Regional Infectious Diseases Unit, Western Infirmary, Edinburgh (www.med.ed.ac.uk/ridu/History.htm).

Box 11.10 Co-infection of HIV and hepatitis viruses

Hepatitis B

- 10% of drug users will be carriers of hepatitis B
- Re-emergence of carriers (HbsAg) with CD4 counts of < 200 cells/μl may occur

Hepatitis C – data contradictory and may vary with risk groups because of length of HCV infection which is often unknown

- HCV RNA levels rise with falling CD4 counts
- Increased progression of both HIV and HCV disease reported in haemophiliacs
- Studies on drug users have reported no change of HIV progression
- Increased rate of HCV progression also reported in drug users

co-infected with HCV or not. There are reports in the literature of therapy improving, worsening or having no effect on concomitant HCV. Thus despite a greater risk of hepatotoxicity in HIV/HCV co-infected patients on highly activated retroviral therapy (HAART), this is not a reason to withold therapy from HCV HIV co-infected patients but rather such patients require more carefully monitoring. There is also the problem of additional hepatotoxicity associated with the use of antituberculous drugs in HIV/HCV co-infected individuals.

Although treatment for HCV is now available in the form of interferon and ribavirin, tolerance and interactions with HIV therapy are likely to be problematical since ribavirin has been reported to interfere with the phosphorylation of nucleosides such as zidovudine.

HIV dementia and encephalitis

HIV/AIDS is unusual in that it combines both immunological, neurological and psychiatric disorders and as a consequence patients may develop a variety of disabilites ranging from wasting disorders, severe pain, neurological dysfunction such as paralysis or cognitive impairment and psychological symptoms. These combinations of problems may result in considerable problems for both patients and carers. Autopsy studies in Edinburgh (prior to modern antiretroviral therapy (ART)) have shown that as many as 60% of IDU-related HIV patients have evidence of HIV encephalitis although only 6–7% have frank dementia.

AIDS

As for AIDS itself, little variation between the risk activities with regard to presentation has been reported. In the USA, figures available to the Centers for Disease Control show that conditions such as Kaposi's sarcoma are unusual in the absence of homo/bisexuality. In drug users, Kaposi's sarcoma, cytomegalovirus and chronic cryptosporidiosis are all significantly less common than for all other risk groups notified with AIDS, while PCP, tuberculosis, oesophageal candidiasis and extrapulmonary cryptococcosis are more common.

Progression from HIV to AIDS

No evidence has been found for a role of alcohol, opiates or other psychoactive drugs in accelerating the progression of immunodeficiency in HIV-seropositive homosexual and bisexual men. The major factors identified in the progression of HIV appear to be age and HLA type. Al B8 Dr3 and Bw 35 are all associated with more rapid progression whilst B27 is associated with slower progression to AIDS and death.

Survival after the development of AIDS

Without treatment, in general, around half of patients with AIDS survive for one year but only a fifth for three years and the median survival time is around 18–20 months. In Edinburgh the one- and three-year post-AIDS survival rates were 66% and 25%. Older age at AIDS diagnosis and HLA type A1 B8 DR3 were associated with shorter survival. Whilst it had been generally assumed that the survival for injection drug users with AIDS would be shorter than for other risk groups, several studies including our own counter this presumption.

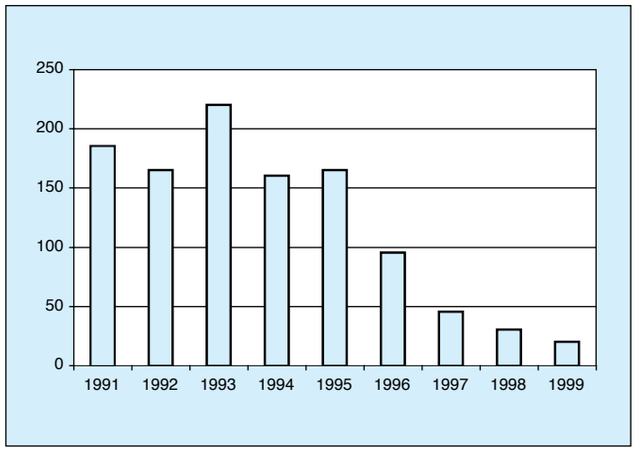


Figure 11.1 Opportunistic events/100 patient-years: Regional Infectious Diseases Unit, Edinburgh

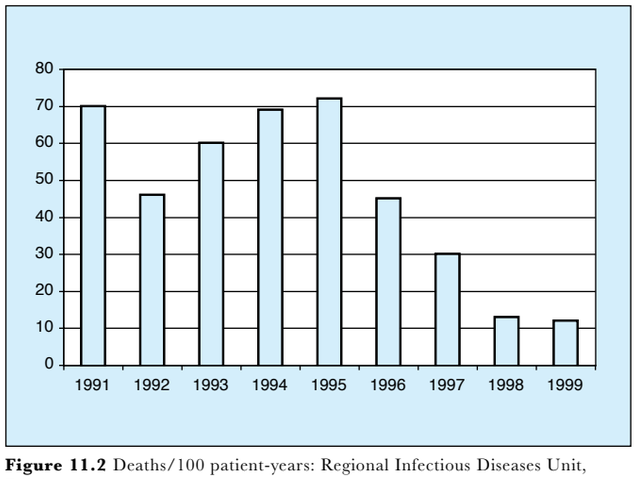


Figure 11.2 Deaths/100 patient-years: Regional Infectious Diseases Unit, Edinburgh

Problems of pain management in IDU-related HIV

The investigation of pain in IDU-related HIV is a major problem for carers, particularly since it is a useful complaint to increase the size of opiate prescriptions. Considerable experience is required in both the investigation and management of pain in order to avoid ever larger prescriptions.

The commonest problem in the management of pain in drug users is that of insufficient doses as a consequence of existing high levels of opiate use and/or disagreement from carers over whether pain is real or being treated adequately. Self-medication is a common problem which increases the uncertainties around prescribing. The use of inappropriate drugs such as Diconal or Temgesic should be avoided because these drugs are highly sought after as recreational drugs in the community.

Our own practice has been to utilise increasing doses of existing drugs such as methadone, oral solutions of morphine, slow-release morphine preperations such as MST (although this also has some street value since it can be injected), morphine/diamorphine solutions via subcutaneous infusions or fentanyl patches.

There is the additional problem of providing adequate sedation during procedures. Patients may require unusually large doses of medazolam as a consequence of their regular intake of benzodiazepines. If adequate doses of medazolam are not used then there is no loss of memory for the procedure which can be quite distressing for the patients. Alternatively, particularly if the patient uses illicit benzodiazepines or opiates, excessive sedation occurs with even quite small doses. As with the elderly, some HIV/AIDS patients also exhibit an unusual sensitivity to neuroleptics such as carbamazepine or antidepressants, possibly because of the concomitant presence of HIV encephalopathy. Care is therefore required in the introduction of such drugs for pain control.

Box 11.11 Common causes of pain in drug users

Dental caries
Traumatic neuropathies
Abdominal pain
- Constipation
- Cholecystitis
- Appendicitis
- Chronic hepatitis
- Lymph node enlargement
  - MAI
  - Lymphoma

Antiretroviral therapy

It is perfectly possible to treat drug users with antiretroviral therapy although there are a number of simple difficulties such as venous access for monitoring of therapy.

When considering combination therapy for recreational drug users a number of important principles need to be understood by the drug users. Whilst these may seem obvious to ourselves this is not the case for the patients.

A number of groups have been exploring drug regimens thought to be particularly suitable for drug users, usually because they provide the possibility of a once-daily regimen, and therefore the option of employing directly observed combination therapy (DOCT) at a suitable location. Whilst DOCT may be offered to a patient as an option it should perhaps be seen as a means to an end rather than as a long-term solution.

Box 11.12 Problems of antiretroviral therapy in drug users

- Regular venous access required – consider external jugular rather than femoral artery
- Improved health may encourage a return to IDU – continue with harm reduction strategy

Drug interactions

Drug interactions are an ever present problem with modern antiretroviral therapy for all patients, but even more so for those taking recreational drugs where there is the ever-present possibility of serious increases in the levels of pharmaceutical or recreational drugs. Of course from the patient's point of view reduced levels of recreational drugs is also an important problem.

Box 11.13 Important principles in modern antiviral combination therapy

- Intermittent combination therapy is a major disadvantage because of the development of resistance which will impair future therapy choices
- Almost total (95%) adherence is required for the best chance of long-lasting success (undetectable viral load)
- Increasing the number of drugs used in combination therapy does not increase "wellness", it simply increases the chance the regimen will be successful for a longer period
- However more antiviral drugs increase the chance of an adverse drug related event
- Intermittent recreational drug use is more dangerous and difficult to adjust for than regular recreational drug use (time and patience required by both patient and doctor) in terms of interactions with combination therapy

The most extensively investigated interactions are with methadone since it is commonly used long-term for heroin substitution. Little other investigational work has been undertaken possibly because of the difficulties of working with illegal drugs such as cocaine or amphetamine and some lack of interest on the part of pharmaceutical companies. One proviso when discussing interactions is our relative lack of knowledge of methadone levels and symptoms of withdrawal. Very little work seems to have been carried out in this area recently and much more is required if we are to better understand the interactions that do occur.

Other drugs used in HIV

It is important also to remember that a number of other drugs commonly used in HIV medicine such as rifampicin or phenytoin also dramatically reduce methadone levels by enzyme induction and cause problems with acute withdrawal. Increased zidovudine levels have also been reported with sodium valproate, a drug that is commonly used to control seizures.

In summary the use of recreational drugs certainly affects the choice of anti-viral drugs and possibly also the time at which therapy starts. Individual regimes to suit particular problems are important. Because of the complexity of the interactions it is important to get over to the patient how vital it is to know what drugs are actually taken rather than what drugs are prescribed. Misinformation may be fatal and they need to understand why the information needs to be accurate. This will only work of course if the patient is truly persuaded of the need for therapy. The risks of not taking therapy have to be very real and to outweigh the risks of the therapy – which after all in the case of ecstasy and ritonavir could be sudden death – not a very good outcome measure for combination therapy. For drug users the risks of disease may not outweigh therapy until the CD4 count is below 200 cells/microgram when the immediate risk of ill health is 20% or one in five for the next 12 months and 80% or four in five for the next three years. By comparison, the risk of a drug-related adverse event lies somewhere between 3% and 30%. At levels of CD4 count of 350 or 500 the risks of an adverse event are likely to outweigh the risk of serious HIV disease.

Despite all these difficulties, in Edinburgh with around 50% of our patients being drug users, we have managed to achieve the same reductions in opportunistic infections and deaths noted in other areas.

Thus recreational drug use related HIV can be managed successfully via attention to drug dependence needs, social needs and the medical care needs.

Box 11.14 Substitute recreational drug therapy and interactions with HAART

NRTI (Nucleoside Reverse Transcriptase Inhibitors)
- Zidovudine
  - ZDV levels increased by opiates (AUC increased x 2)
- Stavudine and Didanosine
  - Absorption decreased by co-administration with methadone
- Abacavir
  - Rate but not extent of absorption of decreased by co-administration with methadone
  - Increased clearance of methadone – dose adjustment may be needed

NNRTI (Non Nucleoside Reverse Transcriptase Inhibitors)
- Nevirapine and efavirenz
  - AUC of methadone reduced by as much as 30% – dose adjustment may be needed
- Delavirdine
  - AUC of methadone increased – to date no reports of dosage adjustment required

PI (Protease Inhibitors)
- Ritonavir
  - No change in methadone dosage required
  - Heroin and morphine levels reduced
  - Dextropropoxyphene and pethidine levels increased
- Indinavir
  - Initially need to reduce methadone doses but after a few weeks return to previous levels
- Nelfinavir
  - Methadone levels reduced and dosage adjustment usually necessary

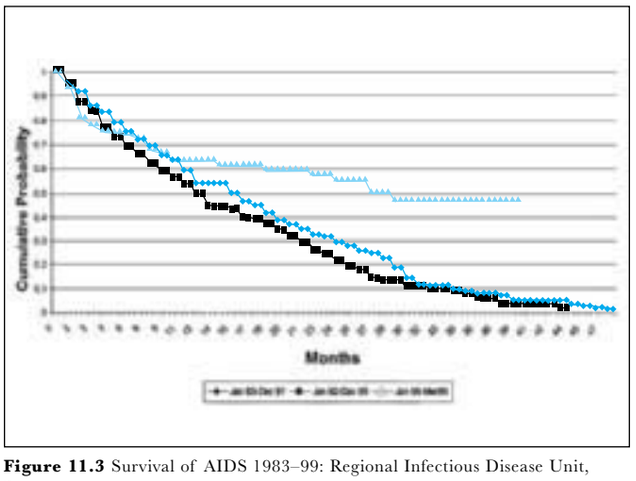


Figure 11.3 Survival of AIDS 1983–99: Regional Infectious Disease Unit, Edinburgh

12 HIV infection in children

Gareth Tudor-Williams, Diana Gibb

Epidemiological aspects

By the end of 2000, there were an estimated 1.4 million children under 15 years of age living with HIV infection worldwide and 4.3 million had died. Of these, UNAIDS estimates that 600 000 became infected during 2000 alone, over 90% via mother-to-child transmission (MTCT). Over 90% of people with HIV live in the developing world and, of these, over two-thirds live in countries in sub-Saharan Africa. Here HIV is reversing gains in child survival and significantly lowering life expectancy. Although the burden of HIV disease borne by African children is enormous, over half of the world's population live in the Asia/Pacific region. The HIV epidemic is at a much earlier stage here than in Africa and the explosive increase seen this decade is alarming.

Around half of women who acquire HIV become infected before 25 years of age and die before their 35th birthday, in the prime of their child-bearing years. As a result, by the end of 1999, the epidemic had left behind 13.2 million AIDS orphans under the age of 15 years. The difficulties that poor communities in Africa face in trying to care for this increase in children without parents is enormous and is largely dependent on existing family and social support structures already greatly affected by the AIDS epidemic.

For a vertically infected child in sub-Saharan Africa the probability of death by 12 months is estimated to be between 23% and 50%, and over 75% will not live to see their fifth birthday. HIV-infected children in Western Europe and North America contribute <1% to the total number of children worldwide living with the disease today. Over the last 5 years the epidemiology of paediatric HIV infection in affluent countries has changed as a result of a number of factors:

- First, there has been a dramatic decrease in MTCT as a result of widespread implementation of interventions for pregnant infected women and their newborns, resulting in few infected children.
- Second, new HIV infections increasingly occur either in children whose mothers acquired the disease in a country with a high prevalence of HIV, or come to light in older children who were themselves born elsewhere.
- Third, with the advent of potent antiretroviral therapy, children are living longer with HIV infection – for example, one-quarter of the 10 000 children living with HIV in the USA are now teenagers and the age of children in Europe is also increasing.

Unlinked anonymous monitoring of HIV through testing newborn dried blood spot (Guthrie) cards provides an unbiased estimate of the prevalence of HIV infection among women having live babies. This is being undertaken in the UK and covers 70% of births (see Figure 12.3). In 1999, the prevalence of maternal infection was 0.25% (1 in every 400 births) in London, compared with approximately 1 in 6000 births outside London. There has been an increase of around 30% in the number of pregnant HIV-infected women being reported in the last 2 years, which may in part reflect an increasing desire for infected women to have children in the new knowledge that the risk of MTCT is low. In London, three-quarters of seropositive

Table 12.1 End 2000 global HIV/AIDS estimates in millions

	Children	Total
People living with HIV/AIDS	1.4	36.1
New HIV infections in 2000	0.6	5.30
HIV/AIDS Deaths in 2000	0.5	3.0
Cumulative HIV/AIDS deaths	4.3	21.8

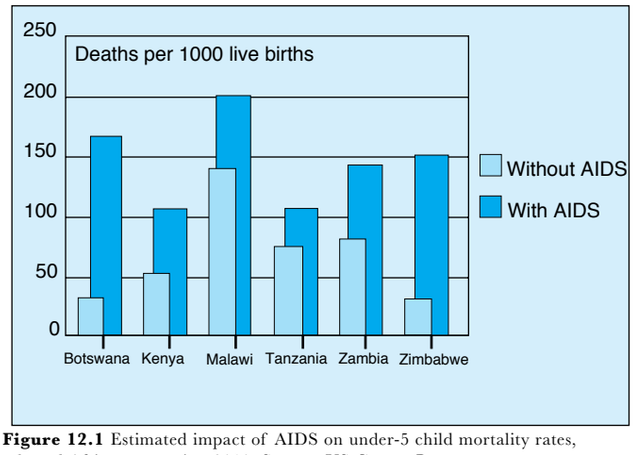


Figure 12.1 Estimated impact of AIDS on under-5 child mortality rates, selected African countries, 2010. Source: US Census Bureau

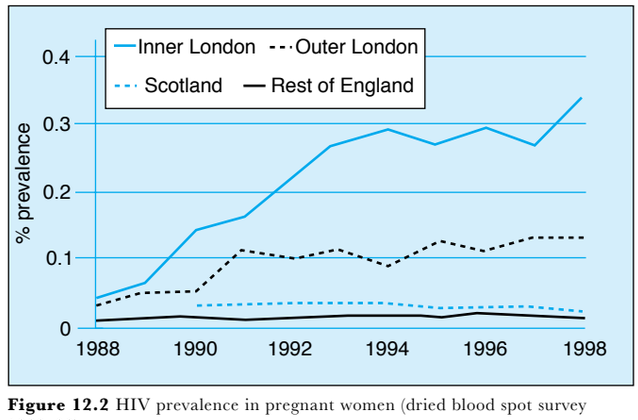


Figure 12.2 HIV prevalence in pregnant women (dried blood spot survey 1988–98)

newborns are delivered to mothers born in sub-Saharan Africa, and similar patterns are seen in European countries such as France and Belgium. In Scotland, Ireland and Southern Europe a high proportion of seropositive children are still born to women with IDU as a risk factor, but here too the proportion of women acquiring HIV from heterosexual transmission is increasing. Romania has the largest number of HIV-infected children in Europe, making up nearly half of the estimated 10 000 children living with HIV/AIDS in East and Western Europe. The majority belong to a cohort of children who were uniquely infected with HIV through contaminated blood products and needles in the late 1980s and early 1990s. Although many have died, there remain a considerable number of these children now entering their teenage years in Romania.

Box 12.1 Mother-to-child transmission of HIV infection

- HIV infection is transmitted to about 15–20% of babies born to HIV infected women *(between 1 in 5 and 1 in 6)*.
- The transmission rate doubles if a woman breastfeeds to about 30% *(1 in 3)*.
- In non-breast fed infants, approximately 70% of transmission occurs around the time of delivery
- Transmission is increased if a women has a high HIV viral load, low CD4 count and/or AIDS.
- Factors around delivery influence transmission.

Antenatal testing and mother-to-child transmission

Most observational studies estimate the risk of MTCT without interventions to be around 15–20% in Europe and the USA and over 30% in African populations. Postnatal breastfeeding doubles the overall risk of transmission and accounts for most of the difference. In non-breastfeeding women, approximately 75% of perinatal transmission occurs around the time of delivery. Other factors independently affecting the rate of transmission include the HIV viral load and CD4 cell count of the mother at the time of delivery, duration of rupture of membranes, prematurity and mode of delivery. In the last five years, the MTCT rate has been reduced to less than 2% in the USA and most European countries by the introduction of antenatal testing, highly active antiretroviral therapy (HAART) for mother's requiring therapy, use of antiretroviral therapy perinatally even if not indicated on the grounds of the mother's disease status, delivery by elective caesarean section, and refraining from breastfeeding. A recommendation that HIV testing should be offered to all women in pregnancy has been successfully implemented in many European countries, notably France, Italy and Spain where the prevalence of HIV in pregnant women was highest.

In the UK, universal offer of HIV testing during the antenatal period has been recommended in London because of the high prevalence since 1992. However, until 1999, it was recommended that antenatal HIV testing should only be offered to women considered at high risk (selective testing) outside London. An economic analysis was published in 1999 showing that a universal offer policy was cost-effective throughout the UK provided that a high uptake of testing was achieved. Department of Health guidelines endorsing this approach were published in August 1999. In low prevalence areas, up to 50 pooled samples can be tested in batches to reduce costs. During most of the 1990s, detection of previously undiagnosed HIV in pregnancy has been low everywhere in the UK. However, during 1999, and more dramatically, during the first half of 2000, there has been a marked improvement in antenatal detection rates, with about 75% of all HIV infected women being aware of their diagnosis before their baby is born in inner London and Scotland, 66% in outer London and about 50% elsewhere in the UK. In most European countries and in the USA, a marked decrease in AIDS cases reported in infancy reflects the high proportion of pregnant women receiving appropriate care to reduce MTCT.

Among UK women who either knew their HIV status before pregnancy or are diagnosed during pregnancy, MTCT rates of 2% or less are being reported among those taking up

Box 12.2 Interventions to reduce transmission from mother-to-child

- Not breastfeeding
- Antiretroviral therapy
- Elective caesarean section delivery (ie before the onset of labour or membrane rupture)
- **IMPLEMENTING ALL 3 CAN REDUCE TRANSMISSION RATES TO 2% OR LESS**

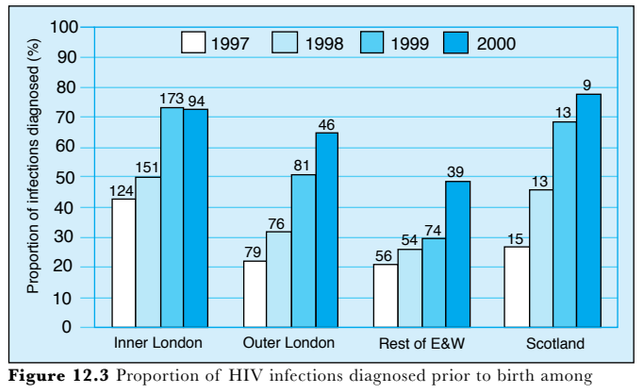


Figure 12.3 Proportion of HIV infections diagnosed prior to birth among pregnant women

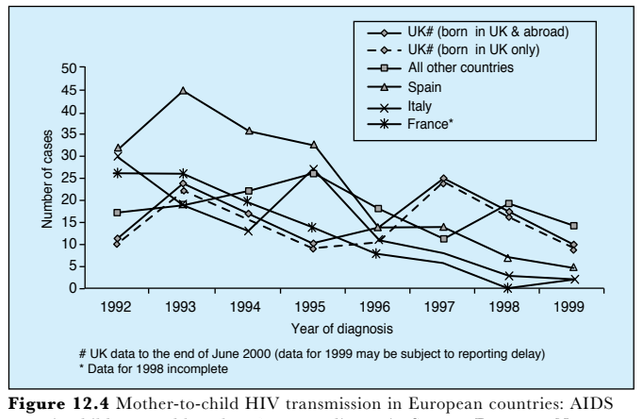


Figure 12.4 Mother-to-child HIV transmission in European countries: AIDS cases in children aged less than 1 year at diagnosis. Source: European Non-aggregate AIDS data set. June 1999. European Centre for the Epidemiological Monitoring of AIDS, Saint Maurice, France

interventions. Women taking HAART for their own disease who have undetectable HIV viral load at delivery have a very low risk of transmitting HIV to their baby. For those women not needing therapy for themselves, most guidelines recommend zidovudine and elective caesarean section (CS) delivery, which limits exposure of mother and baby to antiretroviral drugs and is associated with a transmission rate of <2%. There have been recent concerns about the possible link between mitochondrial dysfunction and perinatal exposure to antiretroviral nucleoside analogues, particularly zidovudine (ZDV) and lamivudine (3TC). This was reported from the French cohort in 1998, but extensive retrospective analysis of US and other European data have not revealed additional cases. In Europe and the USA, it has been agreed that the benefits of antiretroviral therapy (ART) in reducing MTCT outweigh the possible adverse effects, but that it is important to prospectively follow all infants born to infected women as the long-term effects of exposure to ART *in utero* is unknown.

A European trial and a meta-analysis of cohort data from the USA and Europe showed that in women taking no ART or ZDV monotherapy, elective CS delivery decreased the risk of MTCT by approximately 50% compared with vaginal or emergency CS delivery. This approach has been widely adopted in Europe for women taking mono ART in pregnancy to prevent MTCT. However, it has been less widely adopted in some countries such as the USA, where women are more likely to be given triple HAART in pregnancy in order to reduce viral load to below the level of detection. In this situation the additional benefit of an elective CS delivery remains unclear. There is probably no place for dual ART with ZDV and 3TC in pregnancy, as this is rarely able to fully inhibit viral replication, adds little to reducing transmission with ZDV and elective CS, increases the potential for toxicity in the infant and has been associated with rapid selection of 3TC-associated mutants.

In the developing world, a number of major studies have evaluated the efficacy of cheaper and less complicated perinatal ART regimens. These include short-course ZDV and most notably the use of a single dose of the non-nucleoside reverse transcriptase inhibitor (NNRTI) drug nevirapine to the mother during labour and to the infant within the first three days of birth. This extremely cheap regimen has been shown to reduce transmission by nearly 40% compared with a regimen of intrapartum and neonatal ZDV for a week, even in breastfeeding women over a period of 12 months. It is now being implemented alongside antenatal HIV testing programmes in many parts of the developing world. A concern that resistance to nevirapine, which occurred in about 15% of women, might compromise its use in subsequent pregnancies is probably unfounded as virus returns to wild type in the months following delivery. However, there are concerns that giving a single dose of nevirapine during labour in addition to other ART to women in Europe and the USA who fail to achieve undetectable viral load, could compromise the woman's future ART options to any NNRTI drug because of the rapid selection of HIV strains resistant to nevirapine even after a single dose. Resistance testing to guide therapy choice is routinely recommended for all HIV-infected pregnant women in many developed countries because of *de novo* acquisition of drug-resistant strains of HIV-1.

Diagnosis

IgG antibodies to HIV are passively transferred to virtually all babies born to infected mothers, unless they are born extremely preterm or the mother has profound hypogammaglobulinaemia. Standard IgG antibody assays are so sensitive that traces of

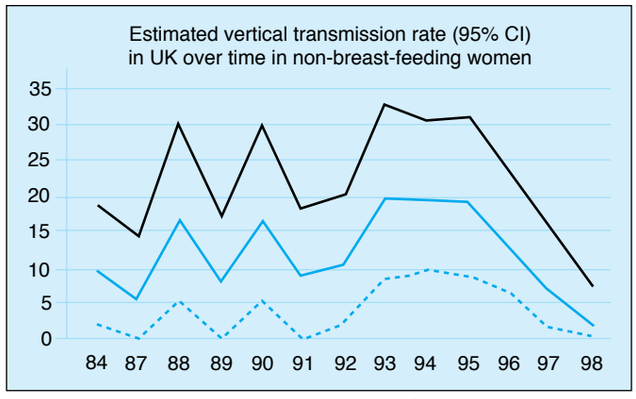


Figure 12.5 Estimated vertical transmission rate (95% CI) in UK over time in non-breastfeeding women (from Doung, BMJ 1999)

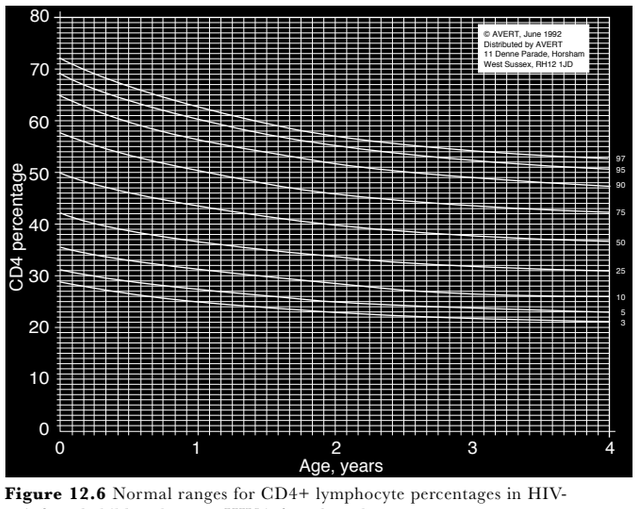


Figure 12.6 Normal ranges for CD4+ lymphocyte percentages in HIV-uninfected children born to HIV-infected mothers.

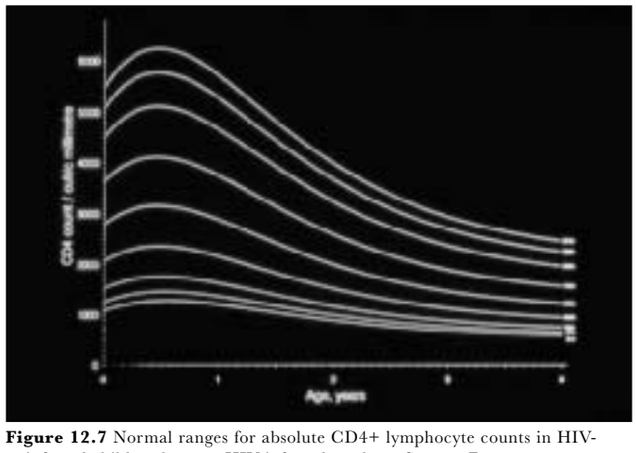


Figure 12.7 Normal ranges for absolute CD4+ lymphocyte counts in HIV-uninfected children born to HIV-infected mothers. Source: European Collaborative Study

Table 12.2 Suggested follow-up of infants born to HIV-infected mothers

Age	Action	Comment
Birth to 4–6 weeks	Give antiretroviral prophylaxis to baby	Usually zidovudine monotherapy, but modified in light of maternal therapy
24–48 hours	Proviral DNA PCR*	If positive, suggests intrauterine transmission or high intrapartum innoculum: may be associated with more rapid disease progression. Not helpful if negative, as less than 50% of infected babies can be detected within 48 hours of birth
3–6 weeks	Proviral DNA PCR	Should detect 95% of infected infants. A positive result must be confirmed on two separate blood samples
4–6 weeks	Stop antiretroviral prophylaxis. Start PCP prophylaxis**	See text for PCP prophylaxis recommendations
3–4 months	Proviral DNA PCR	If all assays are negative and there are no clinical concerns, child is almost certainly uninfected. PCP prophylaxis can be stopped
18 months	HIV antibody test	Performed until seroreversion documented

*Initial infant sample should be tested in parallel with maternal sample obtained around the time of delivery, to ensure maternal strain of HIV can be detected.
**For very low risk infants paediatric specialists increasingly are not recommending PCP prophylaxis.

maternal antibody are frequently detectable in the baby up to 18 months of age. Waiting for antibodies to become undetectable ("seroreversion") is therefore a slow way to establish the child's infection status. Using techniques that detect proviral HIV DNA, by polymerase chain reaction (PCR) or other amplification techniques, 93% of infected infants can be diagnosed by one month of age, and virtually all by three months. Quantitative RNA assays are now widely available but are not licensed for diagnostic purposes because of problems with false positive results and variable performance with non-B clade viral isolates. Whichever test is used, it is essential to ensure that it efficiently detects the maternal strain of HIV.

Virus culture is a highly specialised assay that is available only in research laboratories and has been largely superseded by amplification assays. Immune complex dissociated p24 antigen assays (ICD p24 ag) detect the nuclear capsid antigen of the virus by a commercial ELISA kit. This is a cheap but less sensitive method of diagnosis. Similarly IgA assays are highly specific but lack sensitivity, particularly during the first three months of life.

If PCR assays which reliably detect the mother's strain of HIV are negative on the infant's blood at three different time points, with at least one set performed at or after three months of age, and there are no clinical concerns, the parents/guardians can be informed that their baby is almost certainly not infected (Table 12.2).

T-cell subsets and measurement of immunoglobulins (Ig) are non-specific tests. Reversal of the CD4 : 8 ratio and high Ig (>2 × upper limit of normal) are suggestive of infection but, for diagnostic purposes, should be supported by at least one other test that detects the virus directly. It is important to realise that absolute CD4 counts are physiologically much higher in infants and young children than in adults (Figures 12.7 and 12.8).

All children presumed to be uninfected should be followed until seroreversion is confirmed, and longer term follow-up to ensure normal development until four to five years is advised in children exposed to ART perinatally.

Natural history and clinical manifestations

As in adults, HIV-infected children present with a spectrum of signs and symptoms reflected in the revised Centre for Disease Control classification system (Box 12.3). The differences between adults and children with HIV disease are summarised (Box 12.4). In the absence of HAART, disease progression is generally faster than in adults, with 15–20% of children

Box 12.3 Centers for Disease Control 1994 revised classification system for HIV infection in children less than 13 years old

Category N: no symptoms

Category A: mildly symptomatic
- Lymphadenopathy
- Hepatomegaly
- Splenomegaly
- Dermatitis
- Parotitis
- Recurrent upper respiratory tract infections, sinusitis or otitis media

Category B: moderately symptomatic
Examples of conditions in clinical category B include:
- Anaemia, neutropenia or thrombocytopenia
- Bacterial infections: pneumonia, bacteraemia (single episode)
- Candidiasis, oropharyngeal
- Cardiomyopathy
- Diarrhoea, recurrent or chronic
- Hepatatis
- Herpes stomatitis, recurrent
- Lymphoid interstitial pneumonia
- Nephropathy
- Persistent fever > 1 month
- Varicella (persistent or complicated primary chickenpox or shingles)

Category C: severely symptomatic
Any condition listed in the 1987 surveilllance case definition for AIDS, *with the exception of LIP*. For example:
- Serious bacterial infections, multiple or recurrent
- Candidiasis (oesophageal, pulmonary)
- Cytomegalovirus disease with onset of symptoms at age >1 month
- Cryptosporidiosis or Isosporiasis with diarrhoea persisting 1 month
- Encephalopathy
- Lymphoma
- *Mycobacterium tuberculosis* disseminated or extrapulmonary
- *Mycobacterium avium* complex or *M. kansasii*, disseminated
- *Pneumocystis carinii* pneumonia
- Progressive multifocal leucoencephalopathy
- Toxoplasmosis of the brain with onset at age > 1 month
- Wasting syndrome

developing AIDS-defining illnesses by 12 months. This subset of perinatally infected children typically present with PCP at around three to four months of age (Figure 12.8). Progression rates to AIDS in infancy have been shown to be reduced by the use of primary PCP prophylaxis with Septrin from 4 to 6 weeks of age onwards.

Approximately 70% of perinatally infected children will have some signs or symptoms by 12 months (Figure 12.14). In the absence of antiretroviral therapy, the median age at which children progress to AIDS is about six years, and 25–30% have died by this age. The median age of death is around nine years. In many cases, the child is the first family member to be diagnosed as HIV infected. Some children, however, do not present until the second decade of life. Disease progression in children in developing countries is more rapid (Figure 12.15). Survival following an AIDS diagnosis has greatly improved over the past 10 years, but even where antiretroviral therapy is available the mortality amongst children with PCP and CMV is appreciable (Figure 12.13). This is yet another reason for antenatal HIV testing which can render PCP in infancy wholly preventable.

Children with HIV infection frequently present with signs and symptoms that are common in general paediatrics and are non-specific. The most usual clinical features associated with HIV infection include persistent generalised lymphadenopathy, hepatosplenomegaly, chronic or recurrent diarrhoea, fever, and recurrent otitis or sinusitis.

Persistent oral candidiasis, bilateral parotitis or neurological signs are more specific of HIV infection. Herpes zoster (shingles) in childhood is uncommon and suggests a defect in cellular immunity justifying an HIV test in the absence of other explanations. Similarly, thrombocytopenia can be a presenting feature, and HIV should be considered in the differential diagnosis of idiopathic thrombocytopenic purpura.

Recurrent and often severe bacterial infections are frequent and include pneumonia, cellulitis, local abscesses, osteomyelitis, septic arthritis and occult bacteraemia. The common causative organisms are similar to those seen in children with hypogammaglobulinaemia and include pneumococci, salmonellae, staphylococci, streptococci and *Haemophilus influenzae*. This reflects the B-cell defect that accompanies the destruction of the CD4+ helper T cells. Children with HIV infection frequently have hypergammaglobulinaemia due to dysregulated polyclonal B-cell activation. The antibodies are generally non-functional.

Pulmonary disease is an important cause of morbidity and mortality and may be one of the first manifestations. Lymphoid interstitial pneumonitis (LIP), characterised by multiple foci of proliferating lymphocytes in the lung interstitium, occurs in 20–30% of vertically infected children, but is rare in adults. It presents with persistent bilateral reticulonodular shadowing on chest X-ray (Figure 12.9) and clinical features ranging from asymptomatic to chronic hypoxia. It may be an abnormal response to primary Epstein–Barr virus (EBV) infection. Co-infection with *Mycobacterium tuberculosis* is an increasing problem in children, and can be difficult to distinguish radiologically from LIP. Clinically a child with bilateral infiltrates due to TB would be highly symptomatic, as opposed to LIP which may be clinically silent.

Opportunistic infections, apart from PCP and primary disseminated CMV disease in the subset of children with very rapid disease progression, are usually a late complication of HIV infection and result from severe immunosuppression. The most common are oesophageal candidiasis, multidermatomal varicella zoster, disseminated mycobacterium avium complex (MAC) or CMV infections, cryptosporidiosis, and more rarely,

Box 12.4 Differences between children and adults with HIV disease

- More rapid disease progression:
  - 20% of children develop AIDS by 12 months
  - Child may be the first family member to present
- Higher viral loads at presentation
- Physiologically higher absolute CD4 counts
- Growth faltering common (affects height and weight)
- Encephalopathy presents with developmental delay and hypertonic diplegia
- Opportunistic pathogens encountered for the first time
  - primary illnesses often more severe than OIs in adults
- Poor primary responses to childhood infections/immunisations
- Lymphoid interstitial pneumonitis common
- Malignancy uncommon (accounts for less than 2% of AIDS-defining presentations in children)
- More rapid clearance of antiretroviral drugs, requiring higher than adult equivalent doses particularly in very young children

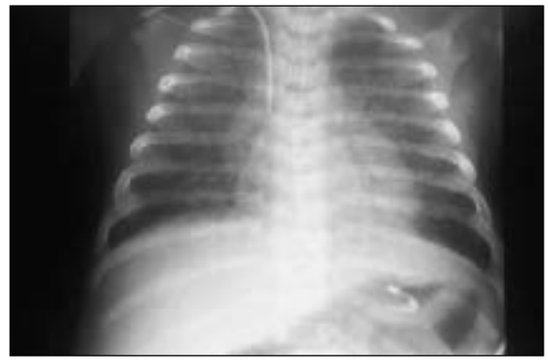


Figure 12.8 *Pneumocystis carinii* pneumonia (PCP) in a three month old. Diffuse bilateral ground-glass opacification, tending to confluence in right upper and both lower lobes. Air bronchograms are seen, which imply air space disease which is a late feature of disease. The earliest infiltrates are usually perihilar. The absence of pleural effusion or hilar adenopathy is typical. Less typical presentations include miliary, coin and nodular lesions, lobar consolidation and cavitations

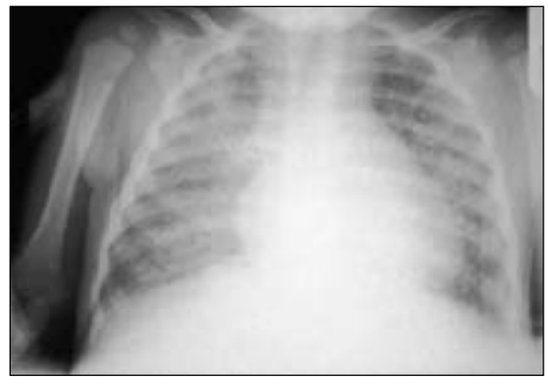


Figure 12.9 Lymphoid interstitial pneumonitis (LIP) in a child aged 12 months. Diffuse, well-circumscribed nodules distributed uniformly throughout both lung fields. May be associated with hilar adenopathy. A radiological spectrum is seen in LIP, ranging from fine linear interstitial infiltrates to large nodules that tend to confluence in the right middle and lingular lobes

toxoplasmosis. MAC should be considered in any child with advanced disease and unexplained fevers, weight loss and abdominal discomfort.

Encephalopathy due to effects of HIV infection on the central nervous system is seen most frequently in the subgroup of children with rapid disease progression. The most common neurological manifestations are hypertonic diplegia, developmental delay (particularly affecting motor skills and expressive language) or acquired microcephaly. Cranial imaging studies may show basal ganglia calcification and cerebral atrophy and MRI scans may show evidence of white matter damage. Seizures are not usually a feature of HIV encephalopathy which does not tend to affect the grey matter. The majority of school age children are attending normal school without requiring additional support in the classroom.

Malignancy, such as Kaposi's sarcoma or lymphoma, is a relatively uncommon feature of paediatric HIV disease, accounting for only 1–2% of AIDS-defining illness in children.

Prognostic markers

The most widely used surrogate markers for predicting disease progression in children, as in adults, are the CD4 values and viral load. Very high viral loads are frequently found in infected children, particularly following perinatal transmission (Figure 12.10). Absolute CD4 counts are physiologically higher in children compared with adults (Figure 12.7). CD4 percentages vary rather less, and according to the CDC classification system can be used across all age ranges; >25% is considered evidence of no immunosuppression, 15–25% moderate and <15% severe immunosuppression (Figure 12.6).

Tables 12.3–12.5 contain data from a pre-antiretroviral trial of intravenous immunoglobulin in the USA that illustrate the independent prognostic value of these markers, although positive predictive values of each are low. Age-adjusted rates of change for viral load and CD4 counts may be of higher positive predictive value for disease progression: an analysis of combined European and US data is presently underway to evaluate this concept.

Management

The aim of any intervention for HIV-infected children should be to maintain the best possible quality of life for the children as long as possible, with the hope that they will be able to take advantage of potential curative therapy in the future. This inevitably means balancing the potential benefits of new treatments against the need for increased monitoring, possible toxicities and limiting future therapeutic options.

As a result of advances in ART, there has been a shift in focus from diagnosing and managing opportunistic infections (OI) to preventing them by restoring and maintaining cellular immunity. For most established opportunistic infections, the best treatment is HAART.

Antiretroviral therapy

Virus replication in children, as in adults, is occurring at all stages of HIV infection and, as improved drugs and drug combinations become available, treatment is likely to be offered increasingly early. Highly encouraging results have been reported with three or more drug combinations in selected, infected infants, which demonstrate that complete viral suppression and maintenance of entirely normal immune development can be achieved and sustained for at least three years. These observations, and studies of adults treated during primary infection, provide a rationale for early aggressive therapy of infants.

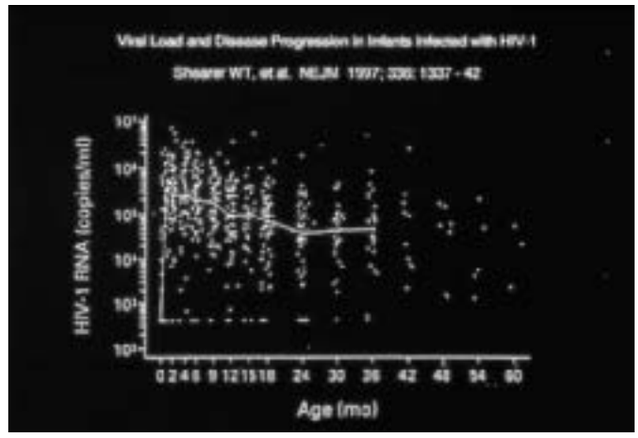


Figure 12.10 Plasma viral load (HIV-l RNA copies/ml) over time in a cohort of perinatally infected, non-breast-fed infants. Solid line represents median values at each time point. Note that median viral load on days 1–3 of life was below the limit of detection (<400.copies/ml). Most non-breast-fed infants are infected during labour or delivery, resulting in a lag before viral replication reaches detectable levels in the circulation

Table 12.3 Association of baseline CD4+ lymphocyte percentage with long-term risk of mortality in HIV-infected children

Baseline CD4+ percentage	Patients (no.)	Deaths (no.)	% mortality
25%	189	50	26
15–24%	93	31	33
5–14%	59	35	59
<5%	33	32	97

Table 12.4 Association of baseline HIV RNA copy number with long-term risk of mortality in HIV-infected children

Baseline HIV RNA (copies/ml)*	Patients (no.)	Deaths (no.)	% mortality
Undetectable	25	6	24
4001–50 000	69	19	28
50 001–500 000	105	34	32
500 001–1 000 000	20	8	40
>1 000 000	35	25	71

*Tested by NASBA RNA QT Amplification system on frozen stored serum (lower limit of detection=4000 copies/ml)

Table 12.5 Association of baseline HIV RNA copy number and CD4+ cell percentage with long-term risk of mortality in HIV-infected children

Baseline viral load/ CD4+ percentage	Patients (no.)	Deaths (no.)	% mortality
100 000 copies/ml			
15%	103	15	15
<15%	24	15	63
>100 000 copies/ml			
15%	89	32	36
<15%	36	29	81

Mean age = 3.4 years, mean follow-up = 5.1 years

* Data taken from the National Institute of Child Health and Human Development Intravenous Immunoglobulin Clinical Trial. Reproduced from Mofenson L *et al. Journal of Infectious Disease* 1997;**175**:1029–38.

Less impressive results have been documented outside clinical trials. Some children have failed due to inadequately defined pharmacokinetics for drugs like nelfinavir in infants, many of the infants have now been exposed to all three classes of ART and have few therapeutic options left, and adverse consequences of early prolonged therapy are unknown. Considerable support is required to enable families to sustain high levels of adherence long term. US guidelines recommend HAART for all infants, but European practice tends to be more conservative (Figure 12.11).

Older children presenting for the first time are a selected group who are not rapid disease progressors. For these children it is reasonable to monitor CD4 counts and only offer treatment if counts are declining steadily below 25%. There is no consensus level of viral load above which treatment must be started.

When starting HAART, most prescribers would initiate triple or even quadruple combination therapy, ideally sparing at least one class of drugs. The protease inhibitors (PI) are more difficult to formulate into palatable suspensions for children compared with the nucleoside analogues and non-nucleoside reverse transcriptase inhibitors. No data in children provide evidence to conclude that PI-containing or PI-sparing regimens have greater long-term clinical efficacy.

The management of heavily pretreated children who are failing therapy requires careful evaluation of past drug history, adherence, unused treatment options, possibly genotypic or phenotypic resistance testing, and pharmacodynamics. It has become clear that many drugs are more rapidly cleared in children. The problem of underdosing in infancy has already been mentioned. Adolescents may require higher than adult doses until reaching Tanner IV or V stages of puberty. Increasingly therapeutic drug monitoring will be used to understand population pharmacokinetics, and to tailor individual therapy.

Short- and long-term toxicities of specific drugs and drug classes are broadly similar in children as in adults, with 58% of children presenting with at least one side-effect in a recent national Italian survey of children on HAART. The most common toxicities are gastrointestinal symptoms and skin rashes. Lipodystrophy is increasingly described, particularly in adolescents who may wish to switch or discontinue therapy as a result. Lipid metabolism abnormalities with significantly raised fasting cholesterol and/or triglyceride levels are particularly associated with PI-containing regimens. Long-term consequences are not yet known and no consensus has emerged regarding the use of statins, but early onset cardiovascular complications are a potential risk.

It is likely that long-term control of viral replication in children will require adjunctive immune-based treatment, and several approaches are under investigation. The role of strategic treatment interruptions is also being evaluated. The long-term goal is to restore the child's HIV-specific immune responses to the point where HAART is no longer needed.

In view of the many uncertainties regarding optimal treatment, it is strongly recommended that children should be offered treatment as part of a clinical trial. Paediatricians in Europe and Brazil are collaborating in a series of studies coordinated by the Pediatric European Network for the Treatment of AIDS (PENTA). Information about the PENTA studies is available through the Medical Research Council Clinical Trials Centre in London (telephone +44 (0) 20 7670 4791/2, fax +44 (0) 20 7670 4814) or INSERM in Paris (telephone +33 1 4559 5201, fax +33 14559 5180).

Box 12.5 Issues to consider when starting therapy in Children

- Parental (and child) readiness
- Likelihood of good longterm adherence
- What formulations could this child take (taste testing: let child chose)?
- What pharmacokinetic data are available for infants/children/adolescents?
- What experience have other family members had on antiretroviral drugs?

Box 12.6 Drug combinations consider:

- Pill/liquid burden
- Ease of adminstration
- With/without food
- Number of times per day (avoid school hours, ?once daily for adolescents)
- Creative use of drug–drug interactions (boosting with ritonavir)
- Pill-swallowing techniques
- Adherence aids (sticker charts, dosette boxes etc.)
- Gastrostomy tubes to improve quality of life if this is severely eroded by difficulties taking medicines orally

Box 12.7 Recommendations for use of HAART in children (adapted from PENTA, 2001)

- Must start HAART if
  Clinical stage C or immunological stage 3 disease (CD4<15%)
- Consider HAART if
  Clinical stage B or
  Steadily declining CD4% falling below 25%, or
  High viral load (>10^6 RNA copies/ml if age <1 year, >10^5 if age over 1 year)
  Infant <12 months, regardless of CD4 or viral load
- Defer HAART if
  Stage N or A disease
  CD4>25%
  Low viral load:
  <10^5 in children between 1 and 30 months
  <50 000 copies/ml in children >30 months

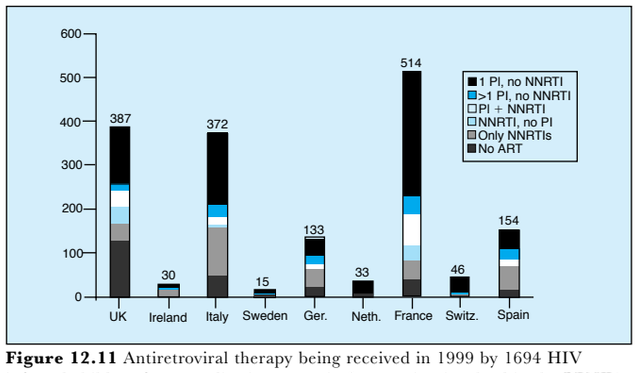


Figure 12.11 Antiretroviral therapy being received in 1999 by 1694 HIV infected children from paediatric centres in 9 countries, involved in the PENTA network of trials (unpublished data, Paediatric European Network for the Treatment of AIDS)

Prophylactic measures

Early-onset PCP is a preventable disease (Table 12.2). Infants at higher risk of acquiring HIV, whose mothers are identified during pregnancy, can be started on PCP prophylaxis from around 4 to 6 weeks of age onwards. Prophylaxis can be stopped once it has been established that the baby is uninfected. Infected children should continue on prophylaxis throughout the first year of life, as CD4 counts are unreliable indicators of risk (see Figure 12.7). Thereafter, it is not unreasonable to stop prophylaxis for children with CD4 counts consistently above 15%, provided the family are reliable clinic attendees and the child's clinical status and immune function can be regularly monitored. Any child with rapidly declining CD4 counts or counts consistently less than 15% should be on prophylaxis. Co-trimoxazole is the drug of choice. Regimens vary, but one convenient dosage regimen is suggested in Table 12.5. Rashes and bone marrow suppression due to co-trimoxazole may require switching to alternative prophylactic agents such as dapsone.

Routine active immunisation schedules should be followed for HIV-infected or -exposed infants, with the exception that BCG should not be given to symptomatic infected children because of the risk of dissemination. There is a theoretical risk of paralytic poliomyelitis in immunocompromised contacts of children excreting live polio vaccine virus. Inactivated polio vaccine (IPV) may be recommended by injection instead of the live oral polio vaccine. In practice it can be difficult to obtain supplies of IPV in the UK, and in view of the very low transmission rate of HIV, many units now condone giving oral polio vaccine (OPV), and advise carers about thorough hand-washing when changing nappies.

Pneumococcal polysaccharide vaccine has been recommended for HIV-infected children over two years of age, but is likely to be superseded soon by conjugate vaccines which can be given to younger children. Influenza vaccine is generally offered each winter, although data demonstrating its efficacy in this population are lacking.

Passive immunisation of symptomatic children is recommended if they are in contact with varicella zoster virus (VZV) and are either VZV naive or have no detectable specific antibodies to VZV. Varicella zoster immunoglobulin (VZIG) ideally should be given within 72 hours of contact. VZIG may prolong the incubation period to 28 days, so clinicians need to consider isolating these patients at clinic visits. Similarly normal human immunoglobulin should be given for susceptible symptomatic children in contact with measles. If children are stable on HAART with CD4 counts above 15%, passive immunisation is unnecessary.

Regular intravenous immunoglobulin infusions (400 mg/kg every 28 days) should be reserved for children with recurrent bacterial infections despite good compliance with co-trimoxazole prophylaxis, or those with proven hypogammaglobulinaemia. Higher doses may be useful in the management of thrombocytopenia (0.5–1.0 g/dose every day, for three to five days).

HIV-infected children who are household or day care contacts of individuals with open pulmonary tuberculosis should be carefully assessed, bearing in mind skin testing is frequently unhelpful because of anergy. If there is no evidence of infection, prophylactic isoniazid for six months, or isoniazid plus rifampicin for three months, is recommended. There is little enthusiasm for prophylaxis against *Mycobacterium avium intracellulare* in children because of adverse reactions and the potential for resistance and breakthrough on single agents such as rifabutin. The most appropriate prophylaxis for all OI is to optimise antiretroviral therapy to restore and preserve immune function.

Table 12.6 Suggested doses of co-trimoxazole for prophylaxis for *Pneumocystis carinii* pneumonia, to be given once daily on three days per week (usually Monday, Wednesday and Friday). Dose is based on 900 mg/m² dose

Surface area (m²)	Dose of co-trimoxazole (mg)
0.25–0.39	240
0.40–0.49	360
0.50–0.75	480
0.76–1.0	720
> 1.0	960 (adult dose)

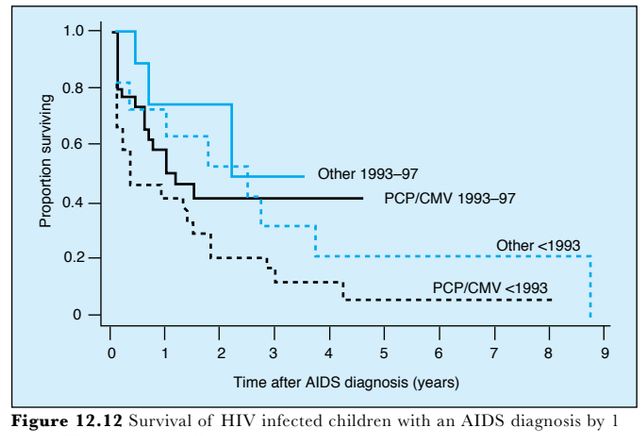


Figure 12.12 Survival of HIV infected children with an AIDS diagnosis by 1 year (UK and Ireland). Survival has improved significantly for those born after 1993 CMV, Cytomegalovirus infection; PCP, pneumocystis carinii pneumonia. Reproduced from Williams A *et al.* PCP and CMV infection in children with vertically acquired HIV infection. *AIDS* 2001;**15**: 1–5

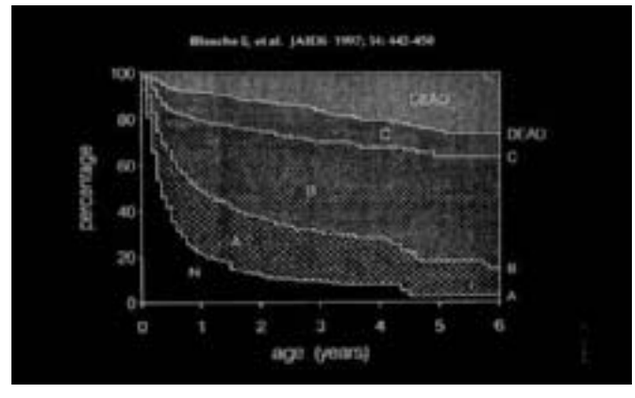


Figure 12.13 Morbidity and Mortality in European children vertically infected by HIV-1

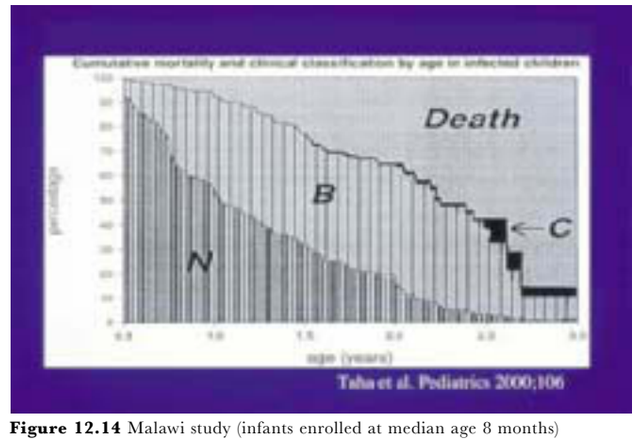


Figure 12.14 Malawi study (infants enrolled at median age 8 months)

Supportive care

Unlike almost any other life-threatening disease of children, HIV simultaneously threatens the parents and other siblings. The parents' own health, their social isolation and feelings of guilt compound the difficulties of caring for a sick child. An effective well-coordinated multidisciplinary team is required to address the changing needs of infected and affected children and their caregivers. Continuity of care between inpatient and outpatient services, local referring hospitals and the community needs to be developed. Ideally adults and children should be treated in family-based units. All too often parents will ignore their own health needs because they put their children first.

Increasingly the work of the multidisciplinary team has shifted towards ways of helping families achieve long-term adherence to HAART. As children survive longer, meeting the needs of adolescents and planning transition to adult clinics is placing new demands on services.

The decision as to who should be informed should be tailored individually. Families may need help in explaining the diagnosis to older children. This needs to be undertaken at the child's pace, and is frequently most effectively achieved in gradual steps. It is not mandatory to tell staff at schools, as universal precautions should be employed for all children with cuts and abrasions. The risks of transmission from casual contacts in school or day care settings are virtually nil. Ensuring that adolescents are well informed and responsible before they become sexually active themselves is a priority.

The child's developmental needs require careful monitoring and support, with access to a clinical psychologist, a physiotherapist, occupational therapist and speech therapist.

The multidisciplinary team should include a dietician, as nutritional problems and growth faltering are very common complications. Balanced supplements are sometimes required and enteral feeding through gastrostomy tubes and occasionally intravenous parenteral feeding may be necessary. Gastrostomy tubes have been used with success to allow unpalatable medicines to be given, even when they were not required for nutritional supplementation.

Because children below the age of eight years very rarely complain of symptoms of unilateral eye disease, regular monitoring of young children with CD4 counts less than 5% by a paediatric ophthalmologist is desirable. Chorioretinitis due to CMV is usually treated by intravenous induction therapy with ganciclovir followed by regular maintenance intravenous treatment five days per week. Paediatric formulations of oral ganciclovir are poorly bioavailable. Intravitreous injections and, in older children, implants have been used.

Pain management is of critical importance in late-stage disease. Complementary therapies such as therapeutic touch and aromatherapy may be useful and require evaluation. It is a testament to the success of HAART that very few children in industrialised countries are needing palliative or terminal care. However unless new treatment strategies become available, the next few years may see some children running out of therapeutic options.

Prevention remains the top priority in managing HIV infection in children. Reducing national perinatal transmission rates to below 2% is an achievable target that can only be realised if HIV-infected mothers can be identified prenatally and offered appropriate interventions. This will require continued effort by health professionals, public health planners and community organisations.

13 HIV counselling and the psychosocial management of patients with HIV or AIDS

Sarah Chippindale, Lesley French

What is HIV counselling?

Counselling in HIV and AIDS has become a core element in a holistic model of healthcare, in which psychological issues are recognised as integral to patient management. HIV and AIDS counselling has two general aims: (1) the prevention of HIV transmission and (2) the support of those affected directly and indirectly by HIV. It is vital that HIV counselling should have these dual aims because the spread of HIV can be prevented by changes in behaviour. One-to-one prevention counselling has a particular contribution in that it enables frank discussion of sensitive aspects of a patient's life – such discussion may be hampered in other settings by the patient's concern for confidentiality or anxiety about a judgemental response. Also, when patients know that they have HIV infection or disease, they may suffer great psychosocial and psychological stresses through a fear of rejection, social stigma, disease progression and the uncertainties associated with future management of HIV. Good clinical management requires that such issues be managed with consistency and professionalism, and counselling can both minimise morbidity and reduce its occurrence. All counsellors in this field should have formal counselling training and receive regular clinical supervision as part of adherence to good standards of clinical practice.

When is HIV counselling necessary?

Pre-test discussion

A discussion of the implications of HIV antibody testing should accompany any offer of the test itself. This is to ensure the principle of informed consent is understood and to assist patients to develop a realistic assessment of the risk of testing HIV antibody positive. This process should include accurate and up-to-date information about transmission and prevention of HIV and other sexually transmitted infections. Patients should be made aware of the "window period" for the HIV test – that a period of 12 weeks since the last possible exposure to HIV should have elapsed by the time of the test.

Patients may present for testing for any number of reasons, ranging from a generalised anxiety about health to the presence of HIV-related physical symptoms. For patients at minimal risk of HIV infection, pre-test discussion provides a valuable opportunity for health education and for safer sex messages to be made relevant to the individual. For patients who are at risk of HIV infection, pre-test discussion is an essential part of post-test management. These patients may be particularly appropriate to refer for specialist counselling expertise. In genitourinary medicine clinics where HIV antibody testing is routinely offered as a part of sexual health screening, health advisers provide counselling to patients who have been identified as high risk for testing HIV positive.

The importance of undertaking a sensitive and accurate sexual and/or injecting drug risk history of both the patient and their sexual partners cannot be overstated. If patients feel they cannot share this information with the physician or counsellor then the risk assessment becomes meaningless; patients may be inappropriately reassured, for example, and be unable to disclose the real reason for testing. Counselling skills are clearly an essential part of establishing an early picture of the patient

Box 13.1 Counselling

Prevention
- Determining whether the lifestyle of an individual places him or her at risk
- Working with an individual so that he or she understands the risks
- Helping to identify the meanings of high-risk behaviour
- Helping to define the true potential for behaviour change
- Working with the individual to achieve and sustain behaviour change

Support
- Individual, relationship and family counselling to prevent and reduce psychological morbidity associated with HIV infection and disease

Box 13.2 Different HIV counselling programmes and services

- Counselling before the test is done
- Counselling after the test for those who are HIV positive and HIV negative
- Risk-reduction assessment to help and prevent transmission
- Counselling after a diagnosis of HIV disease has been made
- Family and relationship counselling
- Bereavement counselling
- Telephone "hotline" counselling
- Outreach counselling
- Crisis intervention
- Structured psychological support for those affected by HIV
- Support groups

Box 13.3 Pretest discussion checklist

Indications for further counselling and referral to counsellor
- People who have been sexually active in areas of high HIV prevalence
- Men who have sex with men
- Current or previous sexual partners HIV positive
- Client presenting with clinical symptoms of HIV infection
- High-risk sexual behaviour
- High-risk injecting drug practices
- Learning or language difficulties

Points for counsellor and/or physician to cover
- What is the HIV antibody test (including seroconversion)?
- The difference between HIV and AIDS
- The window period for HIV testing
- Medical advantages of knowing HIV status and treatment options
- Transmission of HIV
- Safer sex and risk reduction
- Safer injecting drug use
- If the client were positive how would the client cope: personal resources, support network of friends/partner/family?
- Who to tell about the test and the result
- Partner notification issues
- HIV status of regular partner: is partner aware of patient testing?
- Confidentiality
- Does client need more time to consider?
- Is further counselling indicated?
- How the results of the test are obtained (in person from the physician or counsellor)

and his/her history and of how much intervention is needed to prepare him or her for a positive result, and to further reinforce prevention messages. It is at this stage that potential partners at risk are identified which will become an important part of the patient's management if HIV positive.

Post-test counselling

Results

HIV results should be given simply, and in person. For HIV negative patients this may be a time where the information about risk reduction can be "heard" and further reinforced. With some patients it may be appropriate to consider referral for further work on personal strategies to reduce risks, for example one-to-one or group interventions. The window period of 12 weeks should be checked again and the decision taken about whether further tests for other sexually transmitted infections are appropriate.

HIV positive patients should be allowed time to adjust to their diagnosis. Coping procedures rehearsed at the pre-test discussion stage will need to be reviewed in the context of the here and now; what plans does the patient have for today, who can they be with this evening? Direct questions should be answered but the focus is on plans for the immediate few days, when further review by the counsellor should then take place. Practical arrangements including medical follow-up should be written down. Overloading the patient with information about HIV should be avoided at the result giving stage – sometimes this may happen because of the health professional's own anxiety rather than the patient's needs.

Newly diagnosed patients

Counselling support should be available to the patient in the weeks and months following the positive test results. Immediate issues often include disclosure to others which may present a complex challenge to the patient. Current and previous sexual partners at risk will have been identified at the pre-test discussion stage and possible ways of informing these people will be explored with the counsellor. It is also important to discuss safer sex with those diagnosed with HIV. Pregnant women who test HIV positive need information and advice on the management of their pregnancy, including: options on reducing the risk of materno-fetal transmission, options for their own treatment and referral to specialist medical and counselling support. Although testing HIV positive is not a reason *per se* for seeking a termination of pregnancy, all women should be given the opportunity to discuss this if appropriate. Families may be a source of support but in many instances patients need time to come to terms with their diagnosis and to fully understand its implications before they have the capacity or resources to raise it with parents, siblings and/or loved ones who will inevitably be distressed. Being identified HIV positive may facilitate constructive planning for the future, such as deciding on the future welfare and care of children, although this tends to happen later in the counselling process when the early shock has resolved.

Counselling involves understanding a person in their social and familial contexts and many patients will derive crucial support and strengthening of coping mechanisms from this intervention during this vulnerable period. Counselling support can also help a patient engage in wider medical care and monitoring. If a person is inadequately prepared for the test, or a positive result is given inappropriately, he or she may reject further intervention including accessing medical care and therefore the likelihood of psychological morbidity and disease progression may be increased. It seems the "getting it right" for patients at early stages of diagnosis has a profound effect upon

Box 13.4 Counselling skills

- Empathy
- Non-judgemental approach
- Active listening
- Clear discussion and information giving
- Ability to establish working relationship with client
- Facilitating appropriate planning by the client
- Motivating appropriate self-care and reflective abilities

Box 13.5 Post Test Counselling – HIV Positive Result

IMMEDIATE FOLLOW-UP

- Time for "ventilation"
- Awareness of shock factor – keep information to a minimum
- Focus on coping today, tonight, next few days
- Who knows the patient is receiving the result today?
- Safer sex/Partners
- Arrangements for confirmatory HIV test
- Follow-up medical and counselling appointment
- Written information – support numbers

Box 13.6 Psychological issues in HIV/AIDS counselling

Shock

- of diagnosis
- recognition of mortality
- of loss of hope for the future

Fear and anxiety

- uncertain prognosis
- effects of medication and treatment/treatment failure
- of isolation and abandonment and social/sexual rejection
- of infecting others and being infected by them
- of partner's reaction

Depression

- in adjustment to living with a chronic viral condition
- over absence of a cure
- over limits imposed by possible ill health
- possible social, occupational and sexual rejection
- if treatment fails

Anger and frustration

- over becoming infected
- over new and involuntary health/lifestyle restrictions
- over incorporating demanding drug regimens, and possible side-effects, into daily life

Guilt

- interpreting HIV as a punishment; for example, for being gay or using drugs
- over anxiety caused to partner/family

their capacity to cope in the subsequent months and years, and to access help appropriately in later stages of disease.

The importance of encouraging and working towards coping strategies involving active participation (to the extent the patient can manage) in planning of care and in seeking appropriate social support has been demonstrated clinically and empirically. Such an approach includes encouraging problem solving, participation in decisions about their treatment and care, and emphasising self-worth and the potential for personal control over manageable issues in life.

Psychological responses to an HIV positive result

Many reactions to an HIV positive diagnosis are part of the normal and expected range of responses to news of a chronic, potentially life-threatening, medical condition. Many patients adjust extremely well with minimal intervention. Some will exhibit prolonged periods of distress, hostility or other behaviours which are difficult to manage in a clinical setting. It should be noted that serious psychological maladjustment may indicate pre-existing morbidity and will require psychological/psychiatric assessment and treatment. Depressed patients should always be assessed for suicidal ideation.

Effective management requires allowing time for the shock of the news to sink in; there may be a period of emotional "ventilation", including overt distress. The counsellor should provide an assurance of strict confidentiality and rehearse, over time, the solutions to practical problems such as who to tell, what needs to be said, discussion around safer sex practices and adherence to drug therapies. Clear information about medical and counselling follow-up should be given. Counselling may be of help for the patient's partner and other family members.

Counselling can also be offered to the patient and their partner together. This should only take place with the patient's explicit consent, but it may be important for the following reasons listed in Box 13.7.

Partners and family members sometimes have greater difficulty in coming to terms with the knowledge of HIV infection than the patients do themselves. Individual counselling support is often required to manage this, particularly role changes within the relationship, and other adjustment issues that may lead to difficulties. This is part of a holistic approach to the patient's overall health care.

In many cases the need for follow-up counselling may be episodic and this seems appropriate given the long-term nature of HIV infection and the different challenges a patient may be faced with. The number of counselling sessions required during any of these periods largely depends on the individual presentation of the patient and the clinical judgement of the counsellor.

Box 13.7 Advantages of counselling patient with their partner

- Adjustments to sexual behaviour and other lifestyle issues can be discussed and explained clearly to both.
- If the patient's partner is HIV negative (i.e. a serodiscordant couple) particular care and attention must be paid to emotional and sexual consequences in the relationship.
- Misconceptions about HIV transmission can be addressed and information on safer sex given.
- The partner's and the patient's psychological responses to the diagnoses or result, such as anxiety or depression, can be explained and placed in a manageable perspective.
- There may be particular issues for couples who have children or who are hoping to have children or where the woman is pregnant.

Box 13.8 Causes of uncertainty

- The cause of illness:
  - Progression of disease
  - Management of dying
  - Prognosis
  - Reactions of others (loved ones, employers, social networks)
- Effects of treatment
- Long-term impact of antiretroviral therapy
- Impact of disclosure and how this will be managed

The worried well

Patients known as the "worried well" present with multiple physical complaints which they interpret as sure evidence of their HIV infection. Typically, fears of infection reach obsessive proportions and frank obsessive and hypochondriacal states are often seen. This group shows a variety of characteristic features, and they are rarely reassured for more than a brief period after clinical or laboratory confirmation of the absence of HIV infection. A further referral for behavioural psychotherapy or psychiatric intervention may be indicated, rather than frequent repetition of HIV testing.

Box 13.9 Characteristics of the worried well

- Repeated negative HIV tests
- Low-risk sexual history, including covert and guilt-inducing sexual activity
- Poor post-adolescence sexual adjustment
- Social isolation
- Dependence in close relationships (if any)
- Multiple misinterpreted somatic features usually associated with undiagnosed viral or postviral states (not HIV) or anxiety or depression
- Psychiatric history and repeated consultation with general practitioners or physicians
- High levels of anxiety, depression and obsessional disturbance
- Increased potential for suicidal gestures

Linking with community and statutory agencies

Counselling and testing should never be provided without clear, working links with services for back-up and complementary management. Links with these services should be planned as an integral part of any HIV/AIDS counselling initiative from the outset. This is particularly important for patients from ethnic minority communities. Such links must be kept open and flexible to ensure that medical information and advice are consistent across all levels of intervention. Finally, the value of groups in HIV psychosocial and stress management is amply demonstrated. Groups are valuable in reducing an individual's sense of isolation, in providing a safe place to express feelings, to share experiences, and to learn successful coping styles from others, for example support groups for those who are newly diagnosed.

HIV counselling and combination antiretroviral therapy

Significant developments in combination antiretroviral therapy have led to a surge of optimism about long-term medical management of HIV infection and people are now living much longer with HIV. Patient adherence is an important factor in the efficacy of drug regimens. However, taking a complicated drug regimen – often taking large numbers of tablets several times a day – is a constant reminder of HIV infection. The presence of side-effects can often make patients feel more unwell than did the HIV and some may be unable to cope with the side-effects. Counselling may be an important tool in determining a realistic assessment of individual adherence and in supporting the complex adjustment to a daily routine of medication.

Discussions on safer sex are important, as drug-resistant HIV strains are emerging which limit treatment options for those acquiring such strains. Many patients diagnosed with HIV some years ago are now feeling well enough to return to work, to study and are, paradoxically, learning to readjust to living as they had formerly adjusted to the possibility of dying. Patients also have to deal with the uncertainty which remains about long-term efficacy of current medical treatment, and there are some who will fail on combination therapy. Even with the significant medical advances in patient management, counselling remains an integral part of the management of patients with HIV, their partners and family.

Box 13.10 Coping strategies

- Using counselling
- Problem solving
- Participation in discussions about treatment
- Using social and family networks
- Use of alternative therapies, for example relaxation techniques, massage
- Exploring individual potential for control over manageable issues
- Disclosure of HIV status and using support options

Box 13.11 Who is HIV and AIDS Counselling for?

- People worried that they might have HIV
- People considering being tested for HIV
- People who have been tested for HIV (both negative and positive)
- People unaware of the risks involved in behaviours that they are, or have been, engaged in
- People with HIV infection and disease, including AIDS
- People needing support with antiretroviral therapies
- People experiencing practical and emotional difficulties as a result of HIV infection
- Family/Partners/Friends of people with HIV/AIDS

14 Palliative care and pain control in HIV and AIDS

Rob George, Chris Farnham, Louise Schofield

This chapter looks at changes in the palliative care of HIV disease before offering practical guidelines in pain and symptom control and managing the days prior to death.

Managing uncertainty vs. managing death

HIV disease constantly challenged the acute/palliative interface with clinical constellations in which "curable" and palliative elements coexist. HAART (highly activated antiretroviral therapy) is the most recent example since which progression beyond "treatability" is less easy to discern. Similarly in cancer, palliative care's role now includes support early in the disease journey to help manage the uncertainties associated with toxic treatments as well as the worries about disability and mortality. Such psycho-emotional, spiritual and social fallout is beyond the capabilities of linear, curative medicine to manage. Palliative (symptom-based) and therapeutic (pathology-based) approaches are not mutually exclusive. The unifying concept of palliative care these days is the management of uncertainty and suffering, only part of which is care of the dying.

A challenging disease

Many patients with HIV/AIDS come from marginalised groups. Characteristically, these minorities take up services late or sit uneasily with conventional care because of different health beliefs or mistrust. Consequently, we cannot approach them all in the same way; neither can we assume what makes a good death nor what comprises a family or close social group. These are the practical realities of patient-centred care.

Flexibility, collaboration and support

Effective care for patients in whom deterioration or death is a real but fluctuating possibility also confronts us with the need to integrate care flexibly. Patients facing chronic ill health or those with acute, highly symptomatic disease may benefit from specialist advice or shared care as much as those may in the terminal phase of their illness. It is only by close collaboration between teams that good outcomes will be achieved. This brings us to our shared problem: that of uncertainty and the ways in which curative and palliative strategies coexist.

The therapeutic dilemma

When palliation is simply supporting aggressive curative therapy, there is no problem. However, when disease is progressive or debilitating, or where prophylactic and maintenance regimens maintain residual health, but compromise the quality of the patient's life, therapeutic decisions must take account of the burdens and benefits in personal as well as pathological terms. This is the idea that treatments may be futile not just by being ineffective, but also by being so destructive to quality of life that they may be worse than useless. Ethicists call this qualitative futility. In order for us to be of genuine benefit clinically, we must find effective ways of living in this tension between cost and benefit.

The quantity/quality equation

Balancing the costs and benefits of treatment formally and explicitly clarifies a patient's best interest, particularly as health

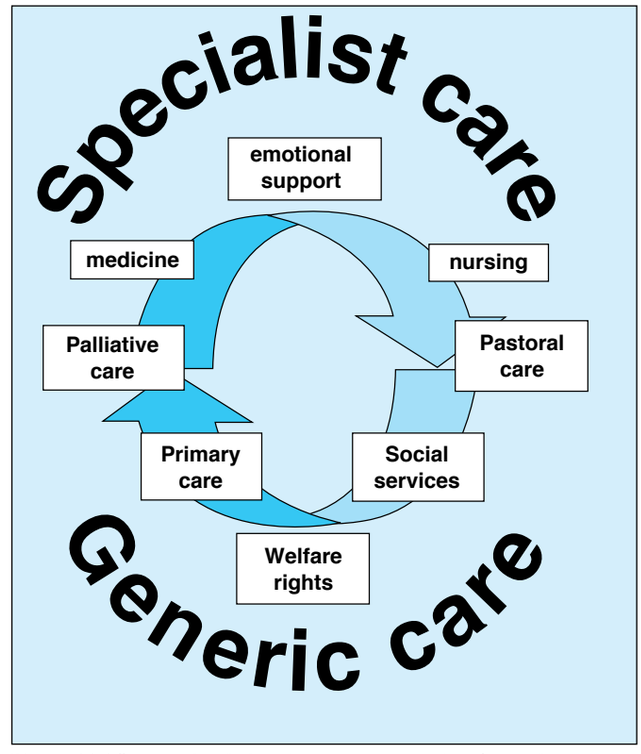


Figure 14.1 Inter- and multidisciplinary care: essential professional groupings necessary for effective supportive and palliative care

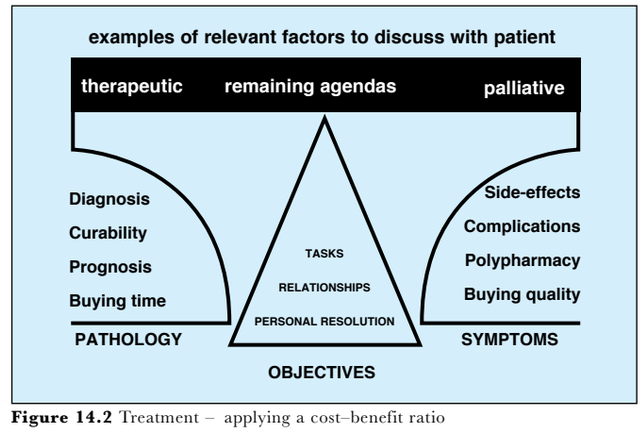


Figure 14.2 Treatment – applying a cost–benefit ratio

begins to fail. Certainties, such as time and energy expended on a course of treatment or immediate and short-term benefits etc. then become increasingly important.

Try not to make assumptions about a patient's views or wishes. The subtle ways in which illness and the individual interact mean that social, psychological and existential/spiritual elements may be every bit as relevant to symptomatology as the underlying pathology. Be sure to discuss these issues with the patient trusting that those not wanting to be involved in decision-making will let you know. Refer on for psychological or pastoral help if you need to.

Box 14.1 The essentials of partnership with patients

- The patient's priorities may be very different from yours
- Try not to make assumptions about a patient's views or wishes
- Quality of life generally, but not necessarily, becomes more important than the quantity as health wanes
- Be sure to discuss costs and benefits openly and in detail

Working with uncertainty

One of the greatest fears for the chronically sick or dying is helplessness. The more a patient and family feel their agendas, wishes and hopes are being taken seriously, the better able they are to cope.

Never make rash promises or be blindly optimistic about all treatments. Patients respond very negatively to hopes that are raised and dashed. Professional denial is the single most common cause of anger expressed by patients against doctors. Avoid presenting options as unchallengeable. There are ways to communicate boundaries or margins of our uncertainty to ensure that patients are informed. Don't forget, consent is a dynamic process.

Box 14.2 The patient is at the centre of decision-making

- Work in partnership with the patient and family
- Share responsibility for making decisions
- Maximise the patient's control over decision-making
- Work in a positive framework
- Agree specific tasks
- Set realistic goals
- Review regularly
- Remain open to creative options

Discussing prognosis

Open sharing of information inevitably leads to questions about prognosis. Always speak to individuals where they are able to express emotions openly. Never do it on an open ward. Include significant others if possible and invite another professional (for example, the key nurse), who can reinforce the discussion and offer support after you have gone. Patients and families need to "re-run" many times and characteristically will hear only the first and last thing that you say.

To gauge a patient's level of knowledge, anxiety or fear, explore their understanding of the situation by starting conversations in an open way, simply by asking what they think is going on or how they feel their disease is fairing.

When you talk about time, according to the stage of illness, break the future into tangible and appropriately small blocks of time such as one to three months. The intervals chosen will depend on each case. Confine your prognostication to this period and use general terms such as better, the same or worse.

Never give a finite prognosis. Always say that the unexpected may happen. If possible arrange to review the discussion to answer outstanding questions. This will also provide the opportunity to revise an opinion.

Box 14.3 Consent is a dynamic process

- Set goals and objectives with the patient.
- Revised them as regularly as necessary – clearly it would be stupid to review weekly when a patient's prognosis is years and equally unhelpful to give a patient a 2-month appointment when you expect them to die in a few weeks.
- Particularly, review aggressive treatment to reach a specific goal immediately that objective is met: a patient's wishes may alter radically as a result of success or failure.
- Be realistic, yet at the same time be prepared to allow a patient to risk things such as travel, provided they are well informed.
- Above all plan positively: people wish to live not to exist.

Box 14.4 Answering questions about prognosis

- Choose a "safe" place
- Include significant others and/or other professionals
- Explore the patient's understanding
- Be honest; don't collude with unrealistic hopes and don't be afraid to say "I don't know"
- Be kind; allow the patient to set the limits on the discussion when exploring painful truths
- Arrange a future contact

Summary

Palliative and curative care are inescapably entwined but changeable and good practice recognises the fluid involvement of different colleagues in care. Palliative care brings: ways of talking about and engaging uncertainty, looking at care planning and dealing with the ethical difficulties around consent and refusals and is valuable at any stage of illness. Palliative care workers should be called on when necessary, not just when you have run out of options and certainly not left until a patient is actually dying.

Symptom Control

General points

The significance of symptoms

Noxious and debilitating symptoms, and pain in particular, can destroy ones quality of life sufficiently to be significant risk

factors for depression and suicide irrespective of whether they are from a disease or its treatment. Symptom control is an essential part of curative treatments.

Interventions may not be conventional, for example using a drug for its side-effect rather than its accustomed indication (for example, opiates for breathlessness), or working psychologically with a patient's perceptions to alter a symptom's impact or its threshold. This idea of a symptom threshold or altering perception may be unfamiliar, but it holds the key to effective symptom control.

Thresholds

Symptoms only become problematical when a threshold is passed and one *perceives* there to be a problem. Anything that changes perception generally (fear, anxiety, etc.) can also alter symptomatology – information (bad news, unexpected deterioration, a new complication etc.) or feelings that are unwanted (fear of failing health, death, guilt, anger, bitterness, etc.) are all examples.

Non-medical symptom control

Equally, symptom thresholds can be raised by psychological interventions and measures that calm, allow patients to unwind or promote a coping mechanism. For example aromatherapy and other complementary therapies, meditation and prayer are effective for some. Massage and acupuncture have solid evidence to support their effects on musculoskeletal and myofascial pain, as do breathing exercises and respiratory pacing in breathlessness. Diet is obviously important in nausea, vomiting and bowel control.

Pain management

Pain is what the patient says it is. Leaving emotion and "soul" for the moment, pain can be classified in several different ways (Table 14.1). This is simplistic, but practical. Nociceptive pain is usually opiate sensitive and neuropathic pain is opioid resistant.

Evaluating pain

Over 90% of pains are controllable. All pain can be improved. Take a proper history and examine the patient to establish exactly what is going on.

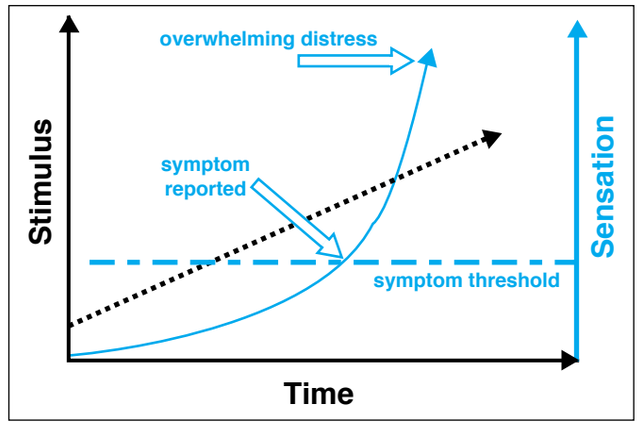


Figure 14.3 Sympton thresholds: the relationship between a noxious stimulus (for example, pain) and the reported symptom is exponential. Once past its threshold, without adequate treatment or a reduction in the process, the symptom will soon become intolerable. If the patient's threshold falls, the symptom will escalate in the same way, even when the underlying disease is stable

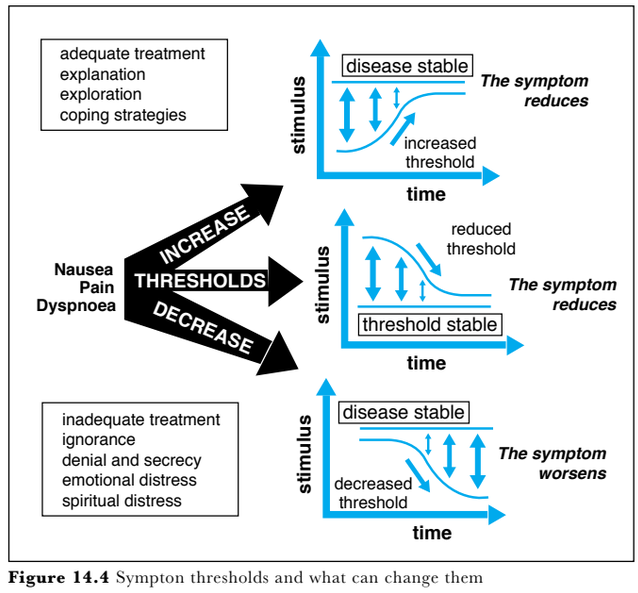


Figure 14.4 Sympton thresholds and what can change them

Table 14.1 Classification of pain

Location	"Source"		Opioid (morphine) Responsiveness			Anti-inflammatory
	Nociceptive	Neuropathic	Full	Partial	Resistant	
Visceral eg liver, bowel, myocardium, pleura	✔ ✔		✔ ✔			✗ (unless from lymph nodes)
Somatic for example, soft tissues damage, inflammatory diseases		✔ ✔		✔ ✔	✔	✔ ✔
Bone pain for example, metastases, infarction	✔ ✔	✔	✔ ✔	✔		✔ ✔
Root irritation for example, compression, inflammation		✔	✔ ✔		✔	✔ ✔ ✔ ✔
Peripheral neuropathy, cord pathology		✔ ✔			✔ ✔	

Table 14.2 Drugs used in pain management

Weaker opioids	Strong opioids	Co-analgesics	Non-opioids
DHC	Oramorph	Anaesthetic agents	Muscle relaxants
Codeine	Diamorphine	Anti-convulsants	NSAIDs
Tramadol (?)	Hydromorphone	Anxiolytics	Paracetamol
Co-proxamol	Fentanyl	Corticosteroids	COX_2 inhibitors
Oxycodone (also used as strong opioid)	Dextromoramide	Tricyclic antidepressants	
	Dipiponone		

To be able to monitor the pain it must be recorded accurately. Use body charts to localise the pain and get an estimate of each element of the pain by asking to score or grade the pain as a score out of 10 for intensity (0 = no pain, 10 = the worst pain you could imagine). This will help in monitoring treatment, but it will also give the patient a sense that pain is not fixed and can improve.

Box 14.5 Fundamentals of pain management

- Any pain, however generated, is a genuine symptom
- Always assume that there is a physical trigger, until proven otherwise
- The vast majority of patients have a combination of pain types
- The correct dose of an analgesic is that which relieves the pain

Opiates and nociceptive pain

Nociceptive pain is managed by using the World Health Organization (WHO) analgesic ladder. Over 90% of such pains are controllable in this way.

When prescribing *any* opiate, nearly all patients need a laxative and nearly 50% need an antiemetic.

Bottom rung: Paracetamol influences other drug's metabolism. However, in general, palliative physicians use doses up to 6 g a day. Similarly gastrointestinal complications (with appropriate prophylaxis) or renal impairment with NSAIDs are not absolute contraindications in patients with a short prognosis and severe pain.

Middle rung: The weak opiates, or compound analgesics (co-codamol, co-dydramol, etc.), may be helpful for mild to moderate pain. Formulations differ slightly in efficacy and prescribing is empirical. Low-dose strong opioid can be used. Dependency is less than 0.1%.

Upper rung: Many strong opiates are available. Morphine is the drug of choice by mouth. Use other opioids *only* where there is a specific problem with morphine. Alternatively using co-analgesics may reduce side-effects.

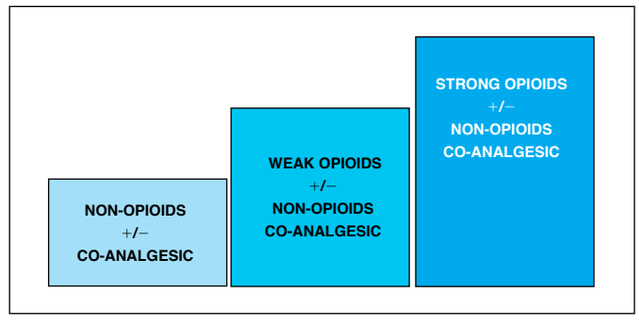


Figure 14.5 The WHO analgesic ladder

Box 14.6 Morphine facts

- Dose range is 1000-fold, (2.5 mg 4-hourly to 2.5 g 4-hourly or more)
- Most patients require less than 200 mg morphine equivalent per day
- It is not addictive when used therapeutically

Box 14.7 Opioid side effects

- *Nausea and vomiting – dose related*
- *Constipation – may be desirable, dose related*
  Peripheral effect
  Always prescribe laxatives
- *Clouding of consciousness – hallucinations rare*
  Central effect, fades with time
  Warn the patient of initial drowsiness
- *Respiratory depression – good for dyspnoea*
  Central + peripheral cough suppression
  Pain has a partially protective effect
  Increase dose carefully where chronic lung disease present
- *Itch – histamine release*

Opiate toxicity

Toxicity is idiosyncratic and dose dependent. For some the therapeutic window may be very small. In poorly controlled pain (usually non-opioid responsive), escalating doses may lead to toxicity: confusion, hallucinations, agitation and myoclonic jerks. Paradoxical pain (loss of analgesia and increasing pain) may occur.

Confusion in the dying is complex and increasing opiates is not the only solution to increasing pain. You may need to stop all opiates until symptoms have subsided. *Seek the advice of an expert*

Box 14.8 Dealing with opiate toxicity

- The half-life of morphine and diamorphine is four hours
- Respiratory rates down to 5 or 10 are acceptable for a few hours
- Central effects can be antagonised, but will lead to rebound agitation and hyperresponsiveness
- It is best simply to stop the opiate and wait
- *In extremis:* naloxone is the specific antidote and reverses all the actions of opiates. Use very small doses
- Physostigmine can be used to selectively antagonise respiratory depression

Good prescribing

Opioid responsive pain can normally be controlled within 24–48 hours. Morphine should be titrated using immediate-release formulations (Oramorph elixir, Sevredol tablets). Their effect peaks within the hour and last four hours. Titrate with a four-hourly regimen with "top-ups" for breakthrough pain of *25–50%* dose and a *25–50%* increase in the next scheduled dose as necessary. In individuals with hepatic or renal impairment the increases should proceed more slowly. Slow-release preparations (for example, MST b.d. or MXL o.d.) are best for maintenance. Immediate release formulations should continue to be available for breakthrough pain at a dose at least one-sixth the final 24-hour dose.

Box 14.9 Prescribing for nociceptive pain

- By the ladder:
  Don't forget that co-analgesics and non-opioids should be added to what you already use
- By the clock:
  NEVER use prn pain control for example, 4-hourly for morphine liquid, 12-hourly for MST
- By the right route:
  Use the mouth where possible, parenteral drug is no more effective

Opioid non-responsive pain

Neuropathic pain: Drugs affecting nerve conduction or central processing usually work. Regimens are empirical as the numbers needed to treat (NTTs) between groups are similar; benefit may take several days to gain; neuropathic analgesics have significant side-effects and dose increases should be made slowly. Patient, family and staff may need support in keeping a steady hand whilst the best combinations are found.

Intractable cases or root or cord problems (for example, CMV) may need a nerve block or long-term epidural. Many different techniques are available. They carry potential morbidity, so use cost–benefit analysis with the patient to decide a plan. Do not make choices on behalf of the patient; immobility and incontinence free of pain may be a valid choice.

Compound pains: Many patients have pain from tissue damage and the nervous system simultaneously. Their treatment requires accurate diagnosis and specific co-analgesics. Morphine may play a part in their management. To read more on this see Further reading on page 95. Make a habit of enlisting specialist support with neuropathic and compound pains.

Total body pain and suffering

This is the difficult area of suffering and the subtle interactions of our psyche, beliefs and body. Some people use the terms "soul", spiritual, or "emotional" pain. It is complex, distressing and very real (Fig 14.6). It stems out of a lowered threshold of distress and may occur with other symptoms as well. For understandable reasons, there is always an element of this in any dying person as they process and face their death and what it means. Fear and guilt are the common roots for many. Don't forget, *paene* (punishment) is the Latin root of pain.

Box 14.10 Neuropathic pain

- Generated in nervous system
- Source can be local nerve to thalamus
- Caused by:
  - Toxins for example, chemotherapy
  - Invasion/compression, for example, by tumour
  - Damage by viruses for example, HSV, CMV, HIV
  - Demyelination of any kind
  - Often coexists with nociceptive pain
  - May not present with classical dysaesthesia
- Seldom opioid responsive

Box 14.11 Neuropathic analgesics

- *Tricyclic antidepressants:*
  - Lofepramine (70 mg 1–3 times/day) or Amitriptylline (10–150 mg nocte +/– day time doses)
- *Anticonvulsants:*
  - Gabapentin (start dose of 300 mg up to 2700 mg)
  - Carbamazepine (100 mg b.d. up to 1600 mg per day),
  - Valproate (ranging from 200 mg to1200 mg per day)
  - Phenytoin (up to 300 mg per day).
- *Benzodiazepines:*
  - Clonazepam (0.5–4 mg nocte),
  - Diazepam, midazolam (parentally)
- *Membrane stabilisers:*
  - Flecainide (100–200 mg b.d.),
  - Lidocaine (lignocaine) (subcutaneous or i.v. infusion in doses of 0.5–2 mg/kg/h)
- *Others:*
  - Clonidine, octreotide, etc. seek advice

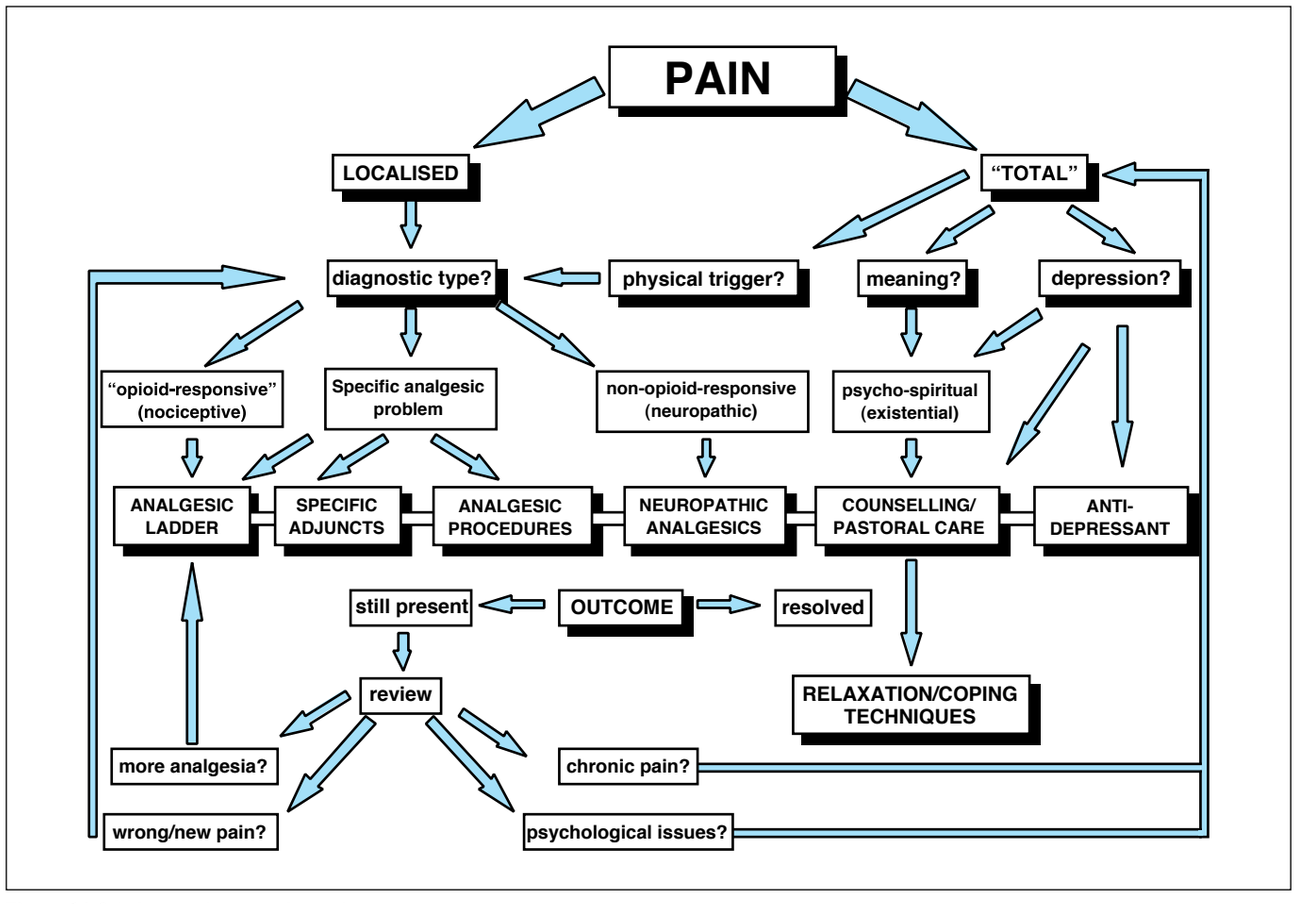


Figure 14.6 A therapeutic approach to pain

Effective management requires one to deal not only with any physical component, but also with the meaning of the symptomatology and the exacerbating effects of fear, anxiety, sleeplessness, loss of future, and of death and its connotations for the individual. In this difficult area do not be afraid to refer for help from a counsellor, psychologist or spiritual adviser.

Nausea and vomiting

These are the second most common symptoms, not least because of the burgeoning numbers of drugs that have gastrointestial side-effects. Careful assessment should ensure a logical and methodical use of antiemetics.

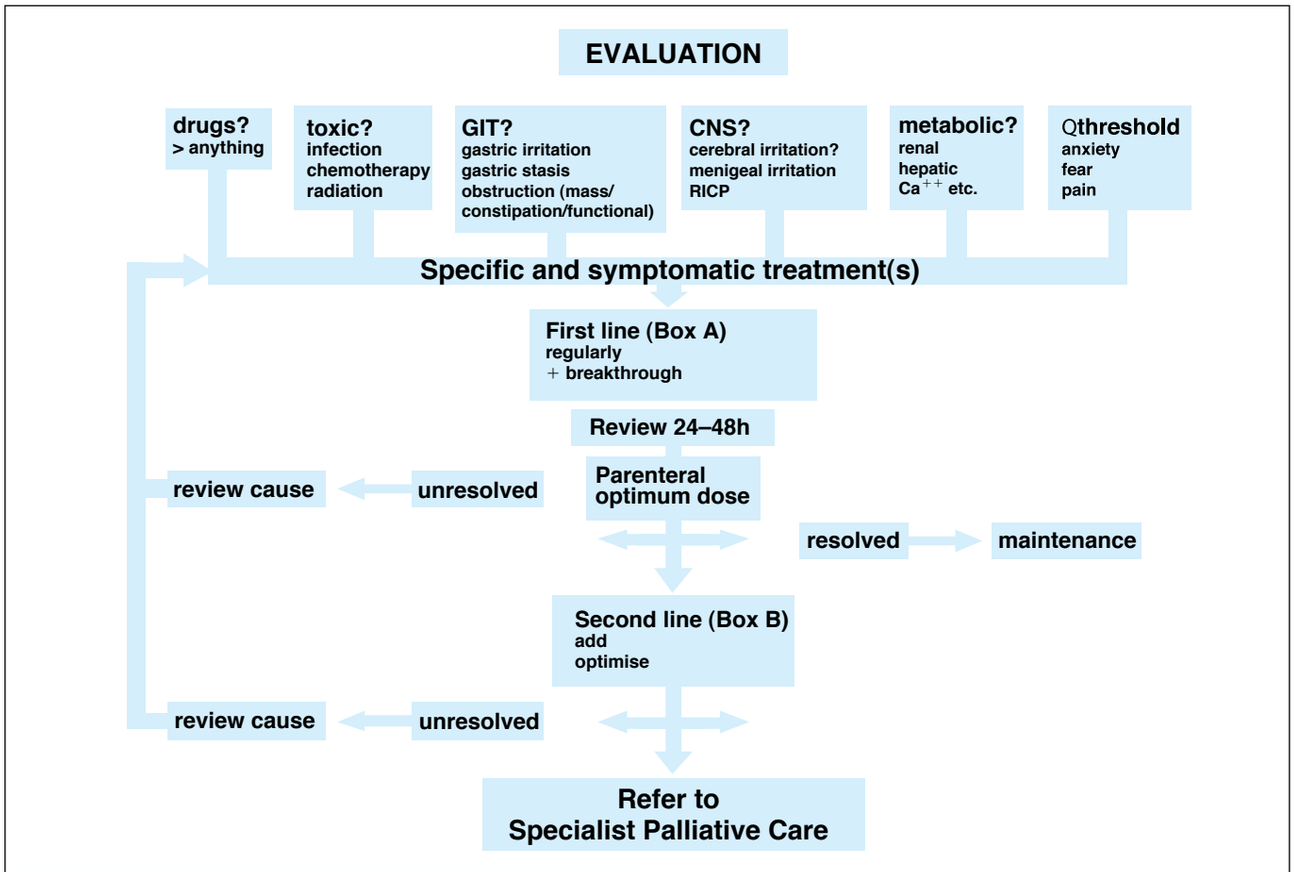
General guidelines

Nausea and vomiting is usually "uncontrolled" because of erratic and illogical prescribing in inadequate doses given orally. These guidelines are therefore common sense, but necessary.

Assess nausea and vomiting separately. Nausea tends not to respond to prokinetics. Treat any potentially reversible causes and stop emetogenic drugs if possible. Anxiety exacerbates nausea and vomiting and may need specific treatment.

Box 14.12 Logical and methodical use of antiemetics

- One-third of patients require more than one antiemetic for satisfactory control
- Select the most appropriate drug for the putative cause
- Use regularly, at optimum dose
- Have a low threshold for parenteral routes
- Careful re-evaluation on a regular basis
- Use adjuvant drugs such as corticosteroids and antisecretory drugs



Box A First-line antiemetics

Prokinetic (gastric stasis, functional obstruction)
Metoclopramide 10 mg stat., then 10 mg t.d.s. or 30–60 mg/24h via subcut. infusion (n.b. inactivated by muscarinic drugs)
Vomiting Centre (RICP, motion, mechanical obstruction)
Cyclizine 50 mg p.o./s.c./i.v. stat then 50 mg p.o. t.d.s. or 150 mg/24h via subcut. infusion
Hyoscine hydrobromide as patch 500 mg/72h
5HT$_3$ Antagonists (metabolic, drugs, radiation, gut distension)
Granisetron 1–3 mg p.o./i.v. stat and b.d. or 9 mg/24 h via subcut. infusion

Box B Second-line antiemetics

Broad spectrum
Methotrimeprazine 6.25 mg stat 12.5–25 mg/24h via infusion
Dexamethasone 4–8 mg daily as one dose mane
Prokinetic:
Cisapride 20 mg p.o. stat and b.d.
Antisecretory drugs:
Hyoscine butylbromide 20 mg s.c. stat and 60–120 mg/24h
Octreotide 100 mg stat and t.d.s. or 300–600 mg/24h via infusion
Cannabinoids:
Nabilone 1–2 mg b.d.

Figure 14.7 Guidelines for pain evaluation

Prescribe the most appropriate first-line antiemetic for the likely cause according to the figure. Prescribe both regularly and as required. If the patient is vomiting, or has been nauseous for some time administer parenterally, preferably by continuous subcutaneous or intravenous infusion preceded by a stat dose. Optimise the dose daily taking into account breakthrough doses and reported level of nausea and vomiting.

If there is no improvement, rather than changing the drug, optimise the dose, and re-evaluate the cause. It may influence your drug choice. After 48 hours, substitute or add an appropriate second-line, broader spectrum antiemetic. A significant minority of patients need more than one antiemetic. Consider non-drug treatments, including acupressure bands, control malodour, and ensure patient avoids foods that may precipitate nausea. If control remains poor, then refer. Only consider converting to equivalent oral regimen after 72 hours of good control and continue antiemetics indefinitely unless the cause is self limiting.

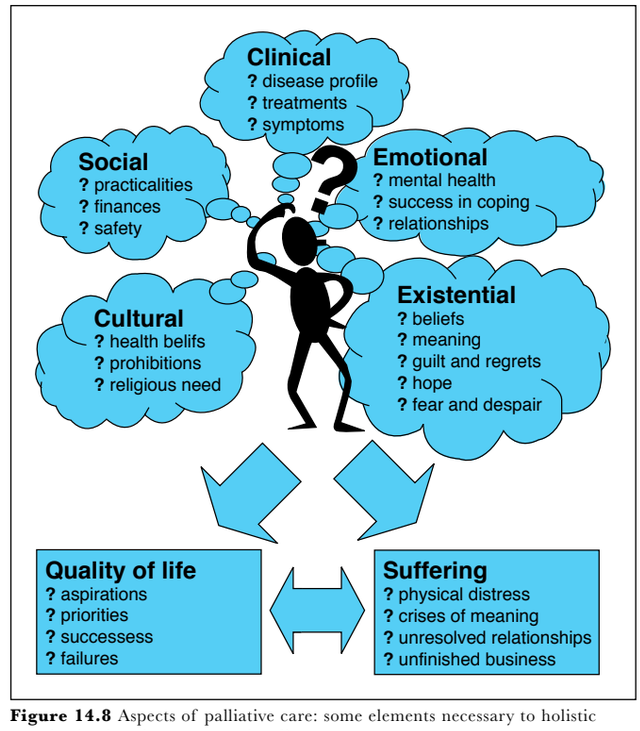


Figure 14.8 Aspects of palliative care: some elements necessary to holistic practice in chronic or progressive disease

Other common symptoms

Other common symptoms are cognitive impairment, weight loss, malaise, weakness, pruritis, cough, diarrhoea, etc. Their differential diagnosis and appropriate investigations and management are covered more widely in the *ABC of Palliative Care*.

Death and dying

When treatment is futile, persevering with treatment and investigation can be obstructive in allowing a patient a dignified and meaningful death. The patient should be at the centre of the decision-making process as much as is possible. It is at this time that the multiprofessional team is so important.

Facilitating choice

If you have been managing your patients properly and involving them in decision-making, the groundwork for managing the last weeks or months should have been done, you will have a good enough relationship to be honest and open and to finish these last preparations. If you have not faced these with your patient in some form, even by flagging that "a time will come…" whilst not failing them as a technician, you will have failed as a doctor. This is the time to check regularly about a patient's wishes. Proper links and services from primary care and social services are essential and friends, family and professionals should be as much "in the know" as possible.

Box 14.13 Preparing for death

- Do they want active treatment if they deteriorate? If so, what level of resuscitation do they want? Is there a time or circumstances in which they wish treatment to be withdrawn or withheld?
- Will they feel more in control if these are written formally? A 'Living Will' or Advance Directive can be of great help to some patients by ensuring that their wishes are known if they become incapable. (See the BMA guidelines "Advance Statements: a guidance to practitioners".)
- Have they said their goodbyes, sorrys and thank yous?
- Are there remaining personal matters to address: a will, funeral preparations, etc?
- Do they want to be at home? If so, is it suitable?

Two additional symptoms

Movement-related pain

This is a common problem in dying patients with HIV and is best managed with NSAIDs. If the patient cannot swallow, then rectal indometacin is very effective.

Pulmonary secretions

Retained secretions in patients too weak to clear them can be controlled with hyoscine 0.6–1.2 mg s.c. over 24 h or glycopyrronium 0.6–1.2 mg s.c. over 24 hours. If they fail to clear, use furosemide (frusemide). Reassure family that noisy breathing of itself is not distressing to the patient.

As death approaches

In the last days of life, pragmatism and sensitivity are essential. Patients have no appetite, are weak and somnolent or unconscious. Altered breathing patterns can last for days. Be calm and reassuring: relieve anxiety for both patient and carers by explanation that these changes are normal and don't cause physical suffering, which is true.

Most importantly continue to visit. The clinical situation can change very quickly. Assess symptoms regularly and change palliative therapeutics as necessary (even several times a day). As swallowing becomes difficult swap to parenteral routes. Most drugs for symptom control can be given continuously via a syringe driver subcutaneously. (Drugs can be mixed, see the charts in Twycross *et al.* 1998.) Figure 14.5 summarises management in the last few days of life.

With the limited communication, problems may manifest themselves as pre-terminal restlessness or distress. Possible physical and psychological/spiritual triggers need to be checked and acted on.

In general encourage the family to talk normally to the patient and to say whatever they need to say. Reassure them that the patient can hear and continue to explain all that you do to the patient and chat normally through procedures. This period of life, when the dying process is actively underway, may be short lived or take many days. In most cases we do not know what is taking place. Where beliefs are unknown or unfamiliar it is best presented neutrally as a time of transition; when our place is to care.

Box 14.14 Pre-terminal restlessness

- Exclude urinary retention
- Treat any suspected pain
- Check that there is not an important visitor that the patient must see or hear
- Check for an important date or anniversary
- Exclude any important religious rite
- Sedate as necessary; midazolam (starting at 10 mg/24 h), levomepromazine (12.5–300 mg/24 h).

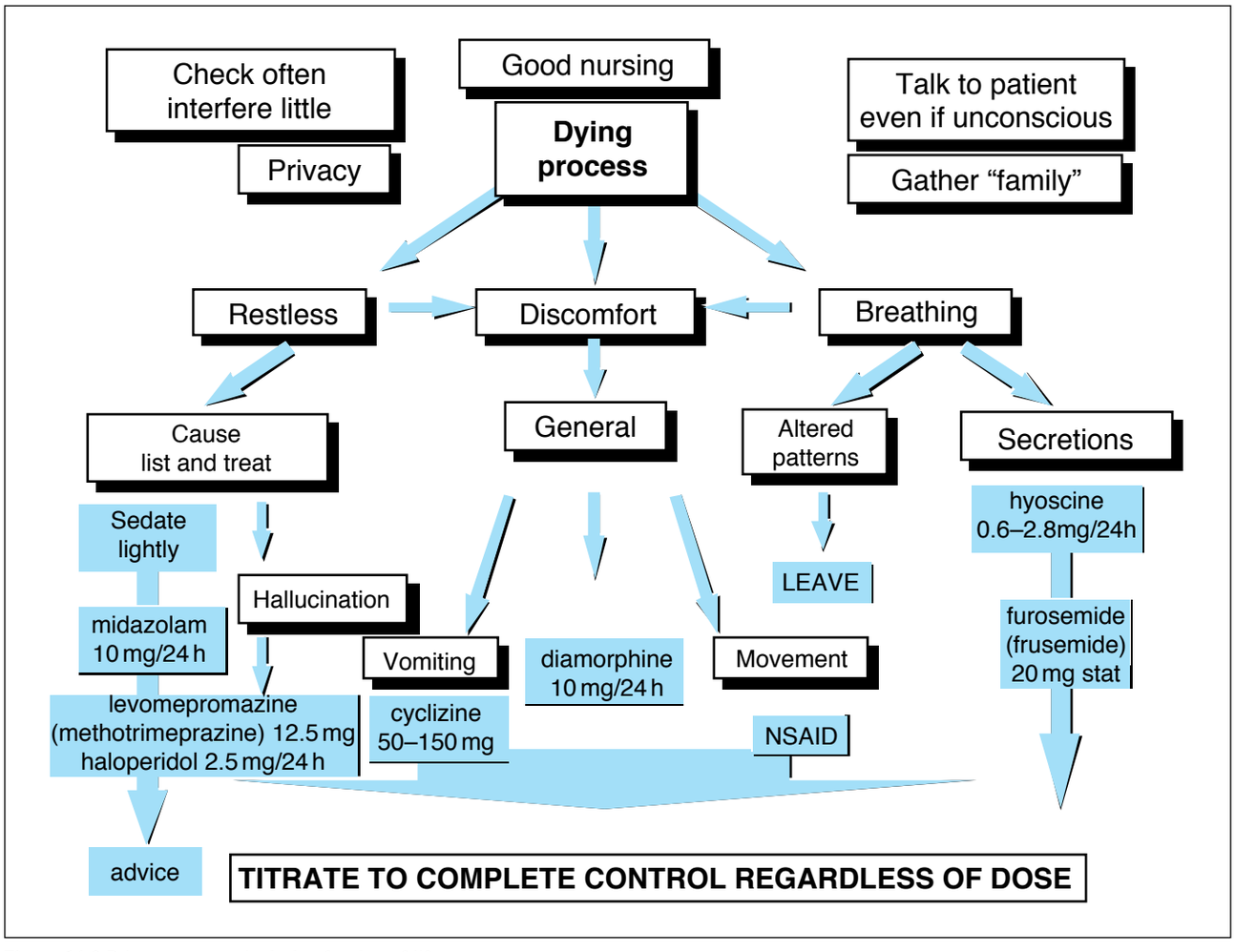


Figure 14.9 Pain management in the last few days of life

It is important to allow those with religious beliefs the opportunity to see their advisors and perform necessary rituals as they wish. This can often lead to conflict if partners and family are of differing opinions. Give time to friends and family to spend talking over what has happened. Obviously you must be aware of the dynamics of the group and you must respect the patient's confidentiality.

Finally, a death affects us and the team involved. Debriefings and supervision work either individually or as a team can be most beneficial.

Further reading

- Fallon M, O'Neill, W, eds. *ABC of palliative care*. London: BMJ Books, 1999.
- British Medical Association *Advance Statements About Medical Treatment*. London: BMJ Books 1995.
- Doyle D, Hanks G, MacDonald N. (eds). *The Oxford Textbook of palliative medicine*, 2nd edn. Oxford: Oxford University Press, 1998.
- George R, Houghton P, Robinson V. *Healthy dying*. London: Jessica Kingsley, 2001 (in press).
- Twycross R, Wilcox A, Thorp S. *PCF1, Palliative Care Formulary*. Oxford: Radcliffe Medical Press, 1998.
- British Medical Association. *Withholding and Withdrawing Life-prolonging Medical Treatment*, 2nd edn. London: BMJ Books, 2001.

15 Control of infection policies

IJ Hart, Celia Aitken

Intensive epidemiological studies of HIV infection have shown that it is not transmitted in the community by casual or intimate non-sexual contact.

As of December 1999 there have been 96 documented instances of confirmed occupational transmission of HIV. There have been, in addition, 171 cases of HIV infection, possibly resulting from occupational transmission in exposed individuals with no other known risk of infection. The rate of transmission after a single percutaneous exposure to HIV positive material is 0.32% (21 confirmed infections after 6498 exposures in 25 studies). The risk of infection after exposure of mucous membranes and/or conjunctivae to infected material is 0.03% (one confirmed infection after 2885 exposures in 21 studies).

It is important to design infection control policies which, while protecting staff against the risk of fection, do not compromise medical and dental care. HIV is one of several blood-borne viruses; carriers of these viruses may be perfectly well and individuals may be unaware that they are infected. Some, including the hepatitis viruses B and C, are potentially more infectious than HIV. Thus, healthcare workers and society in general need to adjust to the concept that direct contact with the blood of others may present a potential, albeit low, risk of infection.

In the UK the Department of Health and many other bodies have issued guidelines to educate and protect healthcare and community workers. Routine HIV screening of antenatal patients is now recommended, and testing of all those at risk is encouraged. Awareness of the risks, education, careful attention to work practices, provision of protective equipment and immunisation against hepatitis B, where appropriate, are measures which will reduce to a minimum the risk of infection with all blood-borne viruses.

Box 15.1 Selected guidelines

- United Kingdom Health Departments. *Guidance for clinical health care workers: protection against infection with blood borne viruses. Recommendations of the Expert Advisory Group on AIDS.* London: HMSO, March 1998
- *A code of practice for sterilisation of instruments and control of cross infection.* London: British Medical Association, June 1989
- *The safe disposal of clinical waste.* London: HMSO, 1992
- United Kingdom Health Departments. AIDS/HIV infected health care workers. *Guidance on the management of infected health care workers and patient notification. Recommendations of the Expert Advisory Group on AIDS.* London: DOH, March 1998
- Advisory Committee on Dangerous Pathogens. *Protection against blood borne infections in the workplace: HIV and hepatitis.* London, HMSO, 1995
- Royal College of Pathologists. *HIV and the practice of pathology.* London: Marks & Spencer Publication Unit of the Royal College of Pathologists, July 1995
- United Kingdom Health Departments. *HIV post exposure prophylaxis: Guidance for the UK Chief Medical Officers Expert Advisory Group on AIDS, July 2000.*
- General Medical Council. *Serious communicable diseases.* London: HMSO, 1997

Hospital care

HIV positivity *per se* is not an indication for isolating a patient in hospital. It may be necessary to consider source isolation, however, if there is evidence of active infection with other agents, such as *Mycobacterium tuberculosis*, varicella-zoster virus, or if there is a likelihood of extensive exposure to body fluids from, for example, haemorrhage or severe diarrhoea.

Medical practices should be of a sufficiently high standard to eliminate any risk of patient-to-patient spread of HIV in hospital. This is achieved, as part of general infection control procedures, by using disposables, and by paying careful attention to decontamination and sterilisation. Attempts to recycle disposables or to bypass accepted disinfection procedures may lead to nosocomial infection.

Staff should adopt sensible precautions if contamination with blood or other body fluids is likely. This applies particularly for the management of known virus carriers but should also be the routine for any patient. The concept of "universal precautions" for all patients is being introduced increasingly into healthcare. In most cases precautions entail no more than wearing disposable gloves and an apron, but in certain circumstances, such as bronchoscopy, protective spectacles and a mask may be necessary to protect the eyes and mouth. Most aspects of patient care and examination do not

Figure 15.1 Bronchoscopy in a patient infected with HIV

expose the staff to body fluids, and protective clothing is not required.

Many staff sustain inoculation injuries while manipulating needles and sharp instruments. Education and careful attention to technique will reduce the risks to a minimum. No attempt should be made to resheathe needles unless a safe resheathing device is available, and needles should be placed immediately into safe sharps disposal containers, which should not be overfilled.

Although there is little epidemiological evidence of increased risk, many hospitals assume that special care should be taken during surgery on known or suspected HIV carriers. This usually means adopting pre-existing policies for hepatitis B carriers and may include the introduction of double-gloving and additional protective clothing. Preventing unnecessary exposure to body fluids and trying to reduce the incidence of penetrating injuries to a minimum are the best defence against infections, which may be present, but unsuspected, in any patient.

Reports of transmission of HIV from a dentist to his patients have raised public concerns about the risks of acquiring HIV and other blood-borne viruses from healthcare workers. Guidelines produced by the UK Health Departments identify work practices known as "exposure-prone invasive procedures" as aspects of medical care that present a potential risk of transfer of a blood-borne virus from healthcare workers to patients.

Exposure-prone procedures are those where there is a risk that injury to the worker may result in the exposure of the patient's open tissue to the blood of the worker. These procedures include those where the worker's gloved hands may be in contact with sharp instruments, needle tips or sharp tissues (spicules of bone or teeth) inside a patient's open body cavity, wound or confined anatomical space where the hands or fingertips may not be completely visible at all times.

Healthcare workers who are HIV positive or HbeAg positive carriers of hepatitis B are excluded from exposure prone procedures. HbeAb positive carriers are excluded if there is $>10^3$ copies/ml of HBV DNA in their blood. There are many reports of hepatitis B transmission from staff to patients but only one report of HIV transmission from a surgeon to one of his patients during orthopaedic surgery. The risk of HCV transmission from staff to patient is still not known but may be higher than previously thought. Clearly, the risks to the patient from HIV in health care workers are extremely low but the frequency of inoculation injury to the surgeon during the course of major surgery highlights the need for continued surveillance.

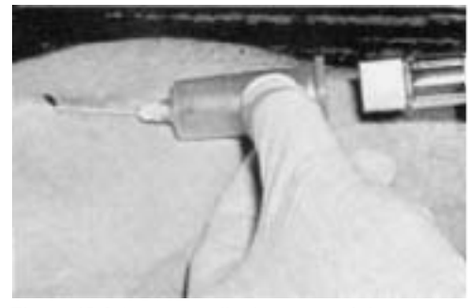


Figure 15.2 A vacuum collection system of the type shown reduces the risk of spillage when large volumes of blood are required

Sharps disposal

Clinical laboratory staff are at risk from certain pathogens which may be present in specimens. The Advisory Committee on Dangerous Pathogens originally produced specific guidelines for work on samples from HIV positive patients. These have now been reissued to encompass potential risks from all blood-borne viruses. The most important aspects of safety in the laboratory are education, training, and prevention of inoculation and skin contact with body fluids. It is important to review all laboratory procedures to reduce the use of needles and the danger of exposure to glass fragments. This may necessitate increased investment in automatic pipetting systems to replace the need for glass pipettes. The absence of evidence of airborne transmission means that HIV positive samples may be handled on the open bench providing the work is conducted in optimal facilities and the operator is free from distraction and

Figure 15.3 Safe sharps disposal

disturbance. The current practice of alerting laboratory staff to samples from known or suspected HIV positive patients by the use of biohazard stickers may be defended on the basis that it reduces risks. It must, however, be emphasised constantly that in the present epidemic no unfixed specimens can be considered free from infection.

Community aspects

HIV carriers m the community present no risk to others from normal day-to-day contact. The combined effects of dilution, temperature and detergent action ensure that standard washing procedures will satisfactorily decontaminate cutlery, crockery and clothing. All blood spillages should be decontaminated with hypochlorite (bleach) and carefully cleaned up. The absence of evidence that saliva can transmit HIV means that nobody should withhold mouth-to-mouth resuscitation from someone who has suffered a respiratory arrest. Members of the rescue services who frequently carry out resuscitation, often in cases in which facial injury exposes them to blood as well as saliva, are provided with masks and other devices. Anyone attempting to use a resuscitation device must be adequately trained as, in the wrong hands, it may prejudice the life of the casualty and in some cases increase the potential risks to the operator by causing bleeding.

Disinfection

An important method of reducing the potential infectivity of viruses is dilution. Thus procedures such as thorough cleaning and handwashing are central to any infection control policy and must never be neglected. HIV has been described as a fragile virus, and this is true to an extent. Although it is effectively inactivated by many different agents, survival of virus may be prolonged at ambient temperatures, and infectious virus may still be present in dried blood after a week. This means that any surfaces and fomites that have been in contact with clinical material must be decontaminated.

The trend towards the use of disposables reduces the need for decontamination in many areas. Thorough cleaning followed by heat sterilisation should be adopted, if at all possible, for any reusable equipment. Although HIV is inactivated by boiling, autoclaving has become the norm in clinical practice. With increasing numbers of HIV carriers in the community it is important for their protection to ensure that instruments are rendered free of all organisms, including bacterial and fungal spores. Organisms that may present no risk to people with normal immunity may lead to opportunistic infections if they are immunocompromised by HIV infection or other agents such as chemotherapeutic drugs.

Liquid disinfectants must always be considered a poor alternative to heat sterilisation. Difficulties exist controlling their potency, most are caustic, and most are rapidly inactivated by organic matter. For hospital or community use, if it is necessary to use a liquid disinfectant, it is sensible to choose one which is known to inactivate hepatitis B and other pathogens such as *Mycobacterium tuberculosis*, as well as HIV.

All waste that is contaminated with blood must be considered potentially infective and treated as "clinical waste" in accordance with the Health Services Advisory Committee's document "The safe disposal of clinical waste". Sharps containers must meet Department of Health specifications and must be incinerated before disposal.

Box 15.2 Community aspects of decontamination

- Cutlery, crockery, clothing
  decontaminated by normal washing
- Decontaminate blood spillages with bleach (hypochlorite)

Figure 15.4 Secure bagging for specimen and request sent to laboratory

Box 15.3 Disinfection

- Autoclave or use disposables if possible
- Hypochlorite (1000 ppm available chlorine) for general decontamination
- Hypochlorite (10 000 ppm available chlorine) if organic matter, including blood, present
- 2% Glutaraldehyde (freshly activated) NB: Beware of dangerous fumes

First aid and inoculation injuries

In the event of exposure to blood, simple first-aid measures should be applied immediately. Any blood or other body fluids on the skin should be washed away with soap and water. Splashes into the mouth or eye should be diluted by washing, and sterile eyewash bottles should be provided in any areas where this is likely to occur. A skin puncture should be encouraged to bleed in an attempt to express any material deposited in the wound. The wound should then be washed thoroughly. Any injury to a member of staff should be reported immediately to the person in charge and then to the occupational health physician or other medical adviser. In hospital this allows for the opportunity to investigate the state of health of the person inoculated and, if necessary, to take protective measures such as hepatitis B prophylaxis or antibiotic cover, or testing the source patient or the use of antiretroviral drugs. At present the recommended drugs for postexposure prophylaxis are zidovudine, lamivudine and indinavir. They should be taken for four weeks. An acceptable recommended alternative regimen is the use of nelfinavir instead of idinavir. However, allowances for pregnancy, drug interactions and potential antiviral resistance in the source may result in some modification to the final regimen. In these circumstances expert advice should be sought. The medical adviser should discuss whether blood samples should be taken for future reference of HIV testing and whether a programme of follow-up consultations should be started.

The medical adviser will need to obtain information about the source patient concerning possible indicators of HIV infection, including risk factors and results of previous HIV tests, medical history suggestive of HIV infection, and details of past and current antiretroviral therapy in patients known to be HIV infected. The source patient should be asked to consent to testing for HIV infection. This will entail pre-test discussion and obtaining fully informed consent. If the patient is unconscious when the injury occurs consent should be sought once the patient has regained full consciousness. If the patient refuses testing, is unable to give consent because of mental illness or disability, or does not regain full consciousness within 48 hours, testing should be considered in exceptional circumstances only, such as where there is good reason to think that the patient may be HIV infected. In this case testing an existing blood sample for HIV infection may be done but only after consultation with an experienced colleague. The decision to test may be challenged in courts so be prepared to justify the decision. Only the source patient and those exposed to the infection may be told the result of the test and the result can only be entered into the patient's personal medical record with the patient's consent. If the patient dies HIV testing can be done if there is good reason to think the source patient may be infected. It is usual to seek the agreement of a relative before testing.

Those concerned with counselling people who have sustained inoculation injuries should have enough knowledge to provide current information about the risks of occupational exposure and should be able to advise on changes in lifestyle such as the adoption of safer sex practices.

In summary, the risk of transmission of HIV within hospitals and to carers in the community is low. Education of staff, good infection control procedures and safe working practices can help to minimise this risk. Due attention to these measures at all times will ensure the protection of patients and staff.

Box 15.4 First aid

- Body fluids on skin, in eyes, or in mouth
  - wash away immediately
- Penetrating wounds
  - encourage bleeding
  - wash with soap and water
  - report to the supervisor and medical officer

16 Strategies for prevention

John Imrie, Anne M Johnson

Introduction

Limiting the spread of HIV relies on health promotion activities to encourage and help sustain behavioural changes that reduce the risk of acquiring or transmitting the virus. Despite advances, the prospect of a widely available effective vaccine remains some distance off and behavioural interventions are likely to remain the backbone of HIV prevention for the foreseeable future. Appropriate prevention strategies are required in both developed and developing country settings and must be specific to the cultural, epidemiological and socio-economic environment of each country. This chapter focuses on HIV prevention strategies in the UK although some of the principles outlined are generalisable to other countries (chapter 10). This chapter deals with sexual and parenteral transmission of HIV. The prevention of perinatal transmission is addressed in chapter 12.

General health education

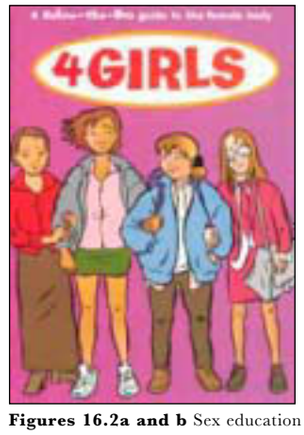
Government information campaigns and media attention in the 1980s raised the general public awareness of HIV/AIDS. Knowledge of transmission routes and risk reduction strategies (for example, condom use and reducing partner numbers) remains high, although public campaigns for HIV risk-reduction no longer have the same profile in the UK. Recent increases in sexually transmitted infections (STI) (for example, chlamydia and gonorrhoea) and high teenage pregnancy rates indicate that safer sexual practises are not consistent among young people. Age at first intercourse continues to decline and while there is some evidence of increased condom use in many countries, there has been little change in the numbers of people reporting multiple sexual partners. Those at greatest risk of poor sexual health outcomes are men who have sex with men, the under 25s, injecting drug users and their partners, inner city populations and some ethnic minority populations

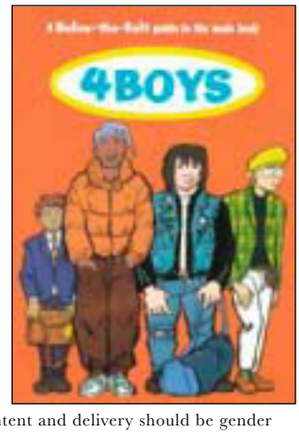
Epidemiological data show an increasing trend in the number of heterosexually acquired HIV infections diagnosed in many developed countries. In the UK in 1999, for the first time the number of newly diagnosed HIV infections acquired hetrosexually exceeded those acquired through sex between men. However the majority of heterosexually acquired infections in the UK remains among those with sexual partners in Africa (chapter 1).

These trends indicate the continued importance of general health education strategies for HIV prevention and sexual health promotion. Prevention messages can be delivered in many different settings, ranging from mass media, school sex education, community and youth organisations, through individual interventions in primary care, contraception services and specialist STD services. All health professionals can provide practical information and personally tailored messages to individuals.

Given the particular risk among young people, education for HIV prevention needs to take place in the broader context of sexual health education in schools, before young people become sexually active, as part of Personal Health and Social Education (PHSE). To remain effective over time, however, school-based sexual health and general HIV education strategies need to be

Figure 16.1 Dutch scratch card shows prevention messages can be delivered in different settings using a range of age appropriate techniques reproduced with permission from the Dutch Foundation for STD Control





Figures 16.2a and b Sex education content and delivery should be gender sensitive and take account of the different needs of boys and girls reproduced with permission from the Family Planning Association

sustained, politically supported by central and local government, financially secure, and routinely assessed and revised to meet the changing needs of new generations of sexually active young people.

Current approaches to sex education in schools include both teacher-led and peer-led approaches. Generally outcomes of sex education have been poorly evaluated and the most effective methods of delivering sex education for achieving improvements in sexual health outcomes are uncertain. However, observational studies have indicated some key components of effective sex education programmes. Several randomised trials are currently under way examining a range of approaches, and they will hopefully provide some more definitive answers.

Box 16.1 Approaches to sex education most likely to improve sexual health outcomes in young people:

1. Begin early (i.e. sex education should start with pre-teens)
2. Cover issues in an incremental and age-appropriate fashion
3. Address knowledge and attitudes, and provide practical skills (for example, using condoms)
4. Provide information, improve knowledge and build confidence to access sexual health and contraceptive services
5. Employ participative approaches (for example, role play)
6. Ensure content and delivery are gender sensitive, taking into account the different needs of boys and girls
7. Ensure understanding of different sexual choices (for example, delaying first intercourse, resisting pressure for sex) and different sexualities
8. Deliver interventions in a range of settings across the community (for example, involve parents and youth services)

Preventing sexual transmission

The epidemiology of HIV within the UK indicates that the greatest risk of infection is still associated with particular behaviours or demographic characteristics. Identified behaviours with the highest risk of HIV infection are: sex between men, injecting drug use; sex with injecting drug users, and sexual contact in parts of Africa and other parts of the world, where heterosexual transmission predominates. In other parts of the world commercial sex workers are at greatly increased risk of HIV. In the UK, other than among sex workers who are also injecting drug users (IDUs), high rates of condom use with commercial partners have maintained low HIV prevalence among prostitutes.

While there has been massive expenditure on HIV prevention over the last decade, until recently there has been a dearth of high-quality evaluation and little evidence from randomised trials to demonstrate effectiveness of different interventions. However, there is now a growing evidence base to support targeted HIV prevention interventions, tailored to the cultural context and needs of particular groups. A small number of randomised trials have shown the interventions to be effective in reducing the frequency of specific risk practices (for example, unprotected penetrative vaginal or anal intercourse) and, in a few cases, the incidence of new STI. In general, these interventions have aimed to provide basic HIV/AIDS education (including instruction on correct and appropriate condom use), enhance motivation for behavioural change, and teach risk reduction and safer sex negotiation skills (including the ability to resist pressure for sex) and have been delivered in community, small group and individual settings.

However, effective interventions in a research setting may not yield the same results in "real life". Careful consideration of local HIV epidemiology with a critical view of the generalisability of the intervention, will help to determine whether a specific intervention is appropriate and prevent spending limited resources on a programme that shows little benefit, or worse still, a negative effect. The literature contains examples of both.

No single intervention strategy is likely to be sufficient to address all of a group's prevention needs. There is no evidence that "single-shot" prevention interventions have enduring effectiveness at a population level. Interventions need to be sustained, with careful monitoring to indicate when changes are necessary, and must adapt, particularly, to the evolving epidemiological, social and cultural changes in successive new generations.

Little has changed with respect to the core content of prevention messages: it requires sexual contact involving the exchange of body fluids or blood-to-blood contact for transmission to occur. Those who know they are HIV negative and in a mutually monogamous relationship, are not at risk of

Box 16.2 Practises that reduce the risk for acquisition or transmission of HIV

- Using condoms for all penetrative sexual intercourse
- Using adequate quantities of water-based lubricant for both vaginal and anal intercourse. (Oil-based products will cause latex condoms to perish. Lubricants containing spermicides (for example, Nonoxyl 9) may cause irritation and have not been demonstrated to be effective in reducing HIV transmission *in vivo*)
- Reducing numbers of sexual partners
- Adopting sexual practises that carry a lower risk for HIV transmission (for example, oral sex, mutual masturbation)
- Avoiding recreational drug use during sexual activity, or when sex is likely to happen
- Ensure timely screening and treatment for suspected STI
- For young people, delaying the age at which first sexual intercourse takes place

infection through sex. To limit sexual risk of infection, the most effective strategies are to reduce numbers of sexual partners, know about partners' previous sexual and drug-use history and adopt safer sex practices (for example, oral sex, mutual masturbation and use condoms). Although condoms do not provide total protection, correct and consistent use will substantially reduce the sexual risk of HIV, STI and pregnancy.

The challenge that remains is how to deliver innovative HIV prevention messages through a range of different community, and individual focused interventions to reduce HIV transmission.

The following sections examine some effective strategies in relation to specific populations at high risk.

Gay, bisexual and other men who have sex with men

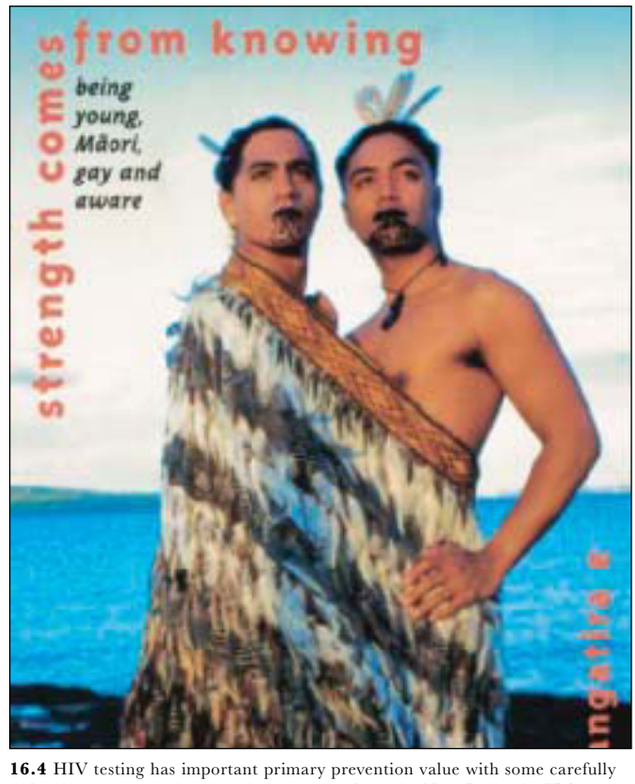
In most industrialised countries, homosexual and bisexual men have been disproportionately affected by HIV/AIDS. Unprotected anal sex is the primary mode of transmission and receptive intercourse carries the greatest risk. The success of early safer sex promotion campaigns primarily led by gay organisations has been highlighted as one of the greatest early successes in HIV prevention with evidence of falling STI rates, stabilisation in HIV prevalence and rapid uptake of safer sex practices. Over time these changes have proved difficult to sustain. Although condom use has become a social norm within the gay community, in recent years increasing proportions of homosexual men are engaging in unprotected anal sex. This particularly involves sex between men who are known to be HIV positive, partners whose HIV status is unknown to each other and younger men (< 25 years) who are likely to have become sexually active in the era of HIV/AIDS, and were not exposed to the intensive campaigns of the mid 1980s.

Factors which may contribute to increasing risk behaviour include: "boredom" with prevention messages; failures to target appropriate messages to a new generation of gay men; and perceived decreased threat of HIV in the era of highly active antiretroviral therapy (HAART).

The content of the successful interventions targeting gay men varies, but have often included motivational training, audiovisual presentations (for example, eroticising safer sex), brief safer sex negotiation skills training, stress reduction training and intensive group counselling. Interventions such as these are likely to attract individuals with particular concerns about their sexual risk behaviours and greater motivation to address them, so such interventions alone are not likely to meet all of the prevention needs of men who have sex with men. These interventions are particularly relevant to men using genitourinary medicine (GUM) clinics and HIV testing services.

Knowing that face-to-face interventions will never be able to reach all of the people at risk, community level strategies offer the potential of reaching those who do not attend services. Broader strategies focused at the community level have been shown to be effective in reaching higher risk and vulnerable men who often do not participate in small group interventions. Prevention programmes involving outreach workers and peer-educators can be used to target men using community venues (for example, bars, clubs, saunas), public sex environments (for example, cruising grounds, public toilets) and other places where homosexual men meet to have sex. These strategies often follow the principals of empowerment or community-building models of health promotion, with the intervention being developed either by gay communities themselves, or in collaboration with public health or sexual health providers. The

Figure 16.3 Face-to-face interventions may be particularly relevant for persons attending health care and other services reproduced with permission from the Terrence Higgins Trust



16.4 HIV testing has important primary prevention value with some carefully defined groups reproduced with permission from the New Zealand AIDS Foundation

content of successful community-focused interventions varies, but among the most important components are peer and opinion-leader delivery of risk reduction messages, community-building activities and peer-outreach providing safer sex materials (i.e. condoms and lubricant). One of the very few rigorously evaluated and effective interventions that specifically targeted young gay men (for example, aged 18–29) was developed using these approaches.

Injecting drug users

HIV transmission between injecting drug users (IDUs) occurs primarily through sharing of HIV-contaminated syringes, needles and injecting equipment. IDUs and their partners are also at risk through sexual transmission. Since many, particularly female, IDUs support their drug habit through commercial sex they may be at risk of sexual transmission both to and from their commercial and non-paying partners.

The epidemiology of HIV infection in IDUs and the social and cultural context of drug use vary substantially between geographical areas. Identifying promising interventions most likely to succeed within a particular setting is reliant upon understanding the local epidemiology and drug-use culture.

Preventing HIV transmission in injecting drug users relies primarily on reducing the frequency of sharing needles, syringes and other paraphernalia used for injecting ("works"), and on ensuring that the risk of sexual transmission for paying and non-paying partners is minimised through safer sex practices. Effective strategies that reduce the risk of HIV transmission through injecting will have other benefits in reducing the incidence of other viral infections (for example, hepatitis B and hepatitis C).

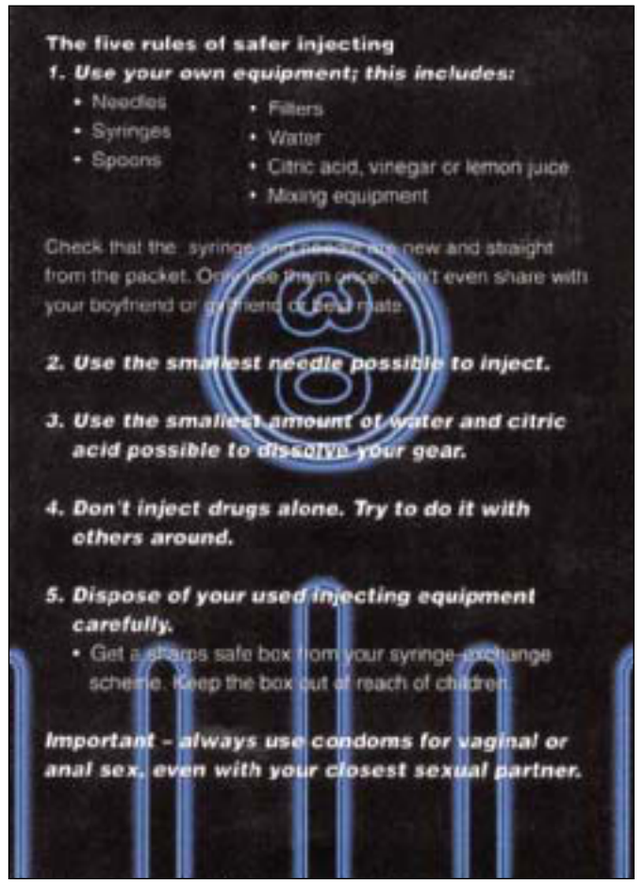
Social, political and legal controversies have hampered prevention strategies to minimise the potential harm of injecting drug use to both the individual and the community, because of particular concerns that increasing the supply of clean injecting equipment would encourage injecting drug use. Research evidence from largely observational evaluations has shown these concerns to be largely unfounded.

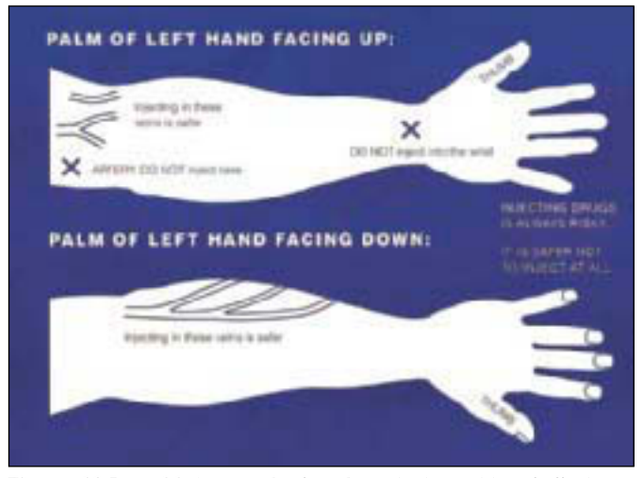
Observational studies have demonstrated that needle exchange programmes (i.e. providing sterile needles and syringes in exchange for used ones) are the most effective base for prevention strategies with drug users. Needle exchange has been successfully delivered within health and social services, through outreach workers, and dispensing machines, and has been demonstrated to be associated with reduced HIV prevalence without increasing levels of drug use. Improved access to bleach cleaning kits (for shared needles and syringes) and training in effective cleaning procedures may reduce HIV transmission through needle sharing. However, the quality of available products and the complex skills required make this a poor substitute to access to clean needles, but better than no intervention at all.

Evidence supports outreach and peer-educators as the most effective way to reach drug users in the community. Former injectors and current injectors have been employed successfully in both roles. Other interventions, specifically low-threshold easy-access drug treatment programmes and oral methadone maintenance, have been shown to reduce overall levels of drug injecting. These interventions bring drug users into regular contact with service providers (whether outreach or service based), where opportunities to deliver other information, education and counselling interventions exist. In particular, treatment and methadone maintenance programmes can offer adjunct social, educational and rehabilitation interventions to break the cycle of drug use and increase the possibility of an individual's integration into routine employment and

Box 16.3 Effective HIV prevention strategies targeting injecting drug users

- Making easily available sterile needles and syringes
- User-friendly, low-threshold drug treatment programmes, including oral methadone maintenance
- Sustained education through outreach programmes and peer education providing information, skills (for example, safer injecting), health services and social support
- Providing access to counselling and HIV testing
- Facilitating access to health care, support and STD services for IDUs with HIV infection
- Special programmes for high-risk subgroups (for example, sex workers, prison inmates, youths in detention)





Figures 16.5a and b An example of good practice in provision of effective accurate information for injecting drug users, reproduced from "A Guide to Safer Injecting". with permission from HIT

mainstream culture. A supportive environment including political, financial and legal support for the programmes at both central government and the local level is essential for the long-term success of comprehensive programmes.

In parts of the world such as the USA where IDUs have suffered a particularly severe HIV epidemic, sexual transmission to the partners of IDUs is a major source of increasing heterosexual transmission. Programmes to prevent sexual transmission among IDUs have received much less attention than those for harm minimisation from injecting and there has been little demonstrable success in changing the sexual behaviour of IDUs. Approaches to changing behaviour in this population clearly need to incorporate those shown to be appropriate to all heterosexual populations. In addition, approaches that have shown some promise specifically among drug users include skills training in correct condom use, voluntary HIV testing and counselling, and sexual negotiation skills. Such programmes need to target in-treatment drug users, sex workers and female sexual partners of male drug users.

African and other ethnic minority communities

African and ethnic minority communities have become an important focus for targeting HIV prevention interventions, however evidence for effective interventions with this group is limited. The few interventions that have been rigorously evaluated have been developed in careful collaboration with the affected communities. They have considered carefully the cultural and social factors influencing sexual attitudes of the communities and treated HIV prevention within the context of wider sexual health, contraception and pregnancy. Within the UK context recent research into the sexual attitudes and lifestyles of diverse ethnic communities has provided guidance for the development of linguistically and culturally appropriate strategies. African communities in the UK are particularly severely affected by HIV. Many within these communities also face the additional challenges of relatively recent migration including problems of language, culture and isolation often along with possible economic and/or legal difficulties associated with refugee status. All these difficulties may in turn limit access to local service for treatment and HIV prevention.

Strategies for people with HIV infection

Until recently targeting prevention interventions to HIV positive individuals has been largely neglected. Affected communities have been understandably concerned about stigmatisation and discrimination, while those responsible for prevention have felt poorly equipped to tackle many of the key issues. As stigma and exceptionalism associated with HIV diminishes, opportunities emerge to build HIV prevention strategies where people living with HIV are partners in development and delivery of the interventions. The advent of highly active antiretroviral therapy (HAART) has led to improved survival and thus to an increasing number of people living with HIV in the population. This brings with it particular public health challenges for individuals and society. Clinicians and policy makers need to be aware that with widespread use of HAART comes responsibility for ensuring that the risk of transmission, and particularly the transmission of resistant or virulent strains, is minimised and that public health is protected. For those living with HIV, HAART may lead to improved quality of life and sexual relationships and increased longevity, but also raises the challenge of maintaining life-long safer sex

Box 16.4 Guidance for enhancing sexual health promotion and HIV prevention in minority ethnic communities

- Facilitating access to appropriate confidential adolescent and adult sexual health and HIV prevention services, including specialist services outside routine clinical settings inline with the expressed needs of the community
- Developing materials using appropriate language and images including materials appropriate for non-native English-speakers
- Early and continued sex education in schools to supplement and support provision in the home
- Assisting parents from cultures where sex in general is rarely discussed to discuss sex education
- Providing focused interventions for young boys in either school or community settings
- Prohibitive messages may be supported by some particularly older generations
- Exploiting and explaining the wider benefits of safer sex in relation to contraception and avoidance of other infections may increase the overall acceptability of messages with all audiences
- Focused work exploring assumptions made about "safe" partners and concurrent relationships in cultures where they are common
- Use of appropriate and community-specific delivery points, for example, settings appropriate to the specific culture
- Awareness of different migration, refugee and acculturation experiences between communities and between generations
- Promoting HIV testing within high-risk ethnic communities is likely to be extremely sensitive and should be treated with careful consideration and caution based on a clear understanding of the individual and community issues

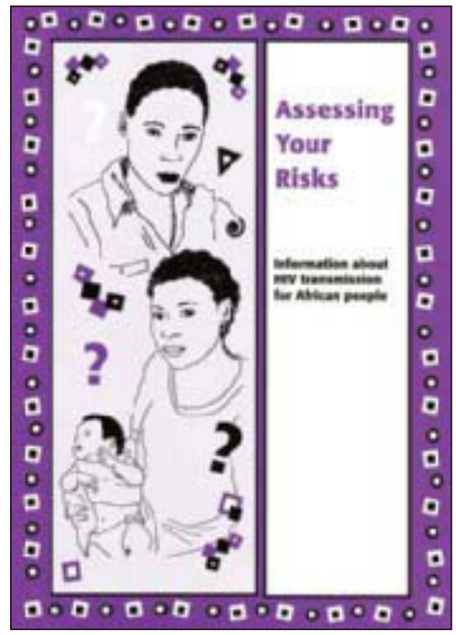


Figure 16.6 Key materials should be available using appropriate language and images for minority groups and for non-native English speakers

practices to avoid infecting others. Collectively, increased survival leads to a larger pool of infected people in the community who may pass on infection and there is already evidence that new HIV infections may once again be increasing in parts of the USA. While there are theoretical reasons to believe that HAART may decrease infectivity by decreasing viral load, at the population level, such gains may be counterbalanced by increased unsafe sexual behaviour, increased incidence of STIs, and the emergence of drug-resistant strains amongst those failing therapy, which in turn may lead to new infections resistant to currently available therapies.

Prevention trials specifically with those living with HIV are scarce. However, recent research and community consultation has provided indications of acceptable primary prevention approaches. Acceptable primary prevention strategies with HIV positive people include providing counselling and support in both one-to-one and small group contexts, providing specialist sexual health and STI screening services for HIV positive people, and offering social, emotional and sexual counselling support within HIV outpatient treatment services. As with other groups, interventions are more likely to achieve success if they occur in genuinely productive partnerships with leadership from the affected communities to overcome wider social prejudice and stigmatisation.

Voluntary counselling and HIV testing

Diagnosis of infected individuals has an important role in secondary prevention, because it allows infected individuals to benefit from treatment to reduce the chance of progression to severe immunodeficiency. Identifying those who are HIV positive in order to work with them to prevent onward virus transmission is also fundamental to primary HIV prevention.

Routine HIV antibody testing of pregnant women is now recommended throughout the UK. Positive women can then benefit from antiretroviral therapy to prevent perinatal transmission of HIV and advice to avoid transmission through breast feeding. Detailed recommendations on the management of HIV positive pregnant women is dealt with in chapter 12.

HIV counselling and testing is widely available in many clinical settings in the UK, particularly in genitourinary medicine (GUM) clinics. Counselling for HIV testing was originally developed in the pre-antiretroviral therapy era and much of the content focused on the nature and interpretation of the test, and the advantages and disadvantages of knowing ones' status in the context of an untreatable infection. All clients were also advised on risk-reduction strategies to prevent the acquisition or transmission of HIV through sex or injecting drug use. In the era of HAART, GUM clinics are increasingly offering routine testing and counselling as part of their clinical services in order to identify HIV positives.

The effectiveness of testing and counselling services in achieving behavioural change for primary prevention is limited and somewhat confusing. Brief client-centred counselling has been shown to be an effective strategy in reducing future STI acquisition in only one large-scale trial, while another study demonstrated its effectiveness in different developing countries using behavioural change endpoints. However, two major reviews of the effectiveness literature have concluded that testing and counselling are only effective as primary prevention strategies for achieving sexual behaviour change within carefully defined groups, including in-treatment drug users, commercial sex workers and post-test counselling and support for those who receive a positive result. Nevertheless, for its secondary prevention benefits and primary prevention value with specific groups, voluntary

Contact addresses and numbers for further information

- National AIDS Helpline: 0800 567123
- Health Development Agency, Trevelyan House, 30 Great Peter Street, London SW1P 2HW. Tel: 020 7222 5300
- National Aids Trust, New City Cloisters, 196 Old Street, London EC1V 9FR. Tel: 020 7814 6767
- The Terrence Higgins Trust, 52/54 Grays Inn Road, London WC1X 8JU. Helpline: 020 7242 1010 (noon to 10 pm every day)
- The Haemophilia Society, Chesterfield House, 385 Euston Road, London NW1 3AU. Tel: 020 7380 6000.
- DrugScope, Water Bridge House, 32–36 Loman St, London SE1 0EE. Tel: 020 7928 1211
- Cardiff AIDS Helpline (10 am to 8 pm Mon-Fri). Tel: 01222 223443
- Northern Ireland AIDS Line, Belfast (7 pm to 10 pm) Mon–Fri. Tel: 01232 326117
- The Sandyford Initiative, 6 Sandyford Place, Sauchiehall St, Glasgow G3 7NP (8.30 am to 4.30 pm Mon, Wed, Fri; 8:30 to 7 pm Tue and Thur.) Tel: 0141 211 8601
- London Lesbian and Gay Switchboard. Tel: 020 7837 7324
- Gay Men's Health, 10A Union Street, Edinburgh EH1 3LU. Tel: 0131 558 9444

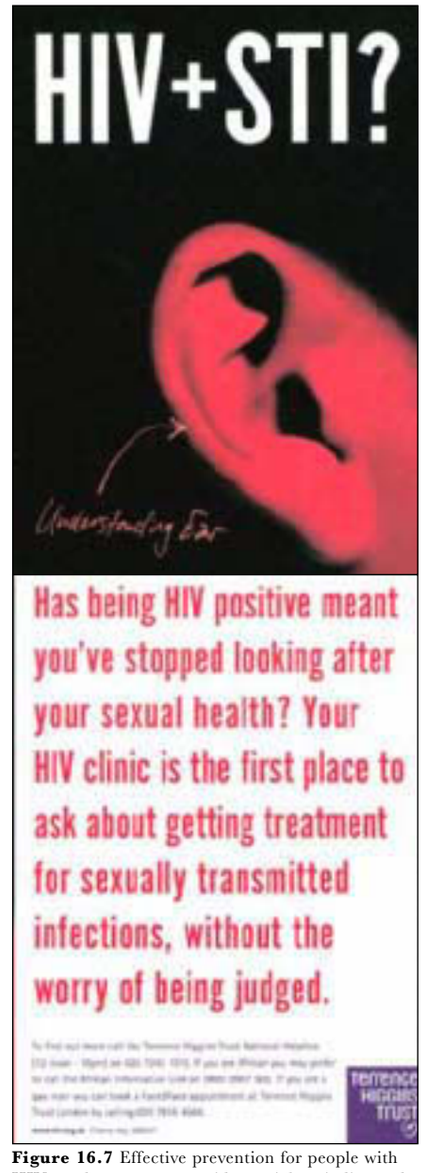


Figure 16.7 Effective prevention for people with HIV needs to overcome wider social prejudice and stigmatisation, reproduced with permission from the Terrence Higgins Trust

counselling and HIV testing is still an important component of any comprehensive HIV prevention strategy.

The testing scenario has much to offer with respect to individually focused prevention. The process of HIV testing offers an opportunity to use a client-centred counselling approach to undertake an individual risk assessment, discuss and develop individually tailored personal prevention strategies and consider the implications of a positive result.

Control of sexually transmitted infections (STIs) and STI screening

There is now substantial evidence from observational, biological and intervention studies to show that STIs (both ulcerative and non-ulcerative) may increase the susceptibility of uninfected individuals to HIV and increase the infectiousness of HIV positive individuals. Control of STIs therefore has an important role in the primary prevention of HIV. In the UK, the network of GUM clinics provides open-access services for screening, treatment and partner notification for STIs. STI control is particularly important among populations at high risk of HIV infection. Screening and treatment offer an opportunity to focus behavioural interventions on those who have STIs. Increasingly, GUM clinics are recognising the importance of offering regular STI screening as part of routine HIV treatment services alongside appropriate counselling on risk-reduction strategies.

In developing countries, where the burden of untreated STIs is much greater and diagnostic and treatment services more limited, syndromic management approaches have been used. These combine clinical history with knowledge of local pathogens to devise treatment algorithms. Such strategies however appear to be most effective in terms of their specificity and sensitivity in identifying STI cases, where the prevalence of STIs is high.

Blood transfusion and blood products

In,the UK and other developed countries, the risk of HIV transmission through blood transfusion has been minimised by testing all blood samples for HIV antibody and excluding those at increased risk from HIV from donating blood. The current categories for exclusion from blood donation in the UK are shown in Figure 16.8.

Travel to countries with high HIV prevalence

In some countries in Africa, HIV prevalence in the general population exceeds 20% and STI rates are much higher than in the UK. Unprotected sex is therefore associated with a high risk of both HIV and STI infections. All travellers to these countries need clear advice on sexual risk reduction through limiting sexual partnerships and always using condoms. There is no risk of transmission from casual contact. However, in some countries, HIV screening programmes for blood transfusions are not always in place, and there may be shortages of sterile medical equipment for injections and intravenous infusions. Sterile needle packs, first-aid kits (including needles, syringes and suture packs) and minor surgery kits are available for purchase or mail order from the Hospital for Tropical Diseases (London) Travel Clinic (2nd Floor Mortimer Market Centre, off Capper Street, London WC1E 6AU, tel. 020 7388 9600). Basic needle packs and larger made-to-order first aid and surgical packs can be ordered from Nomad Travellers Store and Medical Centre (3–5 Wellington Terrace, Turnpike Lane, London N8 0PX, tel. 020 8889 7014).

Figure 16.8 Government advice to those who should not give blood. Reproduced with permission from the Department of Health Publications

17 Being HIV antibody positive

Jonathan Grimshaw

Late in, 1984, when I was tested, HIV had only just been identified as the cause of AIDS. There was no formal counselling before or after testing, no organised emotional or social support in the community and certainly no prospect of treatment. The doctor who gave me the positive test result told me, kindly, that I seemed the sort of person who would be able to cope. I agreed. Never having been confronted by anything like this before I was ignorant of what "coping" would involve.

Initial reactions

I was very frightened. I was convinced I was going to die painfully and soon. I felt very alone. I knew no one else in the same situation. Public fear of AIDS and stigmatisation of "AIDS carriers" were at their height. Confiding in people, even friends, risked hostility and rejection, but I knew equally that friendships would not survive the level of deceit needed to conceal something so devasting. I thought I would never know sexual intimacy or love again. At that time, safer sex was not common behaviour; asking for it could raise the suspicion in a potential partner's mind that you "had AIDS" and no one, I thought, could possibly want to be intimate with or have a relationship with someone who had the "AIDS virus".

I expected, through illness, to lose my income, my security, my independence, my dignity and my self-esteem. I came to realise how much the things that give a life meaning and purpose – aspirations, dreams, motivation, hope, endurance, fulfilment – depend on the unconscious assumption of a future. Coping with HIV meant firstly coming to terms with the loss of that assumed future and secondly trying to give life some meaning and purpose in its absence. This comes with hindsight. At the time I couldn't cope at all and spent much of the first few weeks after diagnosis drunk or tranquillised.

Peer support, counselling and referral

At the end of 1984 the Terrence Higgins Trust established its first support group for people diagnosed with HIV. It gave me and the others there a safe environment in which, for the first time, we could talk openly and honestly about what had happened to us. Most importantly, hearing other people describe feelings and experiences almost identical to one's own made each of us realise that we were not alone. Learning that the frightening and unfamiliar extremes of fear, anger and grief that each of us had felt were a common and natural reaction to the situation we were in was the first step in our being able to see ourselves again as normal people rather than the "AIDS carrier" pariahs of popular perception.

The potential psychological and social impact of a positive HIV antibody test result are now well understood, as is the importance of counselling and referral to agencies that can support people emotionally and practically as they come to terms with the diagnosis and its implications for their lives. For many people, peer support continues to be a key part of that process.

Coping with uncertainty

It took some time for it to sink in that the positive result wasn't necessarily a sentence of imminent death, but no one could tell me how long I had to live. In many ways an AIDS diagnosis would have been easier; it would have given me something concrete to deal with. Being HIV antibody positive was a kind of limbo where you knew the axe would fall, but never when.

How people cope with this kind of uncertainty probably reflects how they cope with uncertainty in other areas of their lives; some avoid thinking about the future if it threatens contentment in the present, some throw themselves aggressively into trying to shorten the odds in their favour, some fatalistically assume the worst and prepare themselves for it.

My way of coping was to throw myself into community work developing services for people with HIV and prevention campaigns and establishing Body Positive. This was the first self-help group in the UK, and perhaps the world, for people with HIV. If I couldn't fight the HIV inside me, I could at least fight the HIV outside me. I became very driven because, like many people confronted at a relatively young age by their mortality, and not knowing how long I had left, I didn't want to die insignificantly. I had a lot to achieve with perhaps very little time.

One would very occasionally hear someone with HIV say that the diagnosis was the best thing that had ever happened to them. More than any other event or crisis, it forced them to think about what was important and re-arrange their lives accordingly. Certainly, the years after my own diagnosis were lived with an intensity and with a sense of fulfilment in my work that would probably not have been achievable without HIV to concentrate the mind.

Retirement

In the early 1990s my CD4 count, which had been declining very slowly over time, suddenly seemed to plummet and I developed some minor illnesses. In fact, the CD4 count never fell below the lower limit of what would be considered a normal range, but I convinced myself that the suddenly rapid decline meant that the deterioration to AIDS had begun. I retired from work, cashed in my pension and bought a nice place by the sea in which to pass my remaining few years. I had achieved what I needed – to feel that I had done something useful with my life – and I was completely ready for death.

After leaving work, my CD4 count stopped declining and I remained well. In retrospect, the retirement was probably necessary as I was almost certainly approaching "burn-out", but it felt at the time as though HIV had fooled me into a premature withdrawal from life.

During 1997 my viral load started to double every three months and I began combination therapy. Since then my CD4 count has dropped below 500 only once, when I became resistant to one of the drugs. Since changing the combination, my viral load has been undetectable.

Living with HIV in the era of combination therapy

It is sometimes assumed that combination therapy has transformed the lives of people with HIV. Well, yes and no. In people who are HIV antibody positive it can postpone illness or an AIDS diagnosis. But in doing so it prolongs the uncertainty. The long-term efficacy of antiretroviral therapy is unknown.

This brings dilemmas of its own. For example, many healthy people in mid-life seeing an advertisement for a pension plan might wish they could put more money aside for their old age. But an HIV antibody positive person has to ask him/herself

"do I spend money and enjoy life now because there may not be an old age to save for, and risk impoverishment if there is; or do I save for old age and risk lying on a hospital bed in a year or two's time regretting not spending my money and living life to the full while I was well?".

I can only imagine how much more acute and agonising dilemmas of this kind – involving trade-offs between present and future – must be in families where a parent and possibly also a child has HIV.

There are other trade-offs, some more difficult than others. Never, for example, during all the years before combination therapy did I have to adapt my life to a medication regimen. It took some time to learn full adherence to the regimen, initially, I would simply forget very occasionally to take a dose when due. But, more fundamentally, adherence involves restricting freedoms that most of us take for granted – to eat what you want when you want, for example. It was difficult to adjust to my freedom being compromised by the treatment rather than the disease itself, although viewed in the light of the benefits of the treatment, these compromises were insignificant.

Although public education has removed much of the fear and prejudice surrounding HIV and AIDS, there are communities where HIV remains highly stigmatised and where people with HIV are discriminated against. Discrimination, real or perceived, restricts the choices one is able to make in life; it limits life's potential – a cruel irony when medicine has found ways to prolong life with HIV.

Nor does combination therapy remove anxieties about falling in love and sexual intimacy. There is still the fear of revealing one's status to a potential partner in case of rejection. Although my viral load is currently undetectable, I can't assume that I'm not infectious. I must still insist on safer sex. The social acceptance of safer sex as normal, or at least sensible, behaviour means that asking for it is less likely to be met with rejection, but having sex with someone entails a risk, however small, that unsafe sex could occur. I know from experience that it isn't always possible to be totally in control of an activity in which someone else is playing an equal part. However much I rationalise that preventing transmission is a shared responsibility, because everyone has a responsibility to protect themselves, and that anyone wanting unsafe sex is probably HIV antibody positive themselves, I know I would feel a tremendous sense of guilt and failure of moral responsibility if unsafe sex did occur.

The HIV "veteran"

I was aware before combination therapy arrived that I had remained well for an unusually long time since diagnosis. Now it seems that combination therapy may keep me alive and possibly well for many years more. During the millennium celebrations it occurred to me that, if adulthood begins at 18 years, I have lived with HIV for over half my adult life.

I read recently that there is sometimes a striking similarity in how long-term survivors of HIV and war veterans describe their feelings about life. Both have had to confront their own mortality in a way that has led them to question, and sometimes reject, many of the assumptions which most people rely on to get through life. Large numbers of their peers and people they loved have died. As time goes on there there are fewer and fewer people with whom they have a shared life experience. War might have made life more intense for a while, but with the perspective of long hindsight there is some bitterness about the damage it has done to their lives. They have a strange sense of not knowing quite where they belong. This describes me pretty well.

18 Having AIDS

Caroline Guinness

I was diagnosed in 1986 when there was very little knowledge of HIV. I had just been diagnosed as having precancer of the cervix, but I felt there was something else wrong – just an instinctive feeling – there was nothing in particular. So I went to my GP, and in fact saw a locum who was very young and enthusiastic. He felt my neck and said my glands were up, which I suppose alerted him to HIV, although he didn't say anything, suggesting it might be glandular fever. He took some blood, and said I should return three days later.

When I went back for the results he said they were negative for glandular fever, but that he had also requested an "AIDS test". I remember feeling really cold when he said that. I knew that maybe that was what it was, because two years beforehand, shortly after my husband left me and I was very vulnerable, I had slept with a bisexual man. I told the doctor that I thought he should have talked to me about it first, and that I wanted the test stopped. He said it was too late as it had already gone to the laboratories. I said in that case I didn't want to know what the result was.

About two weeks later, my own doctor who was back, just turned up at my house. He knew that I didn't want to know the result of the test, but he thought that, as an intelligent woman, I should know that it was positive. Even though I had some suspicions, I found that being told for definite was a different thing altogether. I went into shock. My first reaction was to ask how long I had to live, and he said probably about five years. My next thought was for my daughter, who was three years old at the time, and whether she would be infected too. The doctor didn't think there would be any risk to her as I had obviously contracted it after she was born, but I knew nothing about transmission or anything like that. He suggested another doctor at the practice who had more experience than him, and had been treating a couple of gay men, and that I should go and see her, which I did. She was really sweet, but she didn't know anything about other genitourinary medicine (GUM) clinics, voluntary agencies etc. On the other hand, she was good because she was a very firm believer in complementary therapies, so recommended vitamins and minerals and things which, looking back on it, was actually the best thing she could have done. But not having any counselling and not being in a specialist situation were not good. For the next six months or so I was just in denial – it hadn't sunk in at all. I didn't want to tell anybody because the atmosphere was really bad those days, lots of scaremongering in the Press, calling it the "gay plague" etc.

I did tell a couple of close friends whom I lived with at the time. One, as a gay man, found it very ironic as he thought that if anyone should have tested positive it should have been him, and the other was a girlfriend of mine who sort of panicked. She was OK, but having lost her partner a couple of years before, she couldn't bear the thought of losing somebody else, which of course didn't help me. I didn't want anyone like Social Services to know, as Lee had just started nursery school, and I didn't want it getting out. So I just kept quiet and I continued in my state of denial.

I couldn't cope with work at all – it seemed irrelevant. I told my colleagues that I needed treatment for my cervix which I thought might help explain my lack of concentration. Their reaction was that it wasn't such a big deal, and as I felt I couldn't tell them what was really happening, I resigned. That left me with financial problems, but I didn't want to go to Social Services because of Lee.

I had a partner at the time whom I had been with for six months before being diagnosed, and having to tell him, and him having to get tested was the other thing that was really frightening. Because I didn't know how to tell him, I asked my best friend, whom he got on very well with, if he would tell him. My partner thought that he was going to be told I wanted to split up, so when he realised what it actually was, his initial reaction was one of relief, but the following two weeks, while he got tested and waited for the results, were pretty fraught. We had no information about transmission, but luckily the test came back negative, which was a relief.

My fears about Lee being infected went on for quite some time because I felt I was not getting any real reassurance. I worried about things like her using my toothbrush, and I remembered I had cut my finger and she had helped me put the plaster on, and stuff like that. All those things kept going through my mind. The doctor I was seeing didn't recommend that I had her tested, as she firmly believed Lee would be alright. Looking back on it, I think that if I had just had her tested then I would have felt a lot more reassured, because the whole issue bugged me subconsciously for a long time.

Another very stressful event which happened that year was that a close friend of mine told me he had AIDS. He didn't want anybody to know, and he asked if myself and a couple of other friends could look after him. His health went downhill so quickly and he started getting dementia and incontinence etc, and for me it was like looking in a mirror – very frightening. He did actually go public in the end, but he died shortly before Christmas, so all in all it was quite a bad year.

In 1987, about a year after my diagnosis, through the Terrence Higgins Trust (THT) I finally found out about GUM clinics and I attended James Pringle House, Middlesex Hospital, which made a huge difference to me. I really wanted to meet other HIV positive women – I'd never met any, and still felt as if I was the only woman who had the virus. Someone at THT told me about a support group called Positively Women, who met once a week, so I went along to the group and met a couple of other positive women which helped a lot. I eventually became the Director of Positively Women, and the next three years were really hard work. There was nothing for women at all, so we tried to produce leaflets and information. Despite doing interviews and media work, I never went public about my HIV status. Although our slogan said "For positive women, run by positive women", people never seemed to twig with me; I think they had some vision of what someone with HIV should look like, which I didn't really fit into. Positive women were very much seen as drug users or prostitutes, and most of the women were keeping quiet, usually to protect their families. Through Positively Women, I did many hospital visits to AIDS wards and I used to find that stressful, worrying that I might catch something if I was going to see someone with meningitis or TB.

However, after three years my energy was beginning to dwindle, and I also felt I wasn't spending enough time with Lee, so in 1991, I resigned as the Director, and went part-time. It gave me a bit more time to myself, and because I felt so run down I started using complementary therapies such as acupuncture reflexology, which I'm still having, and which made a difference.

I continued to attend the clinic every three months for a regular follow-up, and the relationship I had with my doctor was very good. She trusted my own judgement on my health, and I found we could work together. She also understood my need for complementary therapies. Seeing the same people on each visit helped maintain continuity and build up a relationship, which was important.

I decided to tell my daughter when she was about 10 years old. She's very bright and reads the newspapers, and it seemed the right time for her. Although I had never gone public about it, I knew it was going to get out at some point, and I didn't want Lee to find out from anyone else. I thought Lee might suspect, but in fact she hadn't. Her first reaction when I told her was to burst into tears, and then she felt embarrassed about crying which made me feel awful as it was quite a natural reaction. For a week or so she kept asking me how I was, and if there was anything in particular that she could do to help. I said she could give me a hand with the housework, but that didn't last very long – I don't think that was what she was expecting! It became immediately apparent that there were no services for children, and she was desperate to meet other kids in the same situation. I suggested to Lee that she didn't tell any friends for a while until she got used to the fact. Anyway she did actually tell a schoolfriend who immediately told everyone else which was exactly what I didn't want to happen.

Her school had been helpful – I had spoken to the Head, her teacher, and the school counsellor before telling her, but she still needed to talk to a trained counsellor, and again she needed to meet other kids in the same situation. None of the organisations offered services for children, but I got a letter from a woman in a similar situation and we met up so that Lee could meet her daughter, which was good for both of them, and at least she knew she wasn't on her own. Lee also started seeing a child psychologist which she really benefited from and she still goes along there when she wants to, but nothing regular.

I think that over the last year or so my energy really hasn't been so good, and as Lee has now reached 13 and is going through everything that 13 year olds do, I could really do with some help now. Her father was in Australia when I was diagnosed, and I didn't want to tell him by 'phone or letter, so I was hoping that he would be coming over to the UK at some point. Because I'd been told that I had about five years to live, I wanted to sort things out as quickly as possible. Anyway, he did come over and I told him, and he had a really odd reaction: he seemed to think I was trying to emotionally blackmail him, which really upset me. It was only later that I found out through a mutual friend that he felt that if he hadn't left me for someone else, I would never have slept with this bisexual man, and so he felt responsible, a thought which had never entered my head. His whole reaction was one of pure guilt, but then over the years that all changed, and for the last few years he's been really supportive.

At the beginning of 1992 I found I was pregnant, and I decided I wanted a termination, and that at the same time I wanted to be sterilised as I didn't want to go through this worry again. I knew enough about transmission at that point to know there was a 10–15% chance that I could pass the virus on, and although that's quite a low risk, I had seen enough other women take the chance and go through the whole nine months and following 18 months not knowing whether the child was infected or not, and I felt I didn't have that in me. I was referred to a hospital and I went there and saw a doctor in outpatients. She knew nothing about HIV – absolutely nothing. She automatically assumed I would want a termination, and before she examined me she removed all the blankets and coverings from the table, so again obviously had no idea about transmission or anything. She also asked the nurse what precautions she should be taking in front of me. It made me feel awful at what was a traumatic time anyway.

About a month ago I was involved in a conference called Living Proof, the first conference for long-term survivors ever, which was really illuminating and quite empowering. There were a lot of other women there which was great to see. I went to three workshops during the two-day event and it was amazing how the experiences of both women and men were so similar. We had all been told first that we probably had five years, then seven years, then 10 years etc. Although my Consultant never said this, it had been the general consensus, and the type of the thing you read in the press, so that when you go past those dates you feel more and more isolated. When you have also suffered so much loss and lost so many people on the way, there is a tinge of guilt that you are still here. Friends whom I told originally have sort of forgotten about it now because it's been going on for so long and they don't seem to realise that I'm still going through it all, and that it takes a large chunk out of my life, that I had to resign my job and go onto benefits.

You feel that people are waiting for you to die. It's still the uncertainty of just not knowing, constantly trying not to be in denial, because there've been enough people I know who have had the virus longer than I have and have died, and I do have definite symptoms. If I was in America I would have been diagnosed as having AIDS a while ago because my CD4 count has been hopping between 150 and 200 over the last two years. Luckily they don't do that here, because psychologically that's a hard one.

When I was admitted into hospital last year, the doctor was trying to be reassuring, saying that it wasn't necessarily HIV-related, but I didn't believe it. I found that most of the nurses had had no specialist training which made me feel a bit vulnerable. One morning I woke to find a young agency training nurse looking at my file; she said, "Oh, you were diagnosed in 1986 and you're still alive – that's amazing", and I thought "I just don't need this, I really don't". I was feeling so ill and didn't really have the strength to deal with it.

When you live in a little closed society like I do medically, where you go into a clinic, where everybody is wonderful and the service is fantastic, you forget about the lack of knowledge and the attitudes outside that world.

Update since 4th edition

It is now the year 2000 and I am still here. I am amazed at just how much my life has changed over the last four years, not to mention the changes in medical advances, from which I have benefitted greatly.

In 1996 I became steadily more and more ill. I was suffering from constant night sweats, my face was covered in molescums, warts appearing everywhere, my hair falling out in huge clumps, my weight plummeting and the most appalling constant fatigue, it was as much as I could do to get a meal together for Lee on her return from school. This was very frightening for her. Finally I came down with pneumonia and having just recovered from that, I immediately got *E. coli* septecaemia and very nearly died. While I was in hospital it was suggested that I start on combination therapy – I felt I had nothing to lose and agreed. The viral load test had just come in and my count was nearly one million, and my T cells were hovering around 50. To my mind I had no choice but to begin treatment.

I started with "Dual Combination" (which is not advisable today). I was taking 3TC and D4T. Almost immediately I started to feel better. My viral load became undetectable, my T cells rose to over 200, my hair stopped falling out, the

molescums disappeared, as did the warts, the night sweats stopped, but the most remarkable thing for me was to suddenly be flooded with masses of energy. I experienced complete euphoria.

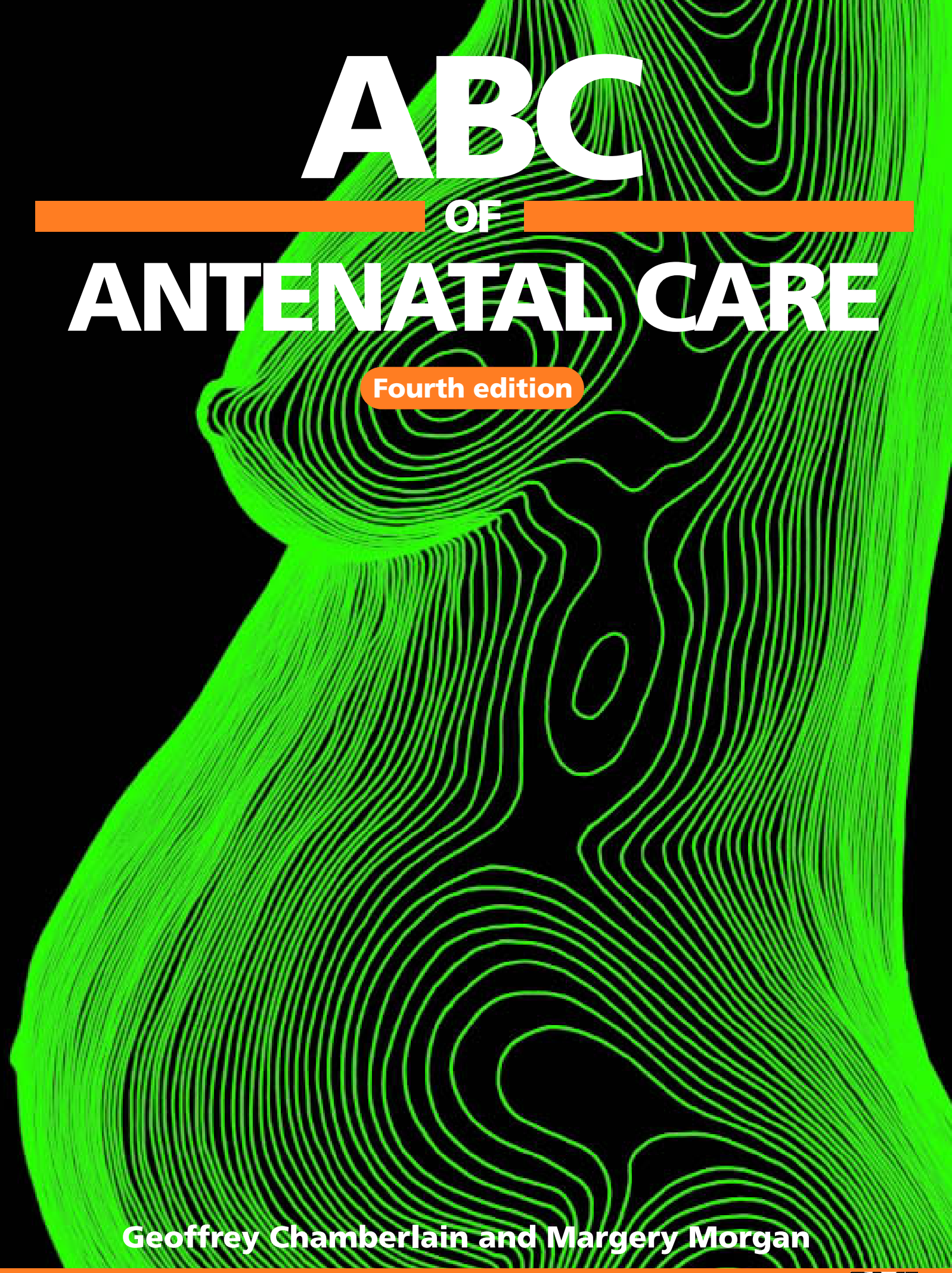
The psychological effects were strange. Having prepared myself for death I found myself strangely afraid of life. I had been forced to opt out of the "rat race", which in a way was rather comforting. Now I needed to join it again. I realised that I had missed years of planning for a future. No pension plans, no savings etc… I also felt a strong sense of "survivor's guilt". I had lost so many close friends and colleagues, and asked myself the question "why me"? Out of the 600 or so women who used Positively Women there were only four of us left from the original group. I decided to start from the beginning again and threw myself into working in the music and film business.

After 18 months I developed a strange side-effect – lipodystrophy. My fat just disappeared off my arms and legs, I had no buttocks to speak of and my face aged about 20 years. My breasts grew enormous and I had a bulging stomach. My viral load was at this point about 5000 and my T cells were dropping again. I decided to see if my body would return to its normal shape and to discontinue any form of treatment. I was still on 3TC and D4T.

After about four months, all my old HIV symptoms were reappearing. My viral load reached 650 000 and my T cells were around 50. The lipodystophy had not changed at all – if anything it was worse – but this could have been "AIDS wasting". Time for a new combination. At this point (September 1998), dual combination was not recommended and I started on triple combination – DMP, AZT and DDI. It took longer for this to work than the last combination, but after about three months my viral load was below 50 and my T cells ranged between 200 and 250. I am still on this combination and still undetectable and, much to my doctor's surprise, my body fat has gone back to normal. My triglycerides are still high, but I have the weight back on my buttocks, arms, legs and face – and it seems to get better by the day.

I have recently married a wonderful man, Lee is now 17 and life goes on. I am more confident of a future, but by no means complacent. I see too many of my peers suffering appalling side effects or complete treatment failure for me to take my life for granted. If I am still around for the next edition of the *ABC of AIDS*, I am sure it will all have changed just as radically again – watch this space.

Index


ABC
OF
ANTENATAL CARE
Fourth edition
Geoffrey Chamberlain and Margery Morgan
BMJ Books

ABC OF ANTENATAL CARE

Fourth edition

GEOFFREY CHAMBERLAIN

Professor Emeritus, Department of Obstetrics and Gynaecology, St George's Hospital Medical School, London and Consultant Obstetrician, Singleton Hospital, Swansea

and

MARGERY MORGAN

Consultant Obstetrician and Gynaecologist, Singleton Hospital, Swansea

BMJ Books

BMJ Books is an imprint of the BMJ Publishing Group

First published in 1992
Second edition 1994
Third edition 1997
Fourth edition 2002
by BMJ Books, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0-7279-1692-0

Cover image depicts body contour map of
a pregnant woman at 36 weeks. With
permission from Dr. Robin Williams/
Science Photo Library.

Typeset by Newgen Imaging Systems Pvt Ltd.

Printed and bound in Spain by GraphyCems, Navarra

Contents

Preface

The chapters in this book appeared originally as articles in the *British Medical Journal* and were welcomed by practitioners. The articles were retuned for publication as a book, the first edition appearing in 1992. Demand asked for more and so the book was updated for a second, a third and now a fourth edition in 2002.

Antenatal care has evolved from a philanthropic service for mothers and their unborn babies to a multiphasic screening programme. Much has been added in the past few years but a lack of scientific scrutiny has meant that little has been taken away. Healthy mothers and fetuses need little high technological care but some screening is desirable to allocate them with confidence to the healthy group of pregnant women. Women and fetuses at high risk need all the scientific help available to ensure the safest environment for delivery and aftercare. The detection and successful management of women and fetuses at high risk is the science of antenatal care; the care of other mothers at lower risk is the art of the subject and probably can proceed without much technology. Midwives are practitioners of normal obstetrics and are taking over much of the care of normal or low-risk pregnancies, backed up by general practitioner obstetricians in the community and by consultant led obstetric teams in hospitals.

This book has evolved from over 40 years of practice, reading, and research. We have tried to unwind the tangled skeins of aetiology and cause and the rational from traditional management, but naturally what remains is an opinion. To broaden this, the authorship has been widened; Dr Margery Morgan, a consultant obstetrician and gynaecologist at Singleton Hospital, has joined Professor Chamberlain as a co-author, bringing with her the new skills used in antenatal care.

We thank our staff at Singleton Hospital for willingly giving good advice and contributing to this book, especially Howard Whitehead, medical photographer, and Judith Biss, ultrasonographer. Our secretaries Caron McColl and Sally Rowland diligently decoded our writings and made the script legible while the staff of BMJ Books, headed by Christina Karaviotis, turned the whole into a fine book.

Geoffrey Chamberlain
Margery Morgan
Singleton Hospital
Swansea

1 Organisation of antenatal care

Looking after pregnant women presents one of the paradoxes of modern medicine. Normal women proceeding through an uneventful pregnancy require little formal medicine. Conversely, those at high risk of damage to their own health or that of their fetus require the use of appropriate scientific technology. Accordingly, there are two classes of women, the larger group requiring support but not much intervention and the other needing the full range of diagnostic and therapeutic measures as in any other branch of medicine. To distinguish between the two is the aim of a well run antenatal service.

Antenatal clinics provide a multiphasic screening service; the earlier women are screened to identify those at high risk of specified problems the sooner appropriate diagnostic tests can be used to assess such women and their fetuses and treatment can be started. As always in medicine, diagnosis must precede treatment, for unless the women who require treatment can be identified specifically, management cannot be correctly applied.

Figure 1.1 New mother and her baby

Background

Some women attend for antenatal care because it is expected of them. They have been brought up to believe that antenatal care is the best way of looking after themselves and their unborn children. This is reinforced in all educational sources from medical textbooks to women's magazines.

Prenatal care started in Edinburgh at the turn of the 20th century, but clinics for the checking of apparently well pregnant women were rare before the first world war. During the 1920s a few midwifery departments of hospitals and interested general practitioners saw women at intervals to check their urine for protein. Some palpated the abdomen, but most pregnant women had only a medical or midwifery consultation once before labour, when they booked. Otherwise, doctors were concerned with antenatal care only "if any of the complications of pregnancy should be noticed". Obstetrics and midwifery were first aid services concerned with labour and its complications: virtually all vigilance, thought, and attention centred on delivery and its mechanical enhancement. Little attention was paid to the antenatal months.

Figure 1.2 Dame Janet Campbell

During the 1920s a wider recognition emerged of the maternal problems of pregnancy as well as those of labour; the medical profession and the then Ministry of Health woke up to realise that events of labour had their precursors in pregnancy. Janet Campbell, one of the most farsighted and clear thinking women in medicine, started a national system of antenatal clinics with a uniform pattern of visits and procedures; her pattern of management can still be recognised today in all the clinics of the Western world.

Campbell's ideas became the clinical obstetric screening service of the 1930s. To it has been added a series of tests, often with more enthusiasm than scientific justification; over the years few investigations have been taken away, merely more added. Catalysed by the National Perinatal Epidemiological Unit in Oxford, various groups of more thoughtful obstetricians have tried to sort out which of the tests are in fact useful in predicting fetal and maternal hazards and which have a low return for effort. When this has been done a rational antenatal service may be developed, but until then we must work with a confused service that "growed like Topsy". It is a mixture of the traditional clinical laying on of hands and a

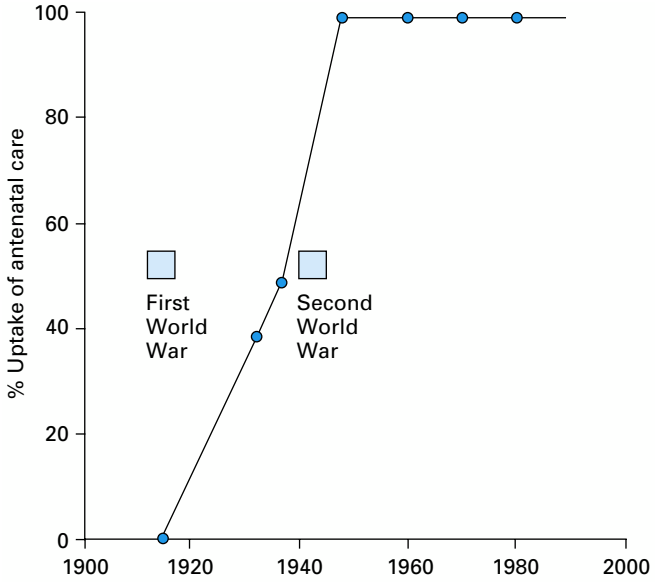


Figure 1.3 Uptake of antenatal care by women in England and Wales

Figure 1.4 Antenatal clinics evolved from child welfare clinics, producing a prenatal version of the infant clinics

Figure 1.5 An antenatal clinic in 2001

patchily applied provision of complex tests, whose availability often depends as much on the whims of a health authority's ideas of financial priority as on the needs of the women and their fetuses.

As well as these economic considerations, doctors planning the care of women in pregnancy should consider the women's own wishes. Too often antenatal clinics in the past have been designated cattle markets; the wishes of women coming for care should be sought and paid attention to. A recurrent problem is the apparent rush of the hospital clinic. The waiting time is a source of harassment and so is the time taken to travel to the clinic. Most women want time and a rapport with the antenatal doctor or midwife to ask questions and have them answered in a fashion they can understand. It is here that the midwives come into their own for they are excellent at the care of women undergoing normal pregnancies.

In many parts of the country midwives run their own clinics in places where women would go as part of daily life. Here, midwives see a group of healthy normal women through pregnancy with one visit only to the hospital antenatal clinic.

To get the best results, women at higher risk need to be screened out at or soon after booking. They will receive intensive care at the hospital consultant's clinic and those at intermediate risk have shared care between the general practitioner and the hospital. The women at lower risk are seen by the midwives at the community clinics. Programmes of this nature now run but depend on laying down protocols for care agreed by all the obstetricians, general practitioners and midwives. Co-operation and agreement between the three groups of carers, with mutual respect and acceptance of each other's roles, are essential.

Janet Campbell started something in 1920. We should not necessarily think that the pattern she derived is fixed forever, and in the new century we may start to get it right for the current generation of women.

Styles of antenatal care

The type of antenatal care that a woman and her general practitioner plan will vary with local arrangements. The important first decision on which antenatal care depends is

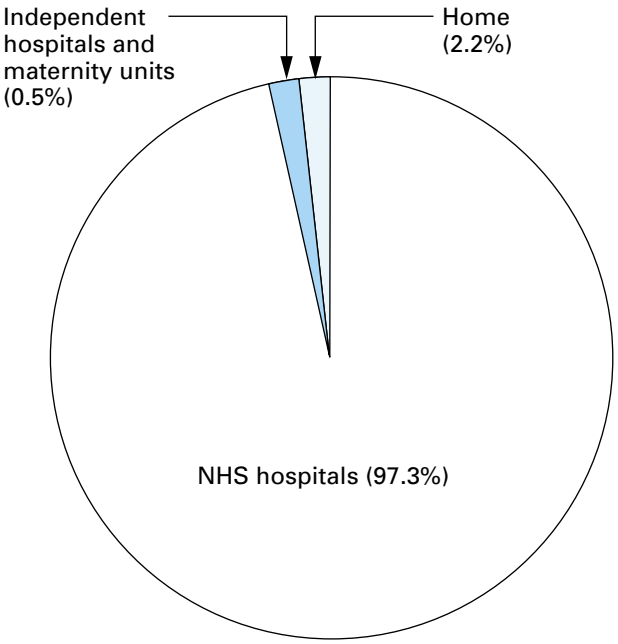


Figure 1.6 Place of birth in England and Wales, 1998

where the baby will be delivered. Ninety seven per cent of babies in the UK are now delivered in institutions, a third of the 2.2% of domiciliary deliveries are unplanned, so about 1.5% are booked as home deliveries. If the delivery is to be in an institution there is still the choice in some areas of general practitioner deliveries either at a separate unit run by general practitioners isolated from the hospital or in a combined unit with a consultant. Most deliveries take place in an NHS hospital under the care of a consultant team. A small but possibly increasing number in the next few years may be delivered in private care, by a general practitioner obstetrician, a consultant obstetrician, or an independent midwife. Recently a series of midwife led delivery units have been established with no residential medical cover.

Once the plans for delivery are decided, the pattern of antenatal visits can be worked out. If general practitioners or midwives are going to look after delivery, antenatal care might be entirely in their hands, with the use of the local obstetric unit for investigations and consultation. At the other end of the spectrum, antenatal care is in the hands of the hospital unit under a consultant obstetrician and a team of doctors and midwives, the general practitioner seeing little of the woman until she has been discharged from hospital after delivery.

Most women, however, elect for antenatal care between these two extremes. They often wish to take a bigger part in their own care. In some antenatal clinics the dipstick test for proteinuria is done by the woman herself. As well as providing some satisfaction, this reduces the load and waiting time at the formal antenatal visit.

During pregnancy there may be visits, at certain agreed stages of gestation, to the hospital antenatal clinic for crucial checks, and for the rest of the time antenatal care is performed in the general practitioner's surgery or midwives' clinic. These patterns of care keep the practitioner involved in the obstetric care of the woman and allow the woman to be seen in slightly more familiar surroundings and more swiftly. In some areas clinics outside the hospital are run by community midwives; these are becoming increasingly popular. Home antenatal care visits also take place, including the initial booking visit.

Delivery may be in the hospital by the consultant led team, by a general practitioner obstetrician, or by a midwife. It is wise, with the introduction of Crown indemnity, that all general practitioner obstetricians have honorary contracts with the hospital obstetric department that they attend to supervise or perform deliveries. About 2% of women now have a home delivery. More than half of these are planned and for this group, antenatal care may well be midwifery led (see *ABC of Labour Care*).

Box 1.1 Fees paid to GPs on the obstetrics list for maternity services April 1997

	£
Complete maternity medical services	186
Antenatal care only from before 16 weeks	100
Confinement	42
Postnatal care only	42

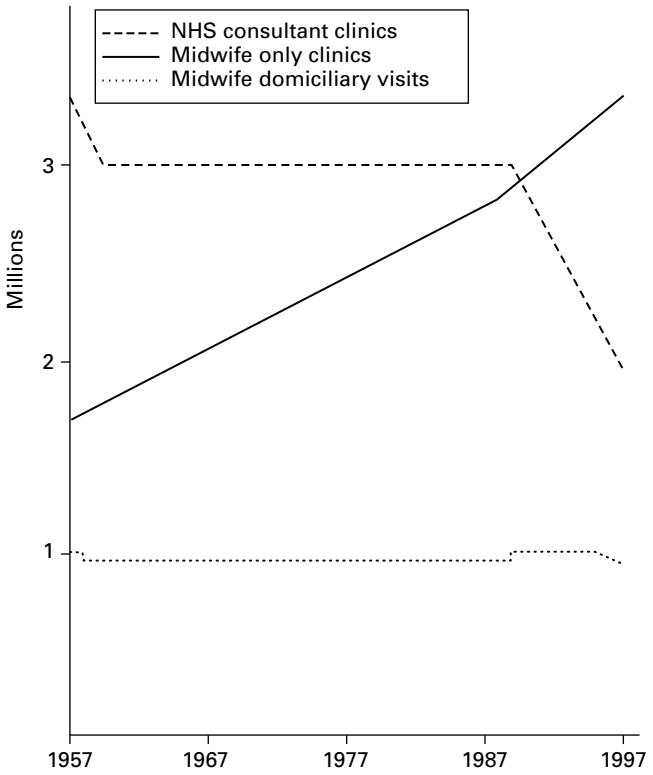


Figure 1.7 Outpatients attendances at antenatal clinics in millions, 1957–97

Early diagnosis of pregnancy

When a woman attends a practitioner thinking that she is pregnant, the most common symptoms are not always amenorrhoea followed by nausea. Many women, particularly the multiparous, have a subtle sensation that they are pregnant a lot earlier than the arrival of the more formal symptoms and signs laid down in textbooks. Traditionally, the doctor may elicit clinical features, but most now turn to a pregnancy test at the first hint of pregnancy.

Symptoms

The symptoms of early pregnancy are nausea, increased sensitivity of the breasts and nipples, increased frequency of micturition, and amenorrhoea.

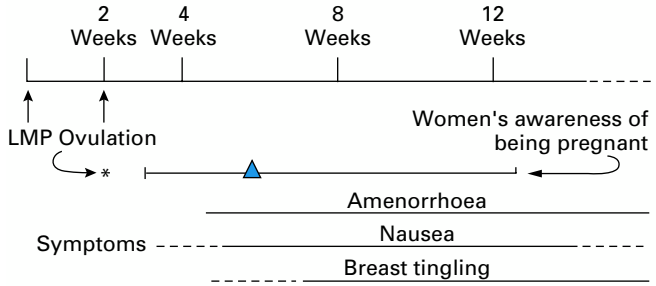


Figure 1.8 Time at which a group of primiparous women first thought that they were pregnant in relation to the more conventional symptoms. The mean (▲) and range are given in weeks of gestation. ---- = extremes.

Signs

The doctor may notice on examination a fullness of the breasts with early changes in pigmentation and Montgomery's tubercuiles in the areola. The uterus will not be felt through the abdominal wall until about 12 weeks of pregnancy. On bimanual assessment uterine enlargement is detectable before this time while cervical softening and a cystic, generally soft feeling of the uterus can be detected by eight weeks. This more subtle sign is not often sought as vaginal examination is not usually performed on a normal woman at this time.

Tests

Mostly the diagnosis of pregnancy is confirmed by tests checking for the higher concentrations of human chorionic gonadotrophin that occur in every pregnancy. The old biological tests using rabbits and frogs are now gone and have been replaced by immunological tests. These depend on the presence of human chorionic gonadotrophin in the body fluids, which is reflected in the urine. The more sensitive the test, the more likely it is to pick up the hormone at lower concentrations—that is, earlier in pregnancy.

Enzyme linked immunosorbent assay (ELISA) is the basis of many of the commercial kits currently available in chemist shops. The assay depends on the double reaction of standard phase antibody with enzyme labelled antibody, which is sensitive enough to detect very low concentrations of human chorionic gonadotrophin. Positive results may be therefore detectable as early as 10 days after fertilisation—that is, four days before the first missed period.

Vaginal ultrasound can detect a sac from five weeks and a fetal cardiac echo a week or so later (Chapter 4), but this would not be used as a screening pregnancy test.

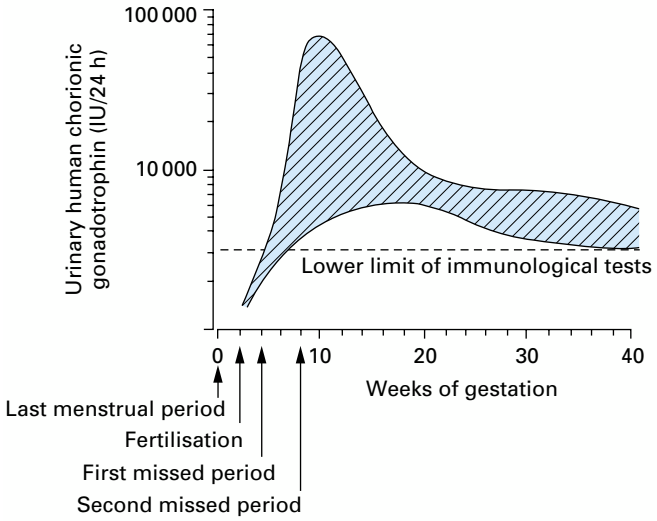


Figure 1.9 Human chorionic gonadotrophin values rise sharply in early gestation but are reduced in the second half of pregnancy. The normal range ±2 SD is shown

Figure 1.10 Clearview pregnancy test results. The horizontal bar in the top chamber shows that a urine sample has progressed satisfactorily from the lower chamber. A horizontal bar in the middle chamber shows a positive result (right) and its absence a negative result (left)

Conclusion

At the end of the preliminary consultation women may ask questions about the pregnancy and the practitioner will deal with these. Most of these queries will be considered in the chapter on normal antenatal management. For most women the onset of pregnancy is a desired and happy event, but for a few it may not be so and practitioners, having established a diagnosis, may find that they are then asked to advise on termination of pregnancy. This they should do if their views on the subject allow; if not, they should arrange for one of their partners to discuss it with the patient. Most women, however, will be happy to be pregnant and looking forward to a successful outcome.

Recommended reading

- Cnattingius V. *Scientific basis of antenatal care*. Cambridge: Cambridge University Press, 1993.
- Cole S, McIlwaine G. The use of risk factors in predicting consequences of changing patterns of care in pregnancy. In Chamberlain G, Patel N, eds. *The future of the maternity services*. London: RCOG Press, 1994.
- Collington V. *Antenatal care*. London: South Bank University, 1998.

Antenatal care has evolved from a hospital based service to a community based service for normal women. Those with a higher risk of problems are best seen in hospital clinics.

The picture of the infant welfare clinic is reproduced by permission of William Heinemann from *University College Hospital and its Medical School: a History* by W R Merrington. The Clearview pregnancy test result is reproduced by permission of Unipath, Bedford.

2 The changing body in pregnancy

Pregnancy is a load causing alterations not just in the mother's pelvic organs but all over the body. Fetal physiology is different from that of an adult, but it interacts with the mother's systems, causing adaptation and change of function in her body. These adaptations generally move to minimise the stresses imposed and to provide the best environment for the growing fetus; they are usually interlinked smoothly so that the effects on the function of the whole organism are minimised.

Pregnancy causes physiological and psychological changes, which affect all aspects of the woman's life.

Cardiovascular system

The increased load on the heart in pregnancy is due to greater needs for oxygen in the tissues.

- The fetal body and organs grow rapidly and its tissues have an even higher oxygen consumption per unit volume than the mother's.
- The hypertrophy of many maternal tissues, not just the breasts and uterus, increases oxygen requirements.
- The mother's muscular work is increased to move her increased size and that of the fetus.

Cardiac output is the product of stroke volume and heart rate. It is increased in pregnancy by a rise in pulse rate with a small increase in stroke volume. Cardiac muscle hypertrophy occurs so that the heart chambers enlarge and output increases by 40%; this occurs rapidly in the first half of pregnancy and steadies off in the second. In the second stage of labour, cardiac output is further increased, with uterine contractions increasing output by a further 30% at the height of the mother's pushing.

During pregnancy the heart is enlarged and pushed up by the growing mass under the diaphragm. The aorta is unfolded and so the heart is rotated upwards and outwards. This produces electrocardiographic and radiographic changes which, although normal for pregnancy, may be interpreted as abnormal if a cardiologist or radiologist is not told of the pregnancy.

Blood pressure may be reduced in mid-pregnancy, but pulse pressure is increased and peripheral resistance generally decreases during late pregnancy.

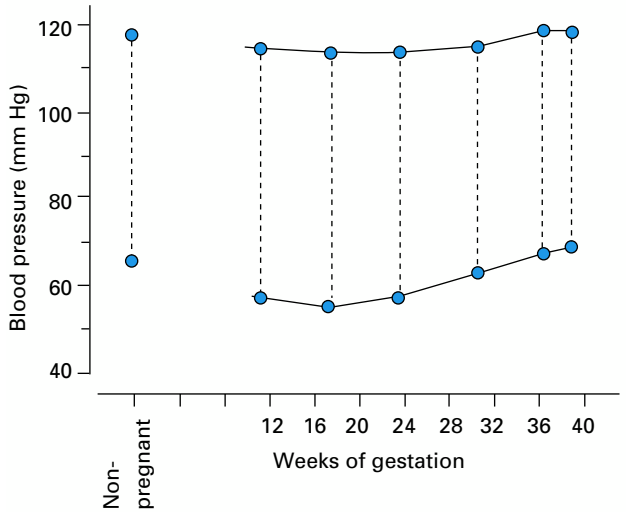


Figure 2.3 Systolic and diastolic blood pressures during pregnancy. The mid-trimester dip found in some women is seen more in the diastolic than in the systolic pressure

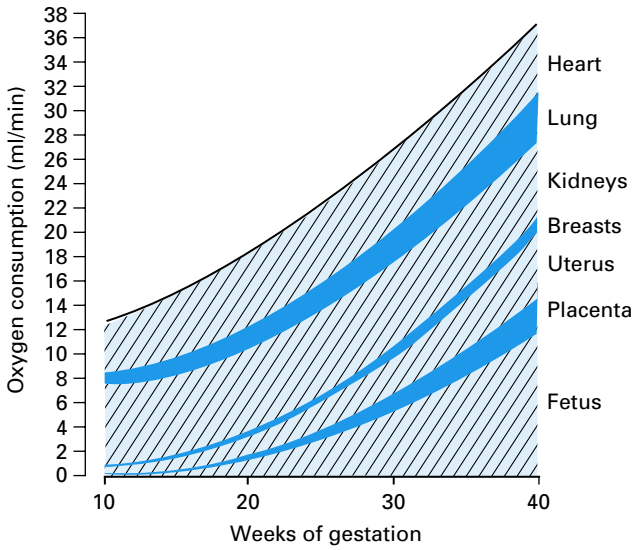


Figure 2.1 Increase in oxygen consumption during pregnancy. A major part of the increase goes to the products of conception (fetus and placenta)

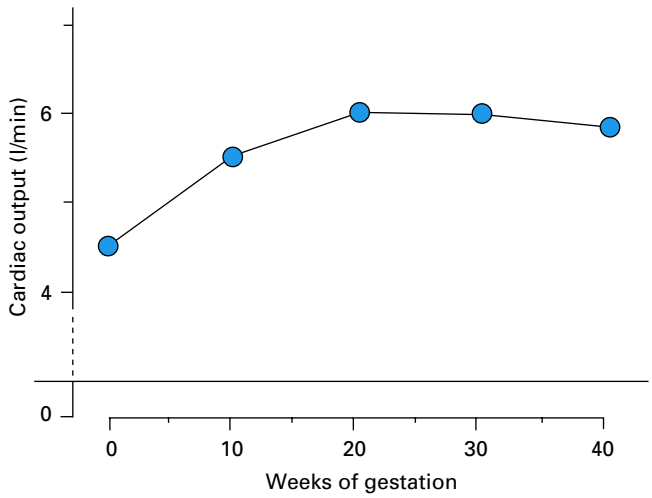


Figure 2.2 Cardiac output in pregnancy. The increase occurs very early and flattens from 20 weeks

Box 2.1 Changes in the ECG in normal pregnancy

- Deep Q waves in I and II
- T wave flattened or inverted in III
- ST segment depressed
- Extra-systolies frequent

Maternal blood volume increases, the changes in plasma volume being proportionally greater than the increase in red cell bulk. Hence haemodilution occurs; this used to be called a physiological anaemia, a bad phrase as it is paradoxical to have a physiological pathological process.

The heart sounds are changed.

- A systolic ejection murmur is common.
- A third cardiac sound is commonly heard accompanying ventricular filling.

The electrical activity of the heart on an electrocardiogram changes.

- The ventricles become hypertrophied, the left to a greater extent than the right and therefore left ventricular preponderance is seen in the QRS deviation.

Heart valves and chamber volumes may change during pregnancy. The heart becomes more horizontal so cardiothoracic ratio is increased and it has a straighter upper left border. These changes can be visualised by cross-sectional echocardiography, which depends on the reflection of high frequency sound from inside the heart.

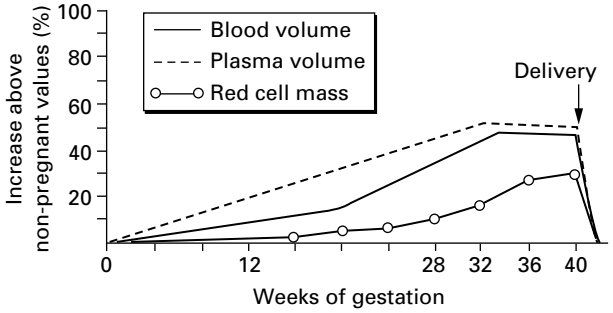


Figure 2.4 Increase in blood volume and its components in pregnancy

Respiratory system

Box 2.2 Changes in chest radiographs in normal pregnancy

Lungs

- Show increased vascular soft tissue
- Often have a small pleural effusion especially straight after delivery

The most common changes seen on chest *x* ray films are shown in the box. Always ensure that the radiology department is told on the request form that a woman is pregnant and give an approximate stage of gestation. Only when there are strong indications should chest radiography be performed in pregnancy at all and then full radiological shielding of the abdomen must be used.

In early pregnancy women breathe more deeply but not more frequently under the influence of progesterone. Hence alveolar ventilation is increased by as much as a half above prepregnant values so that pO_2 levels rise and carbon dioxide is relatively washed out of the body.

Later the growing uterus increases intra-abdominal pressure so that the diaphragm is pushed up and the lower ribs flare out. Expiratory reserve volume is decreased but the vital capacity is maintained by a slight increase in inspiratory capacity because of an enlarged tidal volume. This may lead to a temporary sensation of breathlessness. Explanation usually reassures the woman.

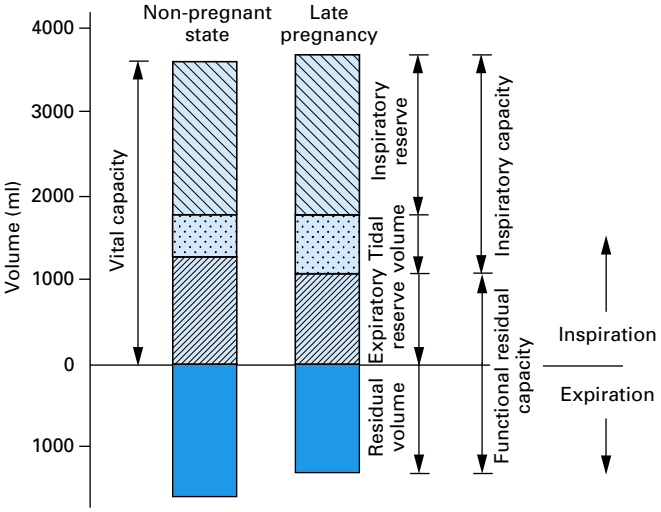


Figure 2.5 Changes in inspiratory and expiratory volumes in pregnancy

Urinary system

Changes in clearance

Renal blood flow is increased during early pregnancy by 40%. The increase in glomerular filtration rate is accompanied by enhanced tubular reabsorption; plasma concentrations of urea and creatinine decrease.

The muscle of the bladder is relaxed because of increased circulating progesterone. Increased frequency of micturition due to increased urine production is a feature of early pregnancy. Later the bladder is mechanically pressed on by the

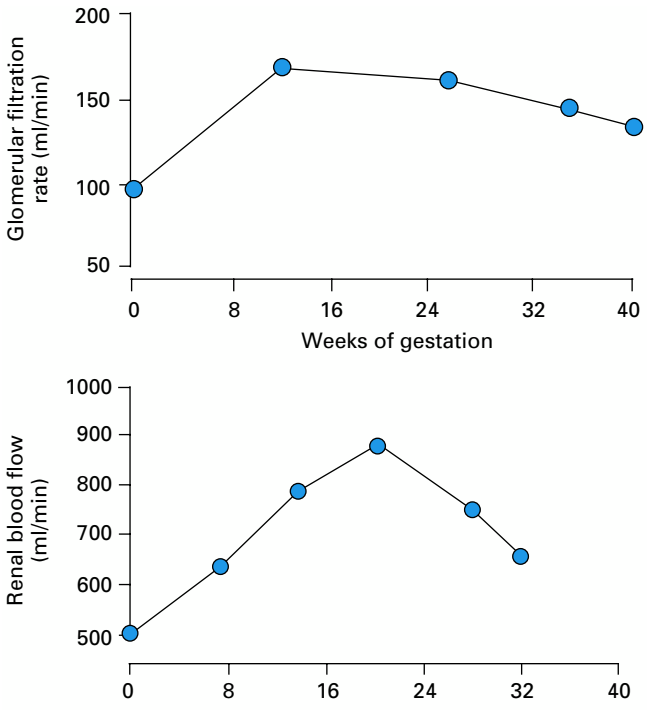


Figure 2.6 Changes in the glomerular filtration rate and in renal blood flow in pregnancy

growing uterus and the same symptoms occur but for a different reason.

The muscle walls of the ureters are relaxed by progesterone so that the ureters become larger, wider, and of lower tone. Sometimes stasis occurs in the ureters; therefore proliferation of bacteria and the development of urinary infection is more likely to occur.

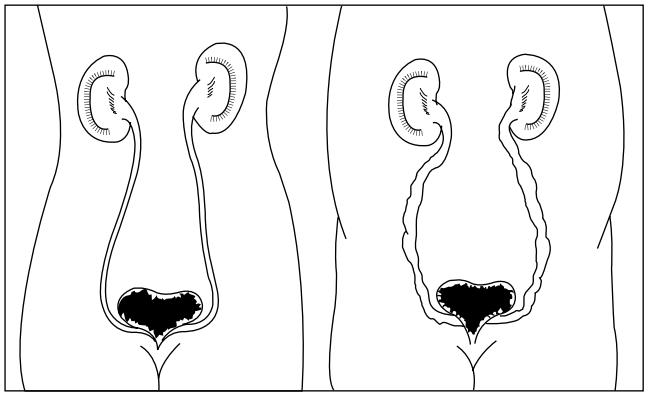


Figure 2.7 Changes in the ureters in pregnancy, during which they lengthen and become more tortuous and dilated

Endocrine system

All the maternal endocrine organs are altered in pregnancy, largely because of the increased secretion of trophic hormones from the pituitary gland and the placenta.

Pituitary gland

The pituitary gland is increased in size during pregnancy, mostly because of changes in the anterior lobe.

Anterior lobe

- *Prolactin*. Within a few days of conception the rate of prolactin production increases. Concentrations rise until term following the direct stimulation of the lactotrophs by oestrogens. Human placental lactogen, which shows shared biological activity, exerts an inhibitory feedback effect. Prolactin affects water transfer across the placenta and therefore fetal electrolyte and water balance. It is later concerned with the production of milk, both initiating and maintaining milk secretion.
- *Gonadotrophins*. The secretions of both follicular stimulating hormone and luteinising hormone are inhibited during pregnancy.
- *Growth hormone*. The secretion of growth hormone is inhibited during pregnancy, probably by human placental lactogen. Metabolism in the acidophil cells returns to normal within a few weeks after delivery and is unaffected by lactation.
- *Adrenocorticotrophic hormone* concentration increases slightly in pregnancy despite the rise in cortisol concentrations. The normal feedback mechanism seems to be inhibited secondary to a rise in binding globulin concentrations.
- *Thyrotrophin* secretion seems to be the same as that in non-pregnant women. The main changes in thyroid activity in pregnancy come from non-pituitary influences.

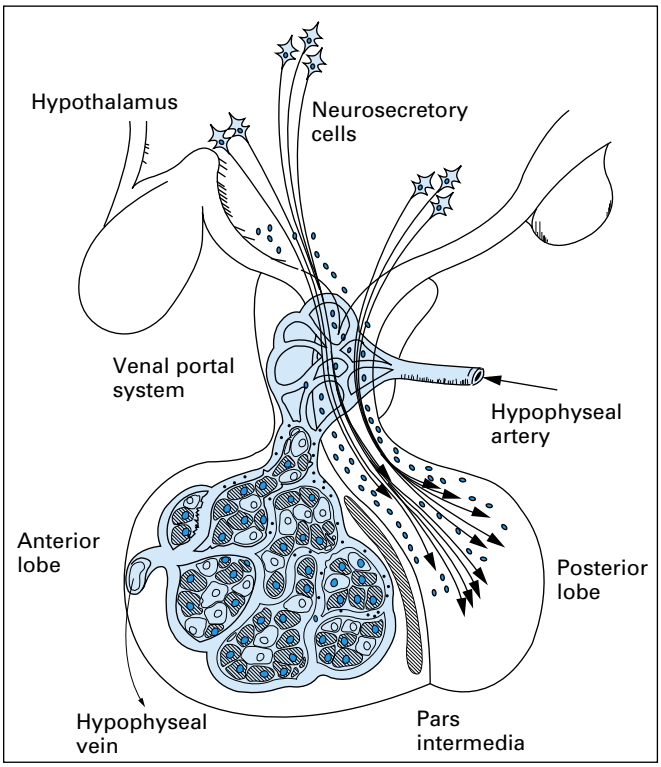


Figure 2.8 Pituitary gland showing secreting areas

Posterior lobe

There are increases in the release of hormones from the posterior pituitary gland at various times during pregnancy and lactation. These, however, are produced in the hypothalamus, carried to the pituitary gland in the portal venous system, and stored there. The most important is oxytocin, which is released in pulses from the pituitary gland during labour to stimulate uterine contractions. Its secretion may also be stimulated by stretching of the lower genital tract. Oxytocin is also released during suckling and is an important part of the let down reflex.

Thyroid gland

Pregnancy is a hyperdynamic state and so the clinical features of hyperthyroidism may sometimes be seen. The basal metabolic rate is raised and the concentrations of thyroid hormone in the blood are increased, but thyroid function is essentially normal in pregnancy.

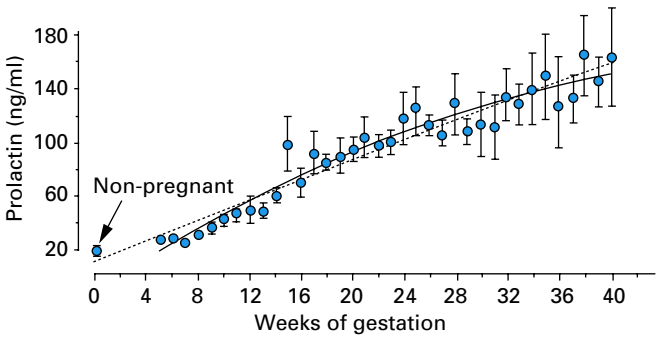


Figure 2.9 Changes in prolactin concentrations in pregnancy (means and SEMs)

In pregnancy the renal clearance of iodine is greatly increased but thyroid clearance also rises so absolute iodine levels remain in the normal range. The raised hCG levels are associated with a reduced (inside the normal range) TSH; hCG probably stimulates the gland in early pregnancy and is capable of stimulating TSH receptors.

Adrenal gland

The adrenal cortex synthesises cortisol from cholesterol. In pregnancy there is an increase in adrenocorticotrophic hormone concentration along with an increase in total plasma cortisol concentration because of raised binding globulin concentrations. The cortex also secretes an increased amount of renin, possibly because of the increased oestrogen concentrations. This enzyme produces angiotensin I, which is associated with maintaining blood pressure. Some renin also comes from the uterus and the chorion, which together produce a large increase in renin concentrations in the first 12 weeks of gestation. There is little change in deoxycorticosterone concentrations despite the swings in electrolyte balance in pregnancy.

The adrenal medulla secretes adrenaline and noradrenaline. The metabolism seems to be the same during pregnancy as before; the concentrations of both hormones rise in labour.

Placenta

The oestrogen, progesterone, and cortisol endocrine functions of the placenta are well known. In addition, many other hormones are produced with functions related to maternal adaptation to the changes of fetal growth.

In some susceptible women, progesterones may soften critical ligaments so that joints are less well protected and may separate (e.g. separation of the pubic bones at the symphysis).

Genital tract

The uterus changes in pregnancy; the increase in bulk is due mainly to hypertrophy of the myometrial cells, which do not increase much in number but grow much larger. Oestrogens stimulate growth, and the stretching caused by the growing fetus and the volume of liquor provides an added stimulus to hypertrophy.

The blood supply through the uterine and ovarian arteries is greatly increased so that at term 1.0–1.5 l of blood are perfused every minute. The placental site has a preferential blood supply, about 85% of the total uterine blood flow going to the placental bed.

The cervix, which is made mostly of connective tissue, becomes softer after the effect of oestrogen on the ground substance of connective tissue encourages an accumulation of water. The ligaments supporting the uterus are similarly stretched and thickened.

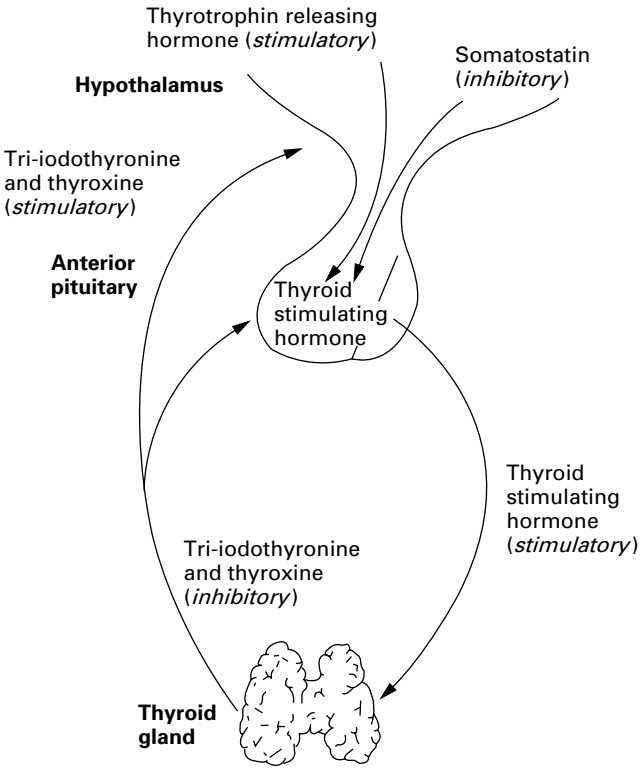


Figure 2.10 Control of thyroxine secretion in pregnancy

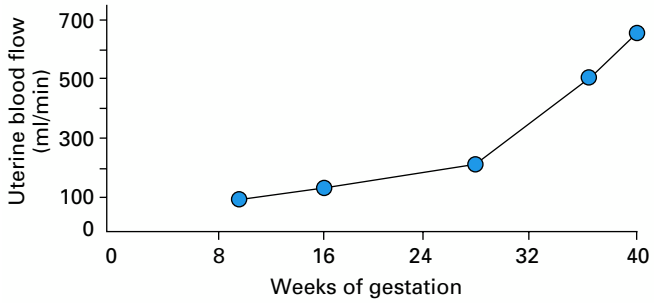


Figure 2.11 Changes in uterine blood flow in pregnancy

The wide range of normal physiological changes of gestation must be allowed for when making clinical diagnoses about diseases in pregnancy.

Recommended reading

- Chamberlain G, Broughton-Pipkin F, eds. *Clinical physiology in obstetrics*. 3rd edn. Oxford: Blackwell Scientific Publications, 1998.
- de Sweit M, Chamberlain G, Bennett M. *Basic science in obstetrics and gynaecology*. 3rd edn. London: Harper and Bruce, 2001.

The figure showing the control of thyroid secretion is reproduced by permission of Blackwell Scientific Publications from *Clinical Physiology in Obstetrics* edited by F Hytten and G Chamberlain. The figure showing prolactin secretion during pregnancy is reproduced by permission of the *American Journal of Obstetrics and Gynecology* (Rigg LA, Lein A, Yen SCC, 1977;129:454–6).

3 Normal antenatal management

Antenatal care has six functions (see Box 3.1). The first two are the same as any performed in an outpatient clinic (treatment of symptoms); the second two relate to multiphasic screening, of which antenatal care was an early example; the third pair are part of health education.

Box 3.1 Aims of antenatal care

- Management of maternal symptomatic problems
- Management of fetal symptomatic problems
- Screening for and prevention of fetal problems
- Screening for and prevention of maternal problems
- Preparation of the couple for childbirth
- Preparation of the couple for childrearing

Antenatal care in the UK is performed by a range of professionals: midwives, general practitioners, and hospital doctors. In many areas up to 90% of antenatal care is in the hands of general practitioners and community midwives. In many parts of the country midwives hold their own clinics outside the hospital or visit women at home. Probably those initially at lower risk do not need routine specialist visits for they offer little or no benefit. Many women now carry their own notes, which leads to greater understanding of what is going on.

In the UK many women book for antenatal care by 14 weeks and are seen at intervals. There is no association between the number of visits and outcome; in Switzerland there are an average of five and in The Netherlands as many as 14, but outcomes are the same. The number of visits depends on a traditional pattern laid down by Dame Janet Campbell in the 1920s (Chapter 1) rather than on being planned with thoughts relating to the contemporary scene. In an ideal world, the follow-up antenatal visits would be planned individually according to the needs of the woman and assessment of her risk.

A more rational plan of care of normal primigravidas and multigravidas is laid down in Table 3.1. With these criteria, antenatal care would be more cost effective and no less clinically useful. When pioneers have tried to reduce the number of visits from the traditional number, however, there has been resistance from older obstetricians, conventional midwives, women having babies, and their mothers, all of whom think that Campbell's by now traditional pattern must be right. A randomised controlled trial in south-east London actually found women in the fewer visits group were more likely to be dissatisfied although outcomes of the groups were the same.

As well as the clinical regimen, antenatal care now entails a whole series of special tests, but these are not generally used for the normal pregnant population.

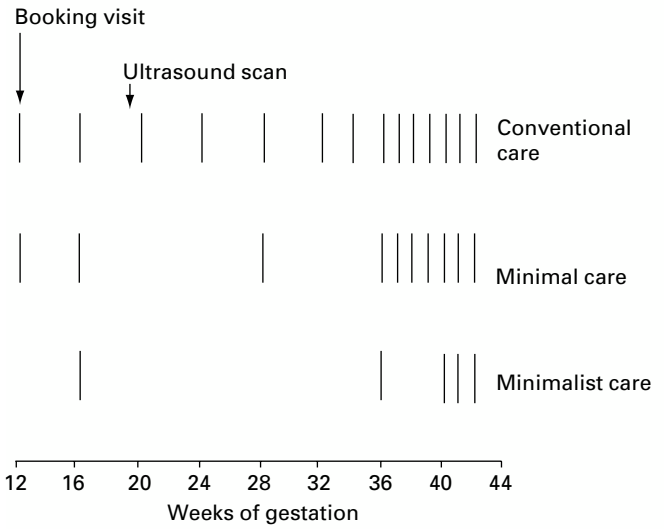


Figure 3.1 Intervals of antenatal visits: conventional pattern (top); current ideas of low risk care (middle); plan for the least number of visits (bottom)

Table 3.1 Care for normal multi- and primigravidas

Week of gestation	Main purpose of visit*
Minimum care for normal multigravidas	
12	History and examination, clarification of uncertain gestation, identification of risk factors for antenatal care and confinement, booking blood tests, booking scan in some units
	Advice on diet, drugs, work, and exercise
15–20	Downs serum screening, α Fetoprotein, anomaly ultrasound scan
22	Fundal height, baseline weight
30	Fundal height, weight gain, identification of intrauterine growth restriction and pre-eclampsia
36	Fundal height, weight gain, identification of malpresentation
40	Assessment if need for induction
Additional visits for normal primigravidas	
26	Blood pressure, urine analysis, discussion of delivery and infant feeding
34	Blood pressure, urine analysis, discussion of delivery and infant feeding
38	Blood pressure, urine analysis, discussion of delivery and infant feeding
41	Blood pressure, urine analysis, discussion of delivery and infant feeding

* Blood pressure reading and urine analysis are performed at every visit.

Prepregnancy care

Some aspects of a couple's way of life may be checked before pregnancy. The man and the woman's medical and social history, and, if relevant, her obstetric career can be assessed. Immunity from infections such as rubella can be tested; alternative treatments to some longstanding conditions such as ulcerative colitis can be discussed. The possibility of a recurrence of pre-existing problems such as deep vein thrombosis can be assessed. Dietary habits and problems at work can be assessed and changes in consumption of cigarettes or alcohol may be considered. Once pregnancy has started the

couple have only two options—that is, to continue or stop the pregnancy. Prepregnancy care allows more time for the correction of detectable problems and the prevention of their repetition—for example, giving supplementary folate to women whose children have abnormalities of the central nervous system. It is now recommended that extra folate is started by all women before pregnancy to avoid deficiency in very early pregnancy when the fetal neural tube is closing (21–28 days of fetal life) so as to reduce the risk of spina bifida.

Box 3.2 Aims of prepregnancy care

- To bring the woman to pregnancy in the best possible health
- To attend to preventable factors before pregnancy starts—for example, rubella inoculation
- To discuss diabetes and aim for excellent glycaemic control
- To assess epileptic medication in terms of fit control and teratogenicity
- To discuss antenatal diagnoses and management of abnormality
- To give advice about the effects of:
  - pre-existing disease on the pregnancy and unborn child
  - the pregnancy on pre-existing disease and its management
- To consider the effects of recurrence of events from previous pregnancies
- To discuss the use of prophylactic folate before conception

Booking visit

Once pregnancy has been diagnosed, the woman usually books a visit at the antenatal clinic, the GP surgery or at home with the midwife who will lead in antenatal care. This is the longest but most important visit. It used to take place at 8–12 weeks' gestation, but in many clinics it has moved to 12–14 weeks. The woman's medical state is assessed so that the current pregnancy can be placed into the appropriate part of a risk spectrum. Baseline data are essential at this point and are obtained from the history, an examination, and relevant investigations.

History

Symptoms that have arisen in the current pregnancy before the booking visit are ascertained—for example, vaginal bleeding and low abdominal pain.

- *Menstrual history.* To assess the expected date of delivery details are needed about the last normal menstrual period including its date, the degree of certainty of that date, and whether cycles are reasonably regular around 28 days. The use of oral contraception or ovulation induction agents that might inhibit or stimulate ovulation should be discussed. A firm date for delivery from the last menstrual period can be obtained from about 80% of women.

  From this calculate the expected date of delivery with a calculator. Do not do sums in the head; this can cause trouble when a pregnancy runs over the end of a year. A woman can be told that she has an 85% chance of delivering within a week of the expected date of delivery, but we must emphasise at this point that this date is only a mathematical probability and, as with other odds, the favourite does not always win the race. Most units now rely on ultrasound to confirm gestation and alter the EDD if the scan date varies considerably, i.e. more than 10 days difference.
- *Medical history.* Specific illnesses and operations of the past should be inquired about, particularly those that entail treatment that needs to be continued in pregnancy—for example, epilepsy and diabetes.
- *Family history.* There may be conditions among first degree relatives (parents or siblings) that may be reflected in the current pregnancy, such as diabetes or twinning.
- *Sociobiological background.* Age, parity, social class, and race of the woman all affect the outcome of the pregnancy. Smoking and alcohol consumption also affect the outcome. Socioeconomic class is usually derived from the occupation of the woman or her partner. It reflects the influence of a mixed group of factors such as nutrition in early life, diseases in childhood, education, and past medical care. It also correlates with potential birth weight, congenital abnormality rates, and eventually perinatal mortality. Less strongly associated are preterm labour and problems in care of the newborn.
- *Obstetric history.* The woman's obstetric history should be discussed carefully as it contains some of the best markers for

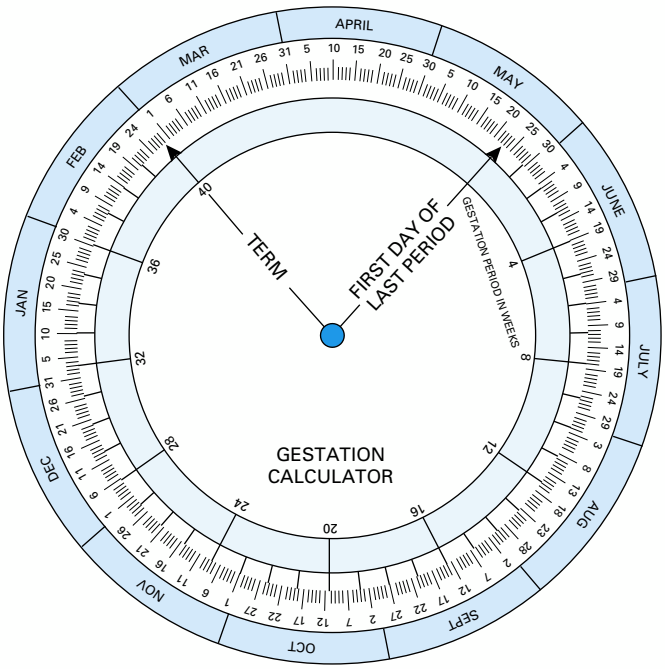


Figure 3.2 An adjustable obstetric calculator should always be used to calculate the current stage of gestation and the expected date of delivery

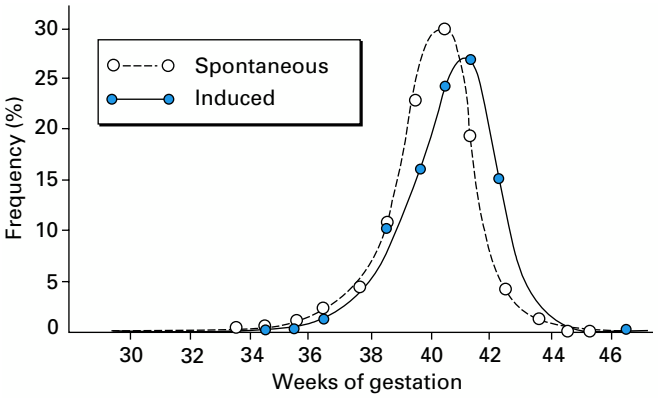


Figure 3.3 Distribution of length of gestation for spontaneous and induced single births when the last menstrual period is known (n=16 000)

performance in the current pregnancy. If the woman has had a previous miscarriage or termination of pregnancy, the doctor should ask about the stage of gestation, and any illness afterwards. Of babies born, the progress of the pregnancy, labour, and puerperium are needed and the stage of gestation and birth weight of the infant. Intrauterine growth restriction and preterm labour may be recurrent and should be inquired about in previous pregnancies. The terms gravidity and parity are often applied to women in pregnancy. Gravidity refers to pregnancy, so anyone who is gravid is or has been pregnant. A woman who is pregnant for the first time is a primigravida. Parity refers to having given birth to a viable liveborn or a stillborn child.

Table 3.2 Proportions of live births in each socioeconomic class in England and Wales adjusted by job description of husband or partner (1998)

Social class	Job description	Population having babies (%)
I + II	Professional & supervisory	25.6
IIINM	Skilled, non-manual	6.7
IIIM	Skilled, manual	16.6
IV + V	Semiskilled & unskilled	10.3
Not classifiable*		3.0
Not married		37.8

* In many surveys unemployed people are classified by their last occupation if they had one.

Examination

A brief but relevant physical examination should be performed. The woman's height is important as it correlates loosely with pelvic size, but shoe size is a poorer predictor. Weight is less often monitored in pregnancy these days, but a booking weight will enable a Body Mass Index (BMI) to be measured. This is the weight in kilograms divided by the height squared (weight (kg)/height (cm^2)). BMI is useful in determining those at increased risk during pregnancy (over 30) who require consultant obstetric care. A value of over 39 (morbidly obese) may indicate that an anaesthetic assessment is necessary to assess potential problems in labour at delivery. The clinical presence of anaemia should be checked and a brief examination of the teeth included, if only to warn the woman to visit a dentist. Tooth and gum deterioration may be rapid in pregnancy and dental care is free at this time and for a year after delivery.

Check whether the thyroid gland is enlarged. The blood pressure is taken, preferably with the woman resting for a few minutes before. The spine should be checked for any tender areas as well as for longer term kyphosis and scoliosis, which might have affected pelvic development; the legs should be examined for oedema and varicose veins.

The abdomen is inspected for scars of previous operations—look carefully for laparoscopy scars below the umbilicus and for a Pfannenstiel incision above the pubis. Palpation is performed for masses other than the uterus—for example, fibroids and ovarian cysts. If the booking visit is before 12 weeks the uterus probably will not be felt on abdominal examination, but in a multiparous woman it may be; this should not cause the examiner to make any unnecessary comments about an enlarged uterus at this stage.

A vaginal assessment was traditionally performed at the booking visit. Its function was to confirm the soft enlargement of the uterus in pregnancy, to try to assess the stage of gestation, to exclude other pelvic masses, and to assess the bony pelvis. Many obstetricians now do not do a pelvic assessment at this stage; no woman likes having a vaginal examination and, if done in early pregnancy, it is associated in the woman's mind with any spontaneous miscarriage which may occur subsequently, even though this is irrelevant to the examination. Fetal size will soon be checked by ultrasound. Even assessment of the bony pelvis in late pregnancy may not be required as the fetal presenting part is available for check against the inlet while the effect of progesterone on the pelvic ligaments is at its maximum. By this time the woman has more confidence in the antenatal staff and is more willing to have a vaginal examination. If the head is engaged, this is a good measure of pelvis size. If it is not, a vaginal assessment may still be needed.

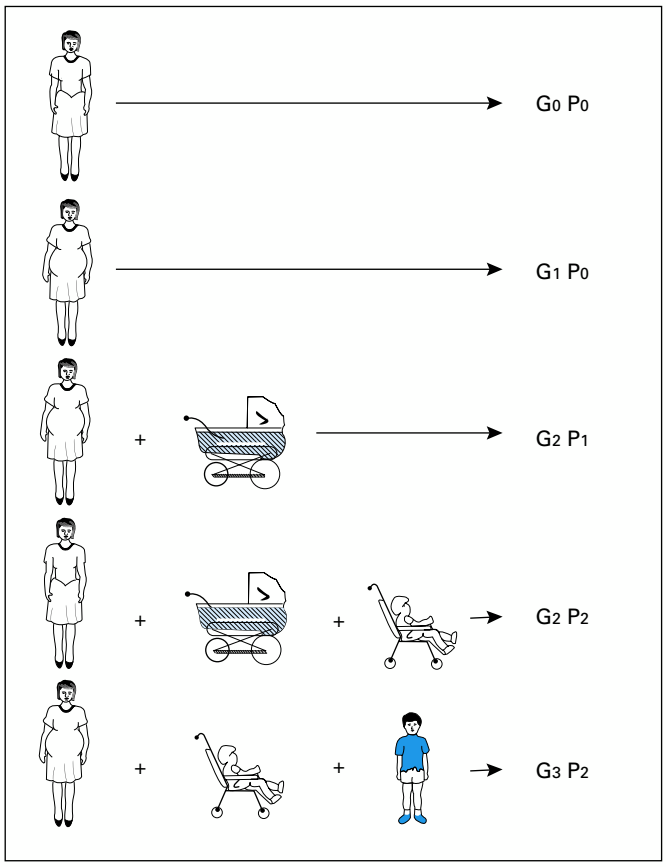


Figure 3.4 Gravidity (G) and parity (P)

Investigations

A venous blood sample is checked for:

- Haemoglobin concentration or mean cell volume (see Chapter 8).
- ABO and rhesus groups and, if relevant, rhesus antibodies. The former is to allow swifter cross-matching of blood if needed in pregnancy or labour; the latter is to warn of problems and be a baseline if a rhesus-positive fetus is in the uterus of a rhesus-negative woman.
- Antibodies to other blood groups—for example, Kell, to give warning of potential incompatibility with the fetus in the presence of less common blood groups.
- Haemoglobinopathies in women originating from Mediterranean, African, and West Indian countries.
- Syphilis. A Wassermann reaction (WR) is non-specific; most clinics now use the *Treponema pallidum* haemagglutination test to investigate more specifically, but no test can be expected to differentiate syphilis from yaws or other treponematoses.
- Rubella antibodies.
- HIV antibodies. If the woman is at risk of infection through intravenous drug misuse, having received contaminated blood transfusions, coming from parts of the world with a high HIV rate (e.g. sub-Saharan Africa), or having a partner who is HIV positive, she may request or be advised to have an HIV test. Full counselling should include her understanding the implications of both having the test and any positive result. In some parts of the UK, antenatal testing is offered to all with a modified advice service beforehand. The mother can opt out.
- Hepatitis B antibodies.
- Toxoplasmosis antibodies (if clinically appropriate).
- Cytomegalic virus antibodies (if clinically appropriate).

Later blood checks are for:

- α Fetoprotein level analysis for abnormality of the central nervous system.
- Down's syndrome serum screening by double or triple test.

The urine is checked for:

- Protein, glucose and bacteria.

Chest radiographs are rarely taken except in women from parts of the world where pulmonary tuberculosis is still endemic.

An ultrasound assessment is now performed on most pregnant women in the UK. It is best done at about

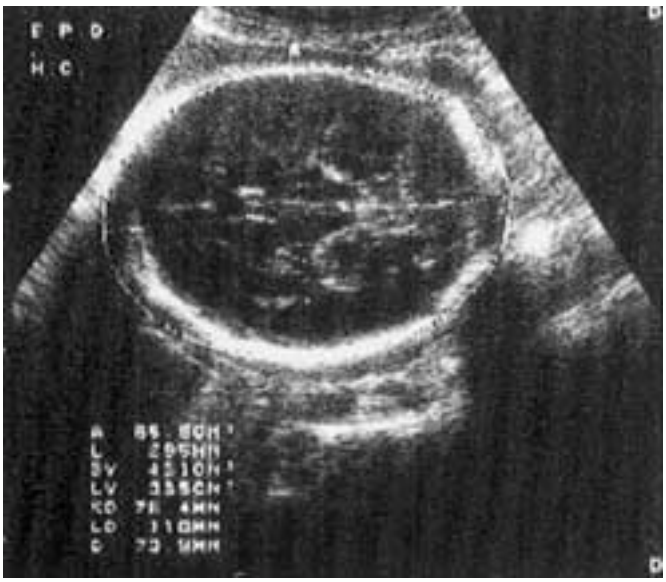


Figure 3.8 Ultrasound of fetal head showing the midline echo, the biparietal diameter of the head circumference outline

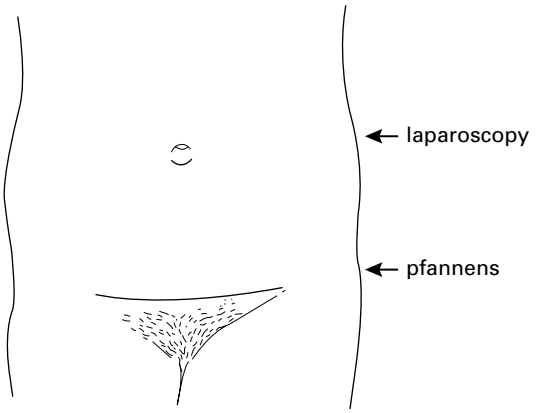


Figure 3.5 Laparoscopy and Pfannenstiel scars

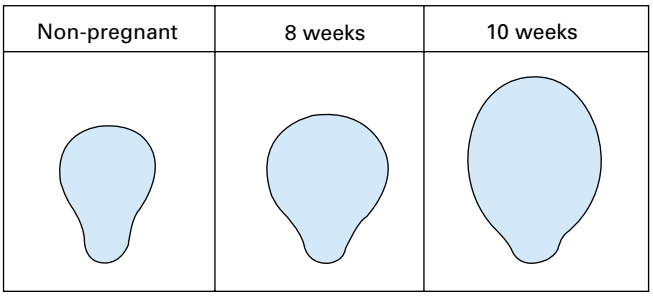


Figure 3.6 Relative growth of uterus in early pregnancy. Growth is usually in width rather than length, so the uterus seems fuller at first. It is also softer and has a cystic quality

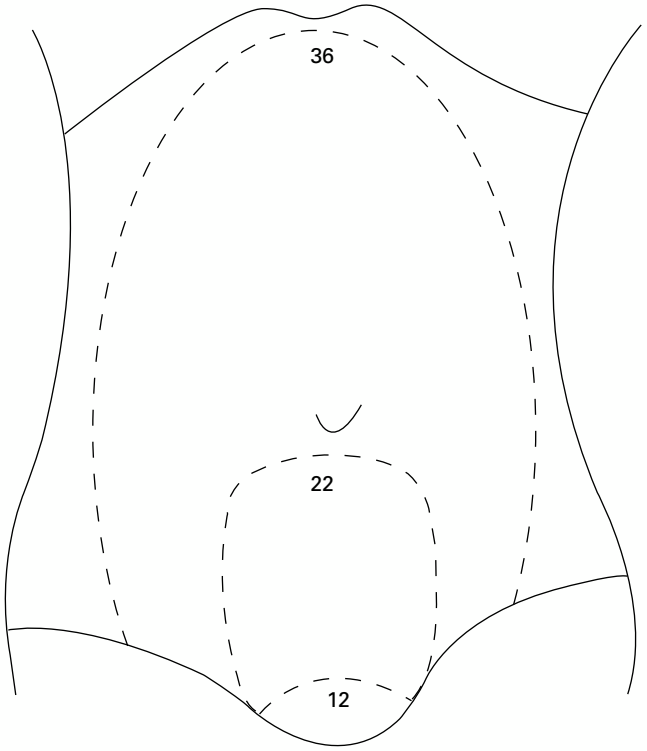


Figure 3.7 Size of uterus at various stages of gestation in pregnancy

18–20 weeks to measure the biparietal diameter and so get a baseline value of fetal size and confirmation of the stage of gestation to firm up the expected date of delivery. Gross congenital abnormalities may be found (Chapter 4).

Ultrasound between 10 and 13 weeks can measure nuchal translucency, which is being evaluated as a screening test for Down's syndrome (Chapter 4). At 18 weeks congenital abnormalities such as spina bifida, omphalocele, and abnormal kidneys may be excluded. A four chamber view of the heart is also possible at this stage to exclude gross abnormalities, but details of cardiac connections may not be obvious until 22–24 weeks. Other conditions which are characterised by decreased growth such as microcephaly or some forms of dwarfism may also not be apparent until late in the second trimester.

Hence, though 16–18 weeks would be a useful time to assess gestational age by ultrasound, much later assessments are needed to assess fetal normality. In addition, more highly skilled ultrasonographers and equipment of high resolution are needed to produce scans to enable assessment of normality. Many of these ultrasound studies of fetal anatomy have been developed in specialist units with highly skilled obstetric ultrasonographers. The ordinary ultrasound service at a district general hospital cannot be expected always to provide such skill or equipment, although with increased training and better machines, some centres are now providing a fuller exclusion service at 20–24 weeks' gestation. Also at 24 weeks Doppler flow

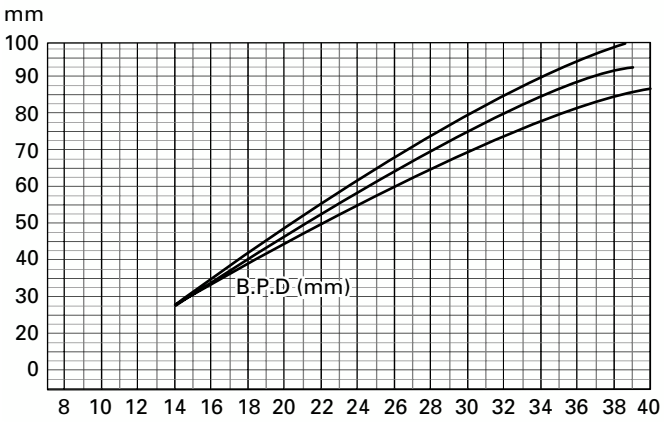


Figure 3.9 Mean (±2 SD) biparietal diameter of the fetal head in a normal population. Note the narrow range of normal values in earlier pregnancy, a great difference from that of biochemical test results

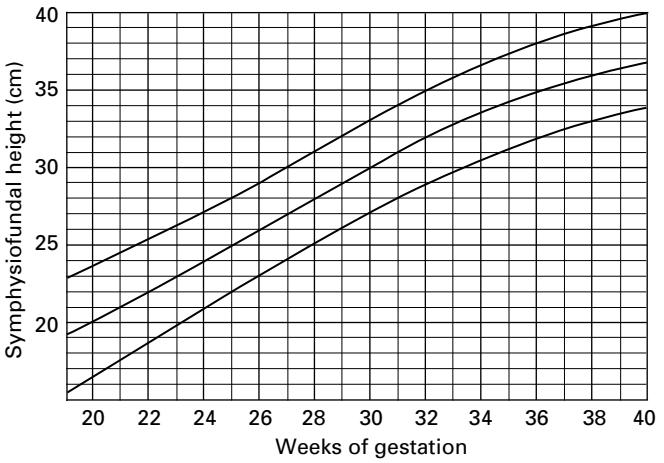


Figure 3.10 Mean (± SD) of symphysio-fundal height by weeks of gestation. Note the wide range of readings for any given week of gestation and the even wider range of expected gestation weeks for any given reading

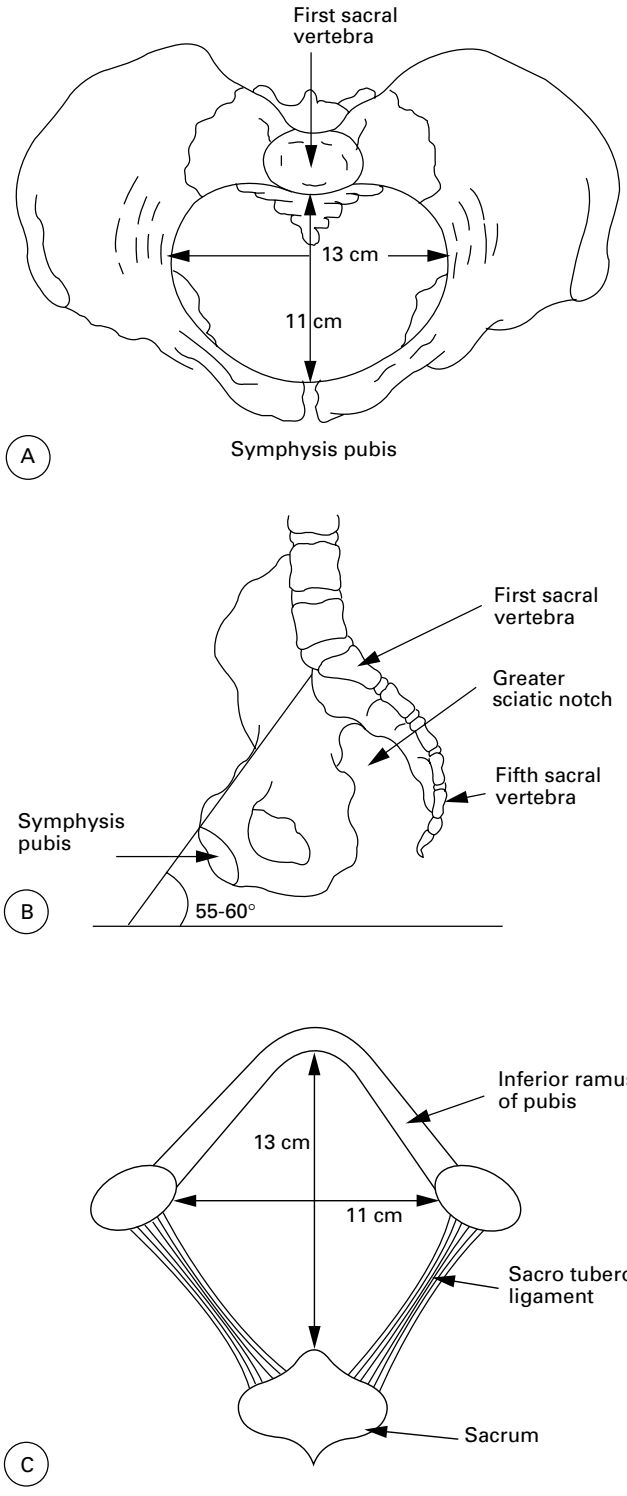


Figure 3.11 Outline of the normal bony pelvis. (A) Inlet seen from above. (B) Side view showing angle of inclination of the pelvic nm. (C) Outlet seen from below

studies may identify those mothers at risk of later hypertension or fetuses for growth restriction (Chapter 4).

Subsequent antenatal visits

At each antenatal visit an informal history is sought of events that have happened since the last attendance. The woman's blood pressure is assessed and compared with the previous readings; proteinuria and glycosuria are excluded each time. Palpation of the abdomen and measurements of the fundus above the symphysis give a clinical guide to the rate of growth of the fetus, especially if they are performed at each visit by the same observer. In later weeks the lie and presentation of the fetus is assessed. In the last weeks of pregnancy the presenting part, usually the head, is checked against the pelvic inlet to ensure that it engages. If the fetal head is not engaged by 37 weeks it is helpful to see if it will engage. To do this, the top of the couch should be propped up to 60° from the horizontal and the lower abdomen re-examined. If this small change in entry angle allows engagement of the fetal head, it will usually go down when labour contractions start. This is a simple test giving useful information about the potential of the fetal head to negotiate the mother's pelvis; it deserves wider usage in antenatal clinics.

The amount of amniotic fluid is assessed clinically and if fetal movements are seen by the observer or reported by the mother, the fetal heart need not be auscultated at the antenatal clinic. If, however, the mother reports reduced movements, the heart should be checked with a hand held Doppler fetal heart monitor and by cardiotocography so that the woman, too, can observe the heart beats and be reassured.

In a visit in the last few weeks of pregnancy a pelvic examination may be performed to check the bony pelvis, the points of importance being shown in Box 3.3. A well engaged fetal head after 36 weeks indicates, however, that the pelvis is adequate in this pregnancy and that digital assessment need not be performed. With a persistently non-engaged head or a breech presentation it should be done. Assessment of the cervix is wise at 32 weeks if the woman is at high risk of a preterm labour or is having a twin pregnancy, although it can be done in many units by vaginal ultrasound. It is also useful to assess cervical ripeness if the pregnancy is postmature after 42 weeks.

Malpositions

By 37 weeks, most fetuses will have settled into a cephalic presentation, but about 3% will still be a breech or transverse lie. Many obstetricians would offer an external cephalic version (ECV). The earlier ECV is done, the easier it is to turn the fetus but the more likely it is to turn back. Most versions are offered from 36 weeks onwards.

Before the version takes place the fetal heart is recorded for about 20 minutes and the lie checked with ultrasound. The fetal breech is then carefully disimpacted from the mother's pelvis. When above the brim, it is grasped in one hand and the head is swung round with the other hand in a series of moves so that the head is pointing downwards.

The fetal heart is checked on a cardiotocograph immediately after the version for about 20 minutes. Success rates vary between 10% and 50%.

End of pregnancy

Traditionally in Britain many obstetricians have been concerned when a singleton pregnancy goes past 42 weeks. In the 1960s the actuarial risk of perinatal mortality did sharply increase after 41 weeks, but this is no longer so and the passage

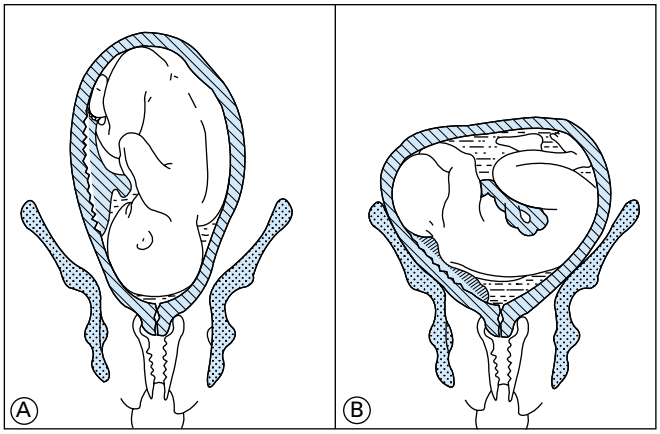


Figure 3.12 Lie of the fetus. (A) Longitudinal lie, which is deliverable vaginally. (B) Transverse lie, which if it persists has to be delivered abdominally

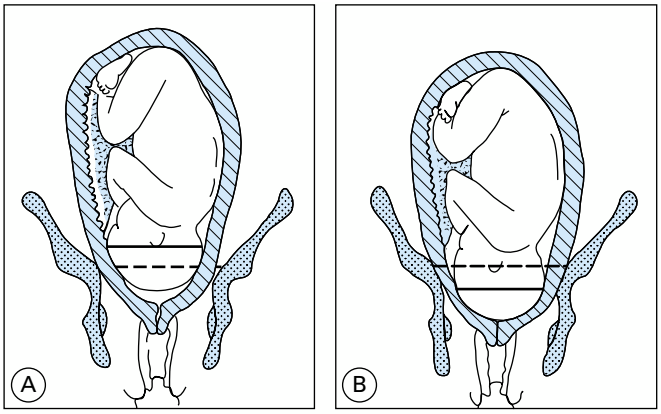


Figure 3.13 (A) The fetal head is not engaged as its maximum diameter (——) is above the inlet of the mother's pelvis (------). (B) The fetal head has descended so that its maximum diameter is below the inlet

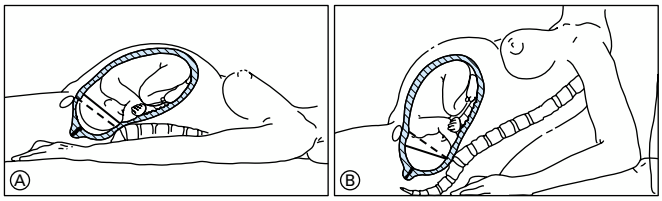


Figure 3.14 (A) The fetal head is not engaged, but when the mother sits up (B) gravity allows the head to sink below the inlet of the mother's pelvis so that the head will engage

Box 3.3 Clinical assessment of bony pelvis should include checking the:

- anteroposterior diameter from symphysis pubis to promontory of the sacrum (S1)
- curve of the sacrum
- prominence of the ischial spines
- angle of the greater sciatic notch
- width of the inferior border of the symphysis pubis
- subpubic angle

of 42 proved weeks is not used by all obstetricians as an indication for induction of labour. For example, if the cervix is not ripe some would consider it unwise to induce merely on calendar dates. Instead, the unusually long length of gestation might be used as an indication for better and more frequent fetal surveillance with Doppler and CTG rather than to take action, but this should be done at the consultant clinic in the hospital rather than in the community. The results of fetal monitoring after 42 weeks should be assessed carefully for the normal reduction of amniotic fluid volume can lead to false conclusions.

Antenatal education

Pregnancy counselling

The visits to an antenatal clinic can be a helpful time for the woman and her partner to learn about pregnancy. Formal antenatal education classes are held in most district hospitals, and couples are encouraged to attend a convenient course of counselling. Furthermore, informal discussions with midwives and doctors at the antenatal clinic are educational and much can be learnt from other mothers in the waiting time at the clinics. This is complemented by many excellent videos, which are often displayed in the antenatal waiting area.

Many good books exist about pregnancy and childbirth, offering a spectrum of styles and detail according to a woman's needs. A woman should be steered towards a well written account of what she needs in a form that best suits her lifestyle and religious observances in a language that she can understand. Plenty of such books are now available, but all hospital and obstetricians should read the material that is offered to the women who visit their clinics to make sure that they agree with and actually offer the services that the books advocate, e.g. it is no good the literature being about epidural pain relief in labour if the hospital at which the woman is booked cannot provide it.

Pregnancy social support

In the welfare state of the UK pregnant women are entitled to several social security benefits, although in many ways this country lags behind many countries in the European Union. The doctors at the clinic would do well to keep up their knowledge from time to time as benefits change rapidly according to the whims of the Department of Social Security and of their political masters. The Maternity Alliance frequently produces excellent pamphlets on these matters to help

Box 3.4 Problems with antenatal ECV

- The fetus may be too big.
- Extended legs may splint the fetus.
- The cord may be wound around the neck or limbs and so anchor the fetus.
- The abdominal muscles may be too tense to allow a grip of the fetal pole.
- Obesity may limit the grip of the fetal pole.
- The uterine muscle may contract and so resist manipulation. Try a uterine relaxant.
- Excess of amniotic fluid will allow reversion to breech presentation.
- A uterine abnormality (e.g. septum or fibroids) may not allow ECV.
- The membranes may rupture.

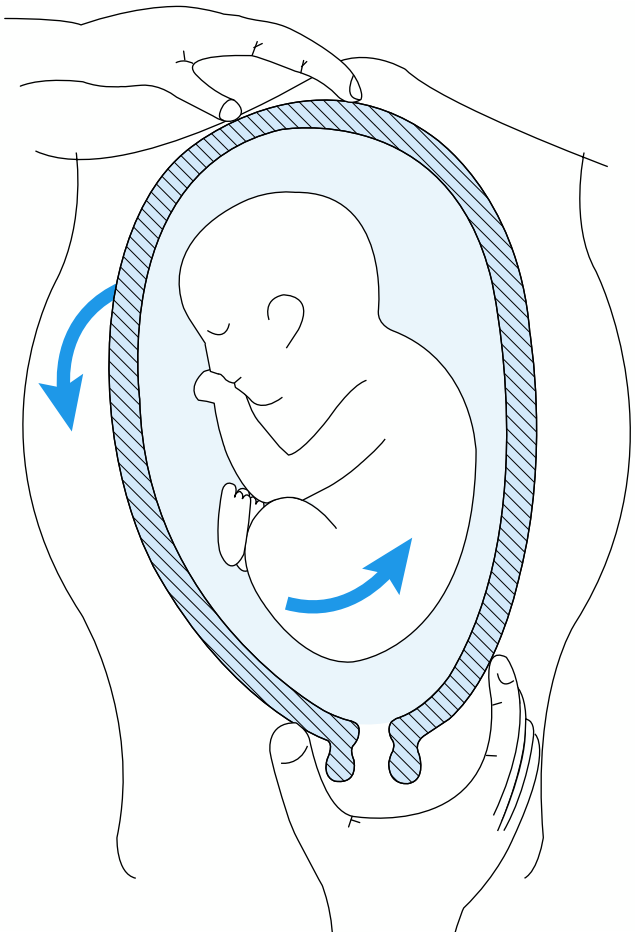


Figure 3.15 External cephalic version is usually performed by disimpacting the breech from the pelvis and then swinging the fetus through 180°

Figure 3.16 Antenatal instruction includes relaxation classes with a physiotherapist

Figure 3.17 A wide variety of antenatal information books is available

both women and professionals keep up to date (Maternity Alliance, 45 Beech St, London EC2P 2LX).

Conclusion

The antenatal visit in the community, general practice surgery, or hospital should be friendly and held at a time when women can mix with others who are also pregnant and so informally discuss their problems. It also provides a nidus for antenatal counselling both formally at the antenatal classes and informally from staff and other women. The medical component is the core of the clinic and consists of the regular screening and assessment of symptomatic problems to bring the woman and her fetus to labour in the best state at the best time.

Antenatal care is now the cornerstone of obstetrics. Though the problems of labour are more dramatic, some of them could be avoided by effective detection and management of antenatal variations from the normal.

Recommended reading

- Fiscella K. Does prenatal care improve birth outcome? *Obstet. Gynec.* 1995;**85**:468–79.
- Hall M. Antenatal care. In Chamberlain G, ed. *Turnbull's obstetrics.* 3rd edn. London: Harper and Bruce, 2001.
- RCOG. Routine Ultrasound Screening in Pregnancy. London: RCOG, 2000.

4 Checking for fetal wellbeing

The great reduction in maternal mortality and morbidity in the past 30 years has allowed more attention to be concentrated on the fetus during antenatal care. Perinatal mortality has been reduced, but still in England and Wales out of 100 babies born, one will die around the time of birth, two have an abnormality, and six have a birth weight under 2500 g. With smaller family sizes in the Western world, parents expect a perfect result. General practitioners and obstetricians are performing more thorough checks to try to detect the fetuses that are likely to be at increased risk. These investigations do not replace clinical examination but provide the fine tuning of assessment. The mother still needs, however, to see someone who can talk to her and discuss the implications and results of these new tests with her.

Some groups of women are at high risk because of their medicosocial background. The extremes of maternal age (under 16 and over 35), high parity (over four pregnancies), low socioeconomic class (Office for National Statistics, social class V), and some racial groups (Pakistan-born women) seem to confer a higher actuarial risk on the babies born to such women. Consequently these women deserve extra antenatal surveillance to detect a fetus with variations from normal. Others show poor growth of the fetus in the latter days of pregnancy or develop raised blood pressure during pregnancy, two manifestations of a poor blood flow to the placental bed. Such fetuses have poor nutritional reserve—a decreased blood flow to the placental bed reduces the amounts of nutrients and oxygen. A series of tests have been developed; some of these are screening tests best applied to the total antenatal population or to a subset considered to be at higher risk. Other tests are diagnostic and specifically used for women with babies thought to be compromised clinically. All these investigations can be done in a day care unit and do not necessitate admission.

Table 4.1 Perinatal mortality in England and Wales in 1995–96 according to various maternal factors

Maternal factor	Rate per 1000 total births
Age (years)	
<20	8·5
20–24	7·1
25–29	6·6
30–34	6·7
35+	8·6
Parity	
0	6·6
1	6·4
2	8·6
⩾3	15·0
Socioeconomic class	
I	5·8
II	6·0
IIIN	6·6
IIIM	7·1
IV	8·9
V	10·6
Place of mother's birth	
UK	8·2
Republic of Ireland	9·8
India	9·5
East Africa	12·4
West Indies	11·5
Pakistan	15·8

Tests in early pregnancy (up to 13 weeks)

Ultrasound

The earliest in pregnancy that the embryo may be visualised by abdominal ultrasonography is six to seven weeks; it will be shown a week earlier with a vaginal probe. At six weeks the embryonic sac can be seen but embryonic tissue cannot be confidently visualised, even with machines of high resolution and skilled ultrasonographers. By seven to eight weeks most ultrasound machines should be able to show the embryo and a fetal heart pulse can often be seen. Most obstetric departments are moving to the use of vaginal probes in early pregnancy because of the better resolution of the image. Nuchal translucency measurements are dealt with in Chapter 5.

Hormone tests

Tests are currently being developed that may be helpful in very early pregnancy to detect women who are likely to miscarry. They mostly measure proteins derived from the placenta, for example, human chorionic gonadotrophin and Schwangerschaftsprotein 1. Oestrogen and progesterone tests are too non-specific to be of prognostic value so early in gestation.

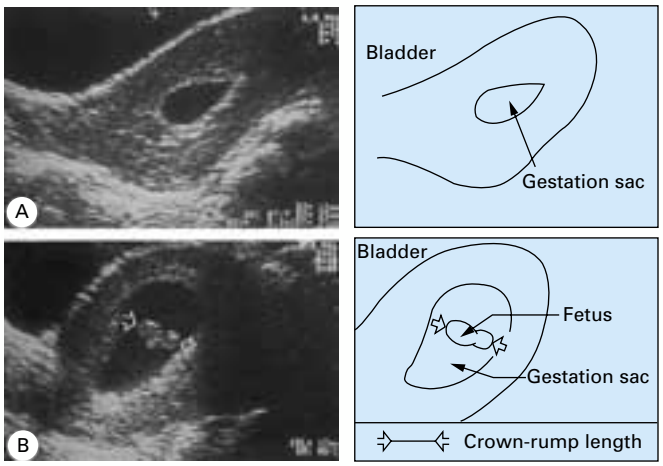


Figure 4.1 (A) The embryonic sac can be seen at six weeks gestation in decidua. As yet no fetal parts can be identified. (B) The same sac two weeks later. Fetal parts can easily be seen between the arrows. The pulsation of the fetal heart may also be seen at this time

Chorionic villus sampling
This is at present mainly used to detect chromosomal abnormalities and is considered in the next chapter.

Isoimmunisation
Maternal immune reactions may be stimulated by ante- or intrapartum fetomaternal bleeding whenever any fetal blood group factor inherited from the father is not possessed by the mother. The emphasis used to be on the Rhesus factor risk but this is rapidly being overcome by preventative anti-D gamma globulin injections given after any potential fetomaternal bleed (delivery, external cephalic version, termination of pregnancy). ABO and other blood groups become relatively more important now and antibodies for these should be screened. Management depends upon an early diagnosis of the blood groups of the mother and the presence of any antibodies. If these are detected at booking, repeat tests of antibodies should be made at intervals until the middle of pregnancy. If the antibody titre is rising the mother should be referred to a special centre capable of dealing fully with isoimmunisation.

If the rise is gradual so that the effect of the maternal antibodies passing back across the placenta is minimal to the fetus, then one might await events or stimulate an early delivery. If the position is worse, then intrauterine exchange transfusions are required. Now these are nearly always done (through a fetoscope) directly into the fetal umbilical vessels. The intraperitoneal transfusions have mostly been abandoned in the Western world. Perinatal survival rates are now reported at over 80% in even severely isoimmunised fetuses but one must remember there are complications of the invasive processes themselves. The procedure related mortality of intravascular transfusion is between 4 and 9%. The value of percutaneous transuterine umbilical artery transfusion should be compared with early delivery and performing extrauterine intravascular exchange transfusions in each centre.

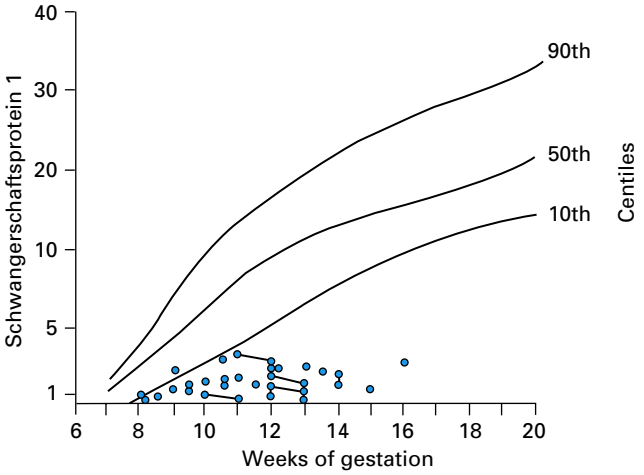


Figure 4.2 Maternal serum concentrations of Schwangerschaftsprotein 1 in pregnancies with no ultrasonic evidence of fetal heart action. This protein is made by the fetus and placenta; concentrations increase steadily through pregnancy. Many fetuses who abort spontaneously have concentrations below the 10th centile in the first weeks of gestation

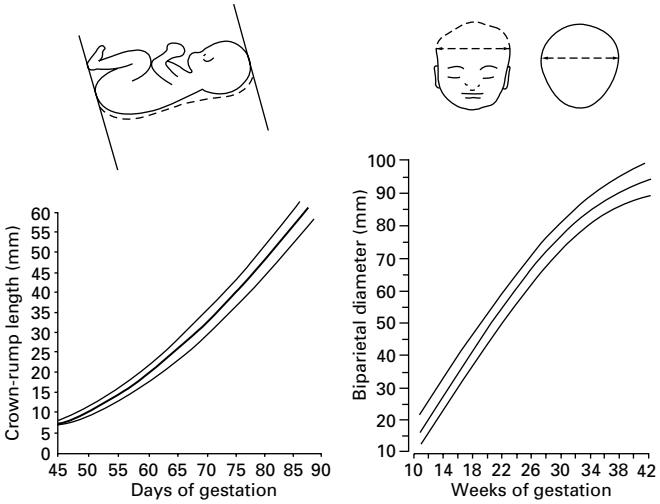


Figure 4.3 Crown-rump length by days of gestation and biparietal diameter by weeks of gestation show a narrow range inside ±2 SD of the mean, indicating a good test

Tests in mid-pregnancy (14–28 weeks)

Ultrasound
Ultrasound has become a more sophisticated tool in the past 40 years, so that by 20 weeks of pregnancy the fetus can be visualised precisely. Two separate sets of measurements are taken of the fetus to assess growth and detect malformations. The detection of malformations is the subject of the next chapter.

Growth may be determined by assessment of a series of measurements of the individual fetus at different times in pregnancy. These may then be compared with a background population to see whether the fetus is growing at the same rate as a statistically comparable group of its peers. Obviously the growth chart should relate to a population from which the fetus comes and not be taken from another population mix, although growth charts generated by ultrasonography are similar for many races except South Eastern Asians.

Crown-rump length
From 7 to 12 weeks the length of the embryo's body can be measured precisely from the crown of the head to the tip of the rump. This measurement is helpful in dating the maturity of an embryo or early fetus, but after 12 weeks it becomes less reliable because the fetus flexes and extends to a greater degree.

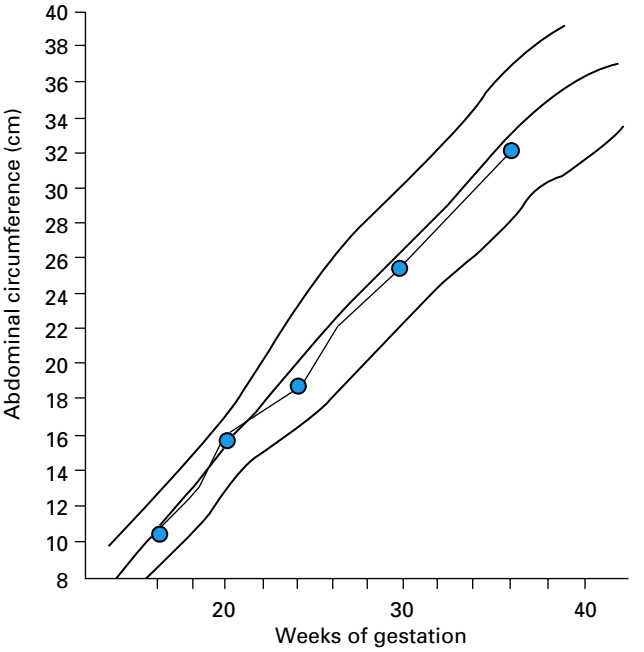


Figure 4.4 Abdominal circumference by weeks of gestation showing the mean ±2 SD. The variability is slightly wider than that in biparietal diameter but growth rates are almost linear until 38 weeks

Biparietal diameter
The distance between the two parietal eminences of the skull gives a precise measurement of fetal head size. From about 16 weeks the range of variation in a normal population widens so that in the last trimester this measurement is less useful. Early biparietal measurements are extremely helpful in dating the pregnancy with more precision even than using the date of the last menstrual period when the woman is certain of her dates. Currently, this is probably the most commonly used technique of ultrasound fetal monitoring in the Western world. If the date by ultrasound differs significantly from that expected by the last menstrual period (usually > 10 days) a revised EDD is usually calculated.

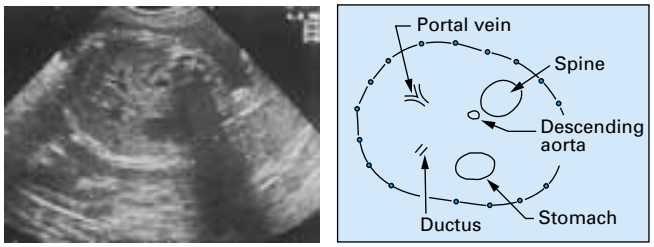


Figure 4.5 Late in pregnancy fetal growth can be detected by examining the fetal abdomen and the circumference can be marked out (by a series of dots)

Abdominal circumference
Measurement of the fetal waist at the level of the umbilical vein provides a good assessment of the size of the fetal liver. Poor fetal nutrition prevents adequate growth of the liver following the failure to lay down glycogen. Serial measurements of abdominal circumference (or area) give good warning of placental insufficiency. A fetus who is growing well is unlikely to die except from an acute event.

Femur length
This can be readily measured from about 12 until 40 weeks. It allows a check on the somatic growth of the fetus. Impaired femur growth with skeletal dysplasia and some chromosomal anomalies invalidates this as a measure.

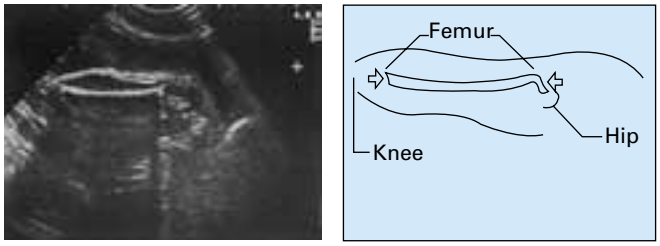


Figure 4.6 Femur length (between two arrowheads) can be measured easily by ultrasonography

Amniotic fluid volume
The estimation of amniotic fluid volume is a measure of fetal metabolism. Volume is estimated by measuring the height of the largest vertical column of fluid detected by ultrasonography. A column <2 cm indicates poor production of amniotic fluid (oligohydramnios).

All these five measurements have different uses at different times of pregnancy.

Early measurements of the biparietal diameter should be used for assessing gestational age. Growth is best assessed by serial circumference measurements of the fetal head and abdomen. In late pregnancy fetal weight can be estimated by using abdominal circumference and biparietal diameter. Assessment of amniotic fluid is an attempt to study dynamic changes as it reflects fetal urine production; this is decreased in placental underperfusion.

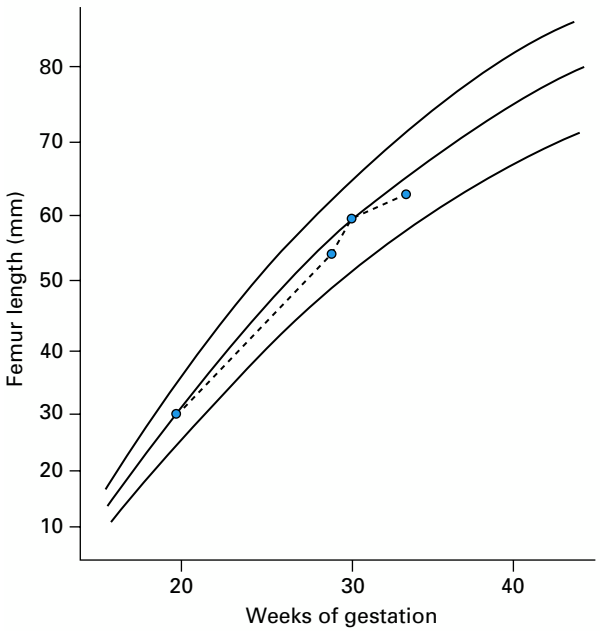


Figure 4.7 Growth in fetal femur length by weeks of pregnancy showing mean ±2 SD

Tests in late pregnancy (29–40 weeks)

One of the main signs of fetal wellbeing in late pregnancy is continued growth, measured by serial ultrasound examination of abdominal circumference. Readings from early pregnancy are needed to give a baseline to the growth measurements in the third trimester. This method of monitoring fetal growth has a high predictive power of detecting poor growth with high sensitivity and specificity.

Movements
Movements of the fetus are felt by the mother from about 20 weeks. In the last 10 weeks of pregnancy they may be used as

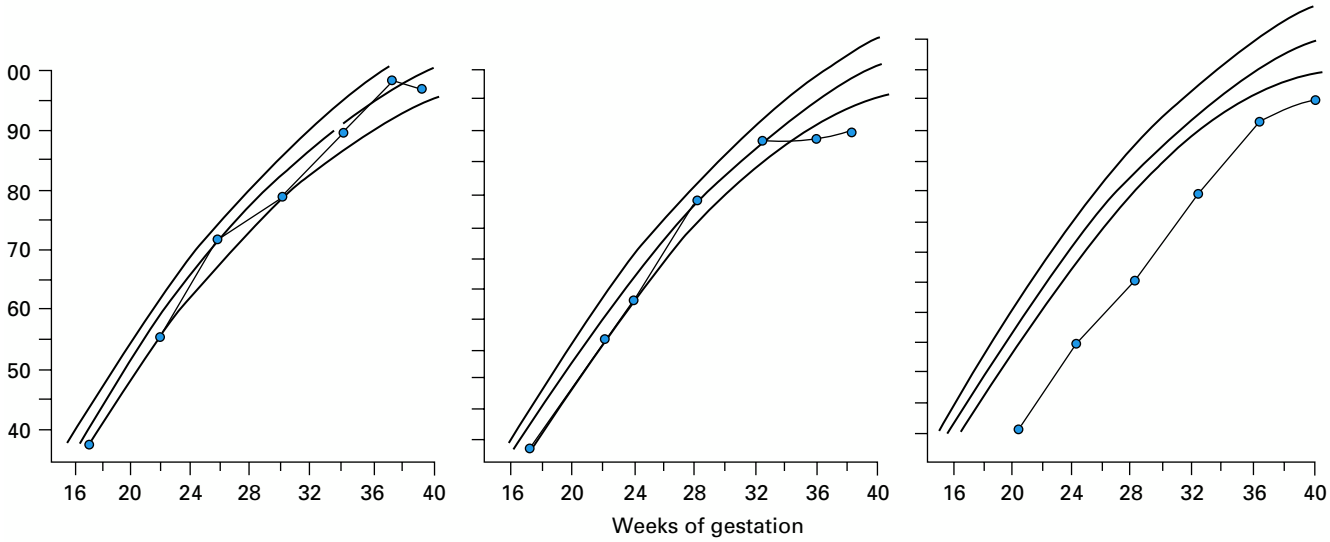


Figure 4.8 Biparietal diameter during pregnancy. Left: Growth follows the normal range of variation and stays within 2 SD of the mean. Middle: Growth tails off from about 32 weeks, the head growing hardly at all in the last weeks. Right: The first reading at 20 weeks is well outside the normal range. If the readings are put back four weeks, growth falls inside the normal range. The woman in this case was probably incorrect in her dates

a coarse measure of fetal wellbeing. Many women feel individual movements distinctly; they record these on a Cardiff Count to Ten kick chart, which estimates how long it takes for the fetus to make 10 movements. In most cases this happens within the first hour or two of the observation period, but fewer than 10 movements in 12 hours may be an early warning sign of problems. The woman should report to her obstetric department for more intensive testing, usually cardiotocography.

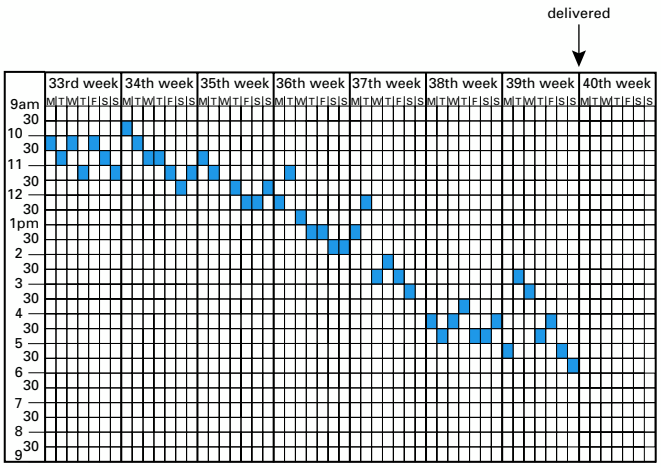


Figure 4.9 Cardiff Count to Ten kick chart. The timing of fetal movements can be graphically displayed on this chart by the mother, who is asked to contact the hospital if there are <10 movements in 12 hours

Cardiotocography

The fetal heart rate bears some relation to the body's response to lack of oxygen—hypoxaemia. This may be measured from 24 weeks by an external ultrasound transmitter and receiver attached to a recording system. The changes in the fetal heart rate in relation to events external to the heart rate such as uterine contractions or fetal movements can be assessed.

The baseline is important, a bradycardia being a warning sign. Episodic changes are more commonly seen, the most healthy being an acceleration; decelerations are of serious import.

Fetal heart rate varies with the balance of the sympathetic and parasympathetic nervous systems, the activity of chemoreceptors in the aorta, and concentrations of adrenaline and acetylcholine. In consequence, when the fetus is awake, baseline variability is normal. Reduced variability so that the trace becomes flat is a sign that the heart is not responding to the interaction of stimuli. This may mean accumulation of metabolic catabolites—that is, fetal acidosis. Sleep patterns need to be excluded from this diagnosis, particularly in less mature babies, as a flat trace can occur for 40 minutes or so when the fetus is asleep. The easiest way to distinguish the two is to wake the baby up by asking the mother to move around and repeat the test.

Other sinister changes include episodic decelerations with or without uterine contractions and a solitary long deceleration lasting for over five minutes.

Cardiotocograph records are widely used in the UK to monitor women at high risk in the later weeks of pregnancy to determine the best time for delivering the baby. Because of the

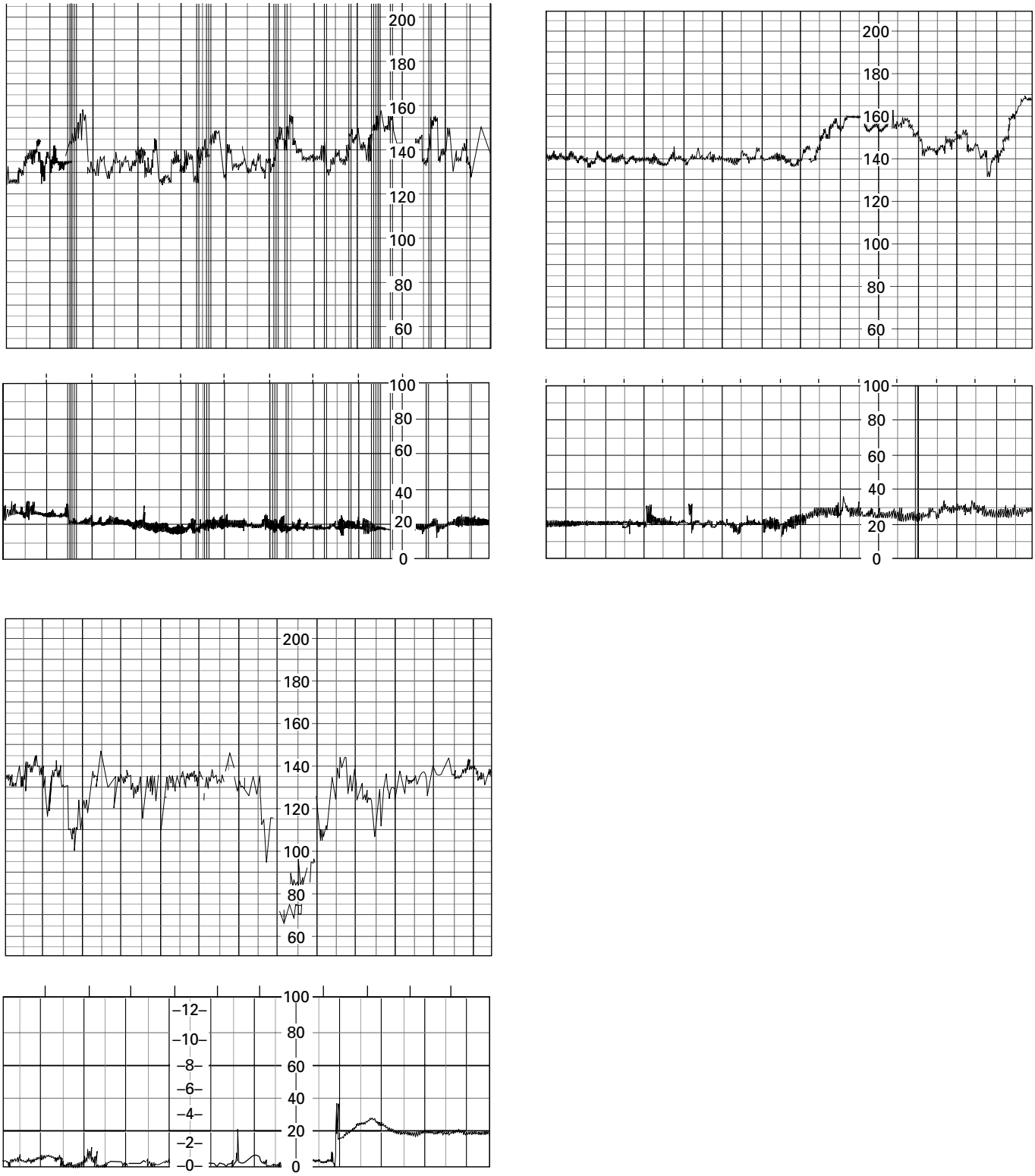


Figure 4.10 Antenatal cardiotocography showing fetal heart rate above and uterine pressure below in each trace. Top left: Fetal movements (shown by the vertical bars) are accompanied by accelerations in the heart rate. Top right: The fetus is asleep but wakes at the end of the trace. Bottom: The heart rate shows episodic decelerations, which have a bad prognosis

poor prognostic value of the individual variables that make up an antenatal cardiotocograph trace, their precise value in prediction is hard to define. A severly abnormal trace probably indicates action but an apparently normal one should not blinker decisions. The trace should be considered with other data from the pregnancy and rarely be regarded as a solitary indicant for action.

Doppler studies

The flow of blood in the arcuate branches of the uterine artery on the maternal side of the placental bed and in the umbilical artery on the fetal side can be measured by the Doppler principle.

- The afferent supply of oxygenated blood to the placental bed through the spiral arteries indicates background nutrition of the fetus in pregnancy giving longer term warning.
- The flow along the umbilical vessels indicates fetal cardiac output, a more acute measure of what is happening at the moment as reduction of flow follows poor fetal cardiac function giving more immediate warning.

Flow in other fetal vessels may help to assess the fetal state. The middle cerebral vessel, the carotid artery, the hepatic or renal vessels may be used. Ultrasound waves are beamed in and their reflected echo patterns vary with the ratio of different flows. This method of monitoring is not yet fully validated, but the interpretation of patterns is beginning to show that it is useful clinically. Abnormal waveforms from the arcuate artery are useful in predicting which women will develop severe hypertension in pregnancy. In a fetus shown to be small by ultrasound measurements, the umbilical artery waveforms help to identify the truly pathological from the constitutionally small baby. Absence or reversal of flow in the umbilical artery during diastole carries a 25–40% mortality, and up to a quarter of survivors have substantial morbidity. Conversely, small fetuses with normal umbilical waveforms have a good outcome. The middle cerebral arteries are the vessels most commonly used to assess cerebral circulation. These show a change from high to low resistance after about 30 weeks, possibly indicating dilatation as normal pregnancy progresses. During hypoxia, blood is redistributed away from the body to vital organs, achieving a brain sparing response. Measuring the ratio of flow in the middle cerebral artery to that in the aorta can give an indication of fetal hypoxia.

Invasive studies

In the second half of pregnancy fetal blood may be sampled by cordocentesis, when the oxygen saturation, carbon dioxide concentration, and concentrations of non-volatile bases such as lactate and pyruvate are measured in small blood samples. Blood is removed from the umbilical vein in the cord; the procedure carries a 1–2% risk of fetal death but the results can be invaluable about the state of fetal acid–base and blood gas concentrations. Furthermore, in some cases chromosome studies yield results that would change management.

Hormone concentrations

The estimation of oestriol concentration (or total oestrogens) in the mother's urine or blood in late pregnancy was used to give some idea of the state of the fetoplacental unit. Unfortunately, the wide variance of results within the normal range did not allow precise enough prediction and the tests have mostly been replaced by biophysical ones.

Testing for progesterone and human placental lactogen has suffered the same fate as that for oestrogen, for the same reasons.

Conclusion

The clinical assessment of the fetus can be further refined by a series of tests. Some are simple and easy to do and are used as screening tests on the whole antenatal population—for example, ultrasound for checking fetal growth. Most fetal

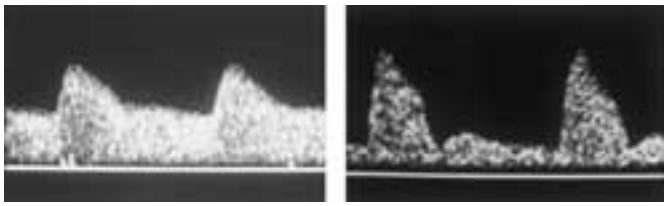


Figure 4.11 Doppler studies showing the waveforms of normal (left) and narrowed (right) arcuate arteries of the placental bed

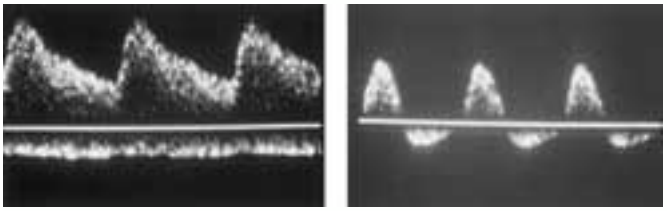


Figure 4.12 Left: Normal waveforms of the umbilical artery. Right: Severely abnormal waveforms of the umbilical artery showing reduced and even reversed flow in the diastolic phase, which suggests that the fetus is compromised

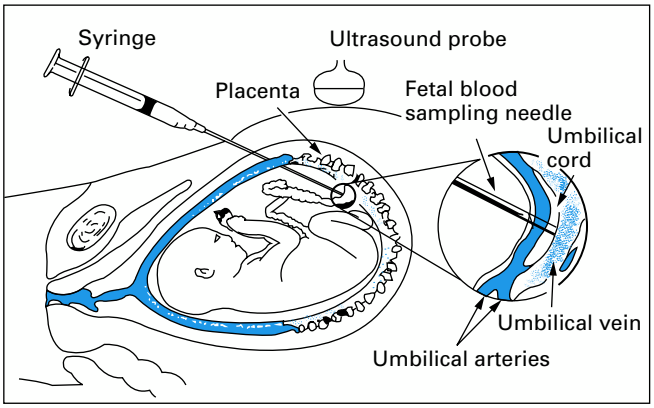


Figure 4.13 Obtaining fetal blood by cordocentesis

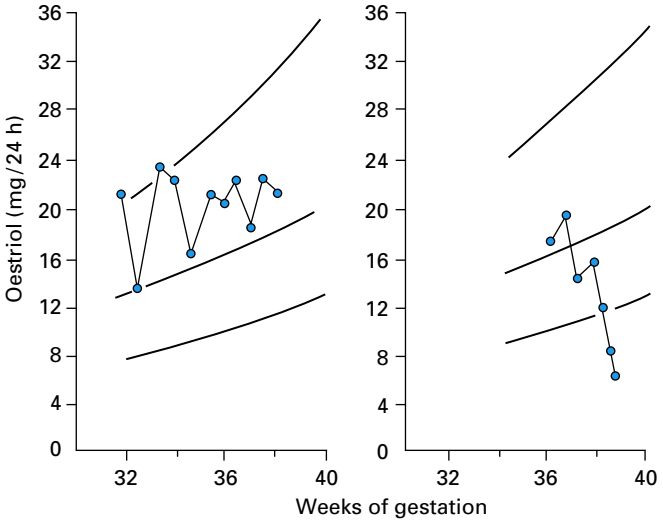


Figure 4.14 Urinary oestriol concentrations (mean (±2 SD)) during gestation. The range is much greater than that of biophysical tests. Left: Variation during a normal pregnancy. Right: Acute placental malfunction in a woman with hypertension. Currently, signs of fetal compromise measured biophysically would have indicated that she be delivered before the last oestriol readings were available

Fetal investigations should be considered to be either screening, for use in large populations, or specifically diagnostic, for use in a selected number of fetuses in which there is clinical suspicion of significant pathological lesions.

investigations, however, should be kept for women who are at high risk of specific conditions—for example, doppler studies in a fetus thought to be small.

The development of more complex biophysical tests has led to a concentration of antenatal care for women at high risk in specialist hospital units. Many district general hospitals do not have all the facilities required so the proper use of regional centres for specialist tests must be encouraged. It may be unpleasant for a woman to have to move from her home area to a centre 40 or 60 km away, but this is usually acceptable if benefits of fetal diagnosis and treatment can be explained by the doctor or midwife so that the woman realises she is helping her baby. Unfortunately, resources and skills cannot be spread uniformly throughout the country. A natural resistance to the new does happen in medicine, but it is Luddite to ignore new investigations merely because they were developed after the practitioner qualified.

Recommended reading

- Gaziano E. Antenatal ultrasound of fetal Doppler. *Clin Perinatal* 1995;**22**:111–40.
- Ingemarsson I, Ingemarsson E, Spencer J. *Fetal heart rate monitoring*. Oxford: Oxford Medical Publications, 1993.
- RCOG. *Use of anti D immunoglobin for Rh prophylaxis*. Guidelines no. 22. London: RCOG, 1999.

5 Detection and management of congenital abnormalities

A congenital abnormality in their expected baby is greatly feared by couples; we are not many generations away from the superstitious who looked on malformation as a retribution for moral misbehaviour. Congenital abnormalities are still one of the major causes of perinatal mortality and morbidity.

The known causes of abnormality are genetic or environmental. Genetic abnormality depends on the chromosomes we inherit from our parents together with the breaks and realignments occurring at fertilisation. As well as the single gene diseases and those following chromosomal rearrangement at meiosis, a large number of conditions appearing in later life, such as hypertension and some cancers, are genetically associated. A whole new philosophy of preventive care is opening. Maternal ageing also increases the risk of abnormalities in genes. A good account of genetic abnormalities is found in the *ABC of Clinical Genetics*.

If parents ask about the risks of recurrence of congenital abnormality, having already had one child with a problem, there are two sets of variables the practitioner has to consider: those who may have a similar defect and those who may have a different defect. Whilst the former are fairly high there is also a small increased risk of new defects.[1]

Table 5.1 Risks of similar and dissimilar congenital abnormalities in the second infants of mothers with an affected first infant*

	Defect in second infant	
Defect in first infant	**Similar defect relative risk†**	**Dissimilar relative risk†**
Talipes	7·3 (5·9–9.1)	1.4 (1.0–1.7)
Limb defect	11.3 (7.2–17.0)	2.4 (1.7–3.3)
Cardiac defect	6.0 (2.2–13.0)	1.1 (0.5–1.9)
Cleft lip	31.4 (19.0–52.0)	2.2 (0.6–2.2)
Cleft palate	44.5 (9.0–13.4)	0.7 (0.1–2.5)

* Based on 1.5 million births in Norwegian Medical Birth Registry.[1]
† 95% confidence intervals for the odds ratios in parentheses.

Environmental factors interfere with embryonic development at a precise stage of organogenesis. They are difficult to pinpoint and often are misassociated. Obvious insults such as exposure to thalidomide, *x* rays, and rubella can be identified; more difficult is the precise place of factors such as organic solvents in the cleaning industry and infections such as toxoplasmosis, cytomegalovirus, and parvovirus.

An antenatal service should aim at diagnosing congenital abnormalities as early as possible. Though the ideal treatment is prevention, this is too late by the time the woman joins the antenatal clinic. If an abnormality can be detected early the couple may be offered the choice of a termination of pregnancy. This is allowed in the UK under statutory grounds of the modified Abortion Act 1967 with no gestational age limit (Fig. 5.3).

Not all couples want to abort their unborn child even if it has an abnormality. Antenatal diagnostic facilities should be made available not only to those who wish a termination pregnancy if an abnormality is found but to affected couples for the reasons given in Box 5.1.

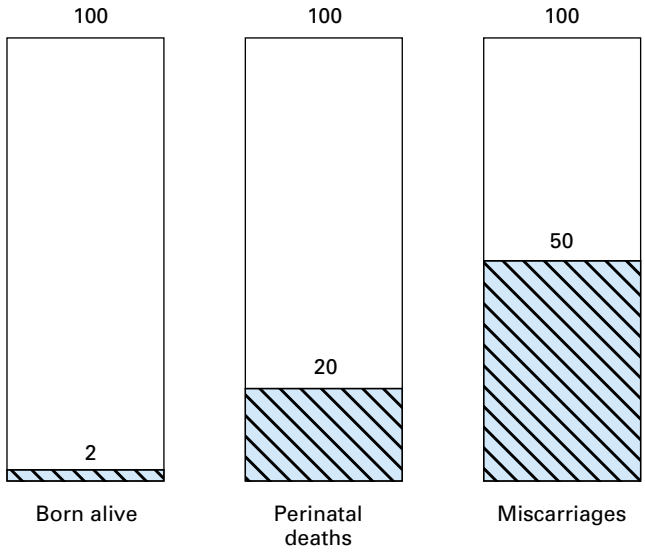


Figure 5.1 Proportions of congenital abnormalities by outcomes of pregnancy

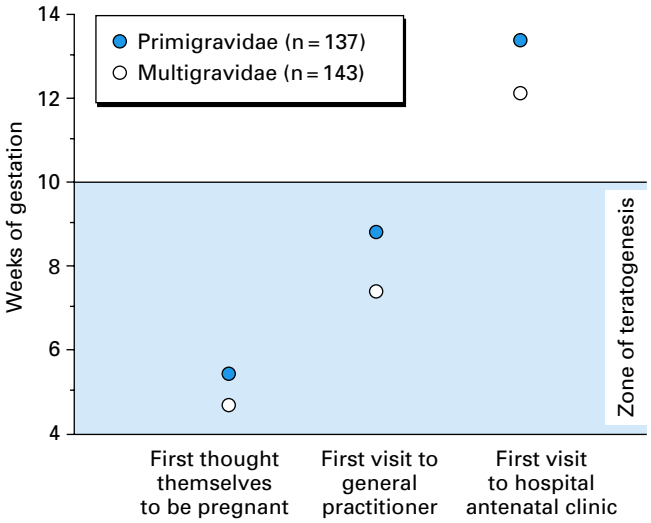


Figure 5.2 When women joined antenatal care. By the time they arrived in the hospital clinic it was far too late for advice to have any influence on the embryo in the first trimester. Maybe the GP visit is the time to provide help

Box 5.1 Reasons for providing an antenatal diagnosis of congenital defect to all relevant couples

- It gives the couple more time to accustom themselves and other children in the family to the idea that an abnormal child is to be born.
- If the abnormality is not lethal, early diagnosis allows plans to be made for delivery in a centre where full treatment may be given early.
- If the abnormality is serious and possibly lethal, counselling for termination of pregnancy can be considered.
- If one of a pair of twins shows a serious anomaly, the option of fetocide can be discussed.

Currently antenatal screening for congenital abnormalities is still mostly concerned with detection of malformations of the central nervous system, the skeleton, and abnormalities of chromosomal origin. The diagnosis of abnormalities of the cardiovascular, alimentary, and urinary tracts is improving; many of these abnormalities are now treatable by neonatologists and paediatric surgeons. Hence, if diagnosed in pregnancy, a woman can be transferred to a hospital where such skills are found. She can have antenatal counselling and the baby can be treated at the best time after the delivery, promptly if necessary.

Testing in early pregnancy

Chromosomal problems

To examine fetal chromosomes, fetal cells are required. In chorionic villus sampling (CVS) a minute piece of trophoblast tissue is removed for examination of the chromosomes in the cell nucleus and, with increasing confidence, DNA assessment. Such sampling is commonly performed at 10–12 weeks of gestation, and a preliminary result is available a couple of days after the test compared with the delay of at least 14 days for amniocentesis. Sampling is under ultrasound control by the transcervical or the transabdominal route (Fig. 5.4). Reports of fetal damage being associated with the transabdominal approach before 10 weeks gestation have led to reservations about its use.[2] There has been a concomitant trend to use amniocentesis at an earlier stage. It is now possible to get a good sample of cells from amniotic fluid safely at 12–13 weeks, so this too is swaying some against CVS.

The attraction of CVS and its quicker results are offset by the higher risks of stimulating a miscarriage. Abortion rates associated with CVS are 2–4% compared with 0.3–1.0% with 16 week amniocentesis. At 10–12 weeks of gestation, however, the rate of spontaneous miscarriage is biologically much higher than at 16 weeks, so the comparison is not only of techniques. Many obstetric units are now using CVS for women at high risk, and with experience the miscarriage rates would be expected to fall.

A nationally organised randomised controlled trial found that CVS had more problems than amniocentesis in diagnostic accuracy, safety, and the need for further testing. However, the obvious advantages of earlier testing and receiving a quicker answer must be weighed against this. Doctors would do well to refer women asking for either procedure to a department of obstetrics that performs both and will give impartial and balanced advice in each individual case. In many centres both tests show similar risk rates.

Trisomy 21 (Down's syndrome) is much commoner in women over 35, but still half of the babies with this condition are born to women under that age. Although the risks to mothers under 35 are less, the overall number of babies born is much greater. To negate this, simple screening is required since both CVS and amniocentesis are unsuitable, invasive procedures

We hereby certify that we are of the opinion, formed in good faith, that in the case

of .....
(Full name of pregnant woman in block capitals)

of .....

.....
(Usual place of residence of pregnant woman in block capitals)

(Ring appropriate letter(s))

A the continuance of the pregnancy would involve risk to the life of the pregnant woman greater than if the pregnancy were terminated;

B the termination is necessary to prevent grave permanent injury to the physical or mental health of the pregnant woman;

C the pregnancy has NOT exceeded its 24th week and that the continuance of the pregnancy would involve risk, greater than if the pregnancy were terminated, of injury to the physical or mental health of the pregnant woman;

D the pregnancy has NOT exceeded its 24th week and that the continuance of the pregnancy would involve risk, greater than if the pregnancy were terminated, of injury to the physical or mental health of any existing child(ren) of the family of the pregnant woman;

E there is a substantial risk that if the child were born it would suffer from such physical or mental abnormalities as to be seriously handicapped.

Figure 5.3 Part of the modified certificate A of the Abortion Act 1967 (revised 1991)

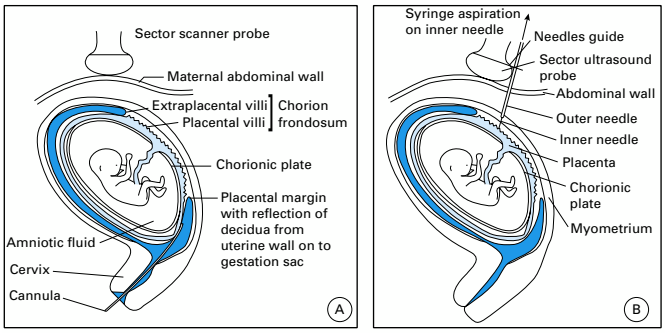


Figure 5.4 Chorionic villus biopsy. Under ultrasound guidance there can be (A) a transcervical approach of the cannula to the edge of the developing placenta or (B) a transabdominal aspiration by needle from the middle of the trophoblast mass

Table 5.2 Refining the risk of Down's syndrome by comparing age alone with results of age and the Triple Test using levels (multiples of mean, MoM) of α fetoprotein (α FP), oestriol, and human gonadotrophin (hCG)

Age alone			**Age (years)**				
			≤20	**21–25**	**26–30**	**31–39**	**≥40**
			1:1530	1:1350	1:900	1:385	1:110
Age and triple test (MOM)							
α FP	Oestriol	hCG					
>0.5	>0.5	<2.0	1:1200	1:100	1:70	1:30	1:10
>2.0	>2.0	<0.5	1:140 000	1:120 000	1:84 000	1:35 000	1:1000

that are very labour intensive. Hence the combination of biochemical tests of maternal blood in early pregnancy to screen for Down's syndrome is widely used. Maternal α fetoprotein, human chorionic gonadotrophin, and oestriol concentrations are measured and the relative risks of each are computed along with the risk advanced by the mother's age at any given gestation. A double test leaves out the oestriol estimation. A combination of risk values for these four markers plus age gives a detection rate (see Table 5.2) and provides the odds of the fetus being affected with Down's syndrome given a procedure screening result. Thus women can be identified who are at high enough risk to justify the hazards and costs of amniocentesis or CVS, the only currently practical ways of getting fetal cells and performing a diagnostic procedure. If the gestational dates used are derived from ultrasound rather than clinical measures, a 10% increase in detection rate is found. Current research is assessing additional markers in both urine and blood at 12 and 15–16 weeks gestation, including pregnancy associated protein A, inhibin A, and urinary β hCG.

Such screening should be offered in all women irrespective of age. It is illogical to restrict it to the over 35-year-olds. As a screening method, maternal age and biochemical estimates would replace the poorer age-only based screen, giving the woman a more precise prognosis of risk and so allowing a more informed decision before going into the diagnostic test of checking fetal chromosomes.

Fetal cells can be detected in the mother's serum as early as 6 week gestation. Though few, they can be identified and isolated. With DNA reduplication, the chromosomal material can be increased and examined. So far male cells have been identified; only female adult cells are present in the mother, any with male chromosomes must have crossed the placenta from the male fetus. Soon DNA manipulation could allow chromosomal abnormalities to be detected and may remove the need for amniocentesis but this is not yet clinical practice.

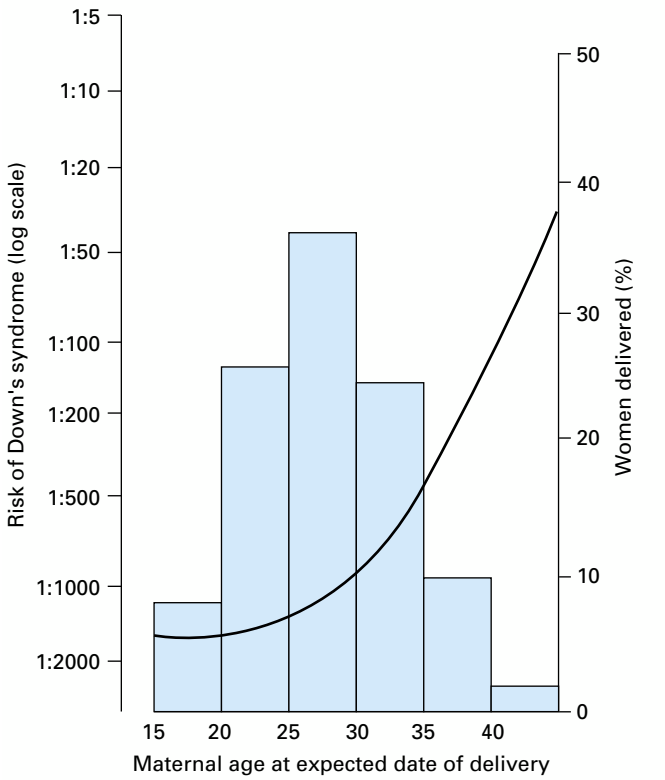


Figure 5.5 Risks of Down's syndrome. The risk increases after 35 and sharply after 40. The percentage of women in the UK who deliver by each age group is also shown. Although the risk is high after 40, the numbers of women delivering are small

Ultrasound tests

The association of ultrasound measured fetal nuchal translucency with Down's syndrome (Table 5.3) and other chromosomal abnormalities has a sensitivity exceeding that of serum screening and a high predictive value. Equally importantly, nuchal translucency measurements correlate with serum screening levels, so that the combination of both tests might increase sensitivity and specitivity.

The big advantage is that this is a non-invasive test done at about 10 weeks of gestation so might be used as a screening test to indicate which women should go forward for fuller chromosome analysis by CVS or amniocentesis.[3, 4] However, sensitivity varies widely according to the skill of the ultrasonographers and the equipment they use.

Those experienced in using the test claim that sensitivity will improve and are now adding a series of other ultrasound markers of Down's syndrome (Box 5.2) in later pregnancy.

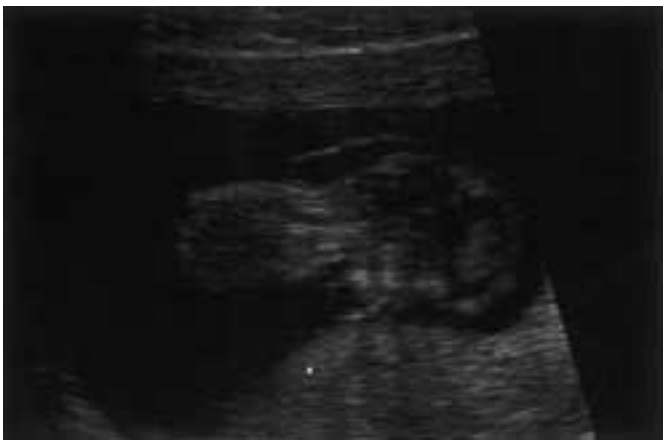


Figure 5.6 Oedema of the back of the neck shows on ultrasound; increased oedema is associated with increased rates of chromosomal abnormalities

Table 5.3 Rising risk of Down's syndrome with increase in nuchal skin fold oedema (from Nicholaides *et al. Br J Obs & Gyn* 1995;101:782–786)

Nuchal oedema (mm)	Relative risk of Trisomy 21
<3	0.2
3	3.2
4	19.8
5	28.6
>5	45.2

Box 5.2 Ultrasound soft markers of Down's syndrome

- Dilated renal pelves
- Nuchal translucency/thickness
- Choroid plexus cysts
- Echogenic bowel
- Short femur

Testing in mid-pregnancy

Structural abnormalities

Open neural tube defects such as anencephaly and open spina bifida allow α fetoprotein to escape from cerebro-spinal fluid into the amniotic fluid, whence it is absorbed into the maternal blood, producing higher than normal concentrations. This is the basis of serum α fetoprotein screening performed between 14 and 16 weeks. It is virtually non-invasive, entailing only a blood sample, and has a high predictive value. Fetal gestational age must be estimated by ultrasonography. False positive results can be caused by multiple pregnancy, a dead fetus, bleeding behind the placenta (which may manifest as a threatened miscarriage), and a few rather rarer abnormalities of the fetus such as gastroschisis.

If the serum α fetoprotein concentration is high a special ultrasound scan may be performed to examine the spine and head carefully at 16–17 weeks of gestation. In some units ultrasound alone is used for the detection of neural tube defects.

By 20 weeks the fetus can be seen clearly on ultrasonography and many neural tube defects will have been detected. At this gestation the heart can be examined and the four chambers identified. The appearance and orientation of

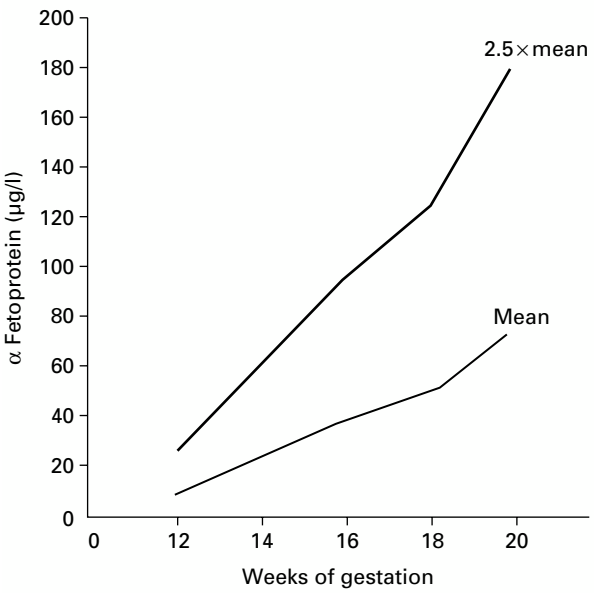


Figure 5.7 Maternal serum α fetoprotein concentration by weeks of gestation. The lower line is taken as the upper boundary of the normal group so it is important to date the pregnancy precisely – that is, by ultrasound measurement of the biparietal diameter. This is magnified by ×2.5 the mean to exaggerate differences

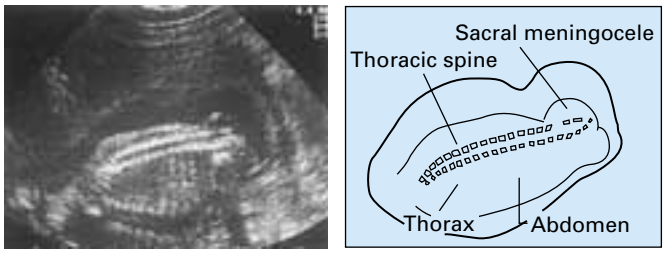


Figure 5.8 Ultrasound scan of a fetus with a sacral meningocele taken at 19 weeks

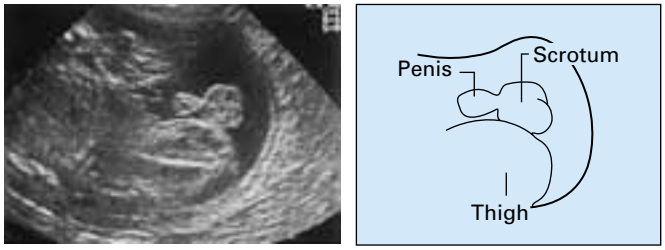


Figure 5.11 Ultrasound scan showing male genitalia resting on the section of the thigh in a fetus of 30 weeks

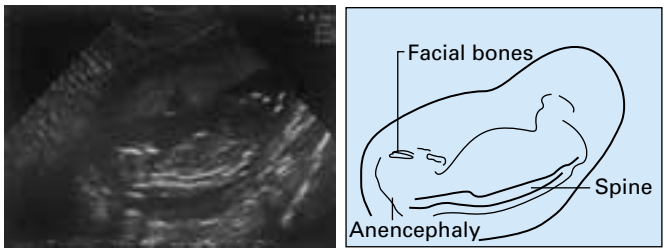


Figure 5.9 Scan of a fetus with anencephaly taken at 17 weeks

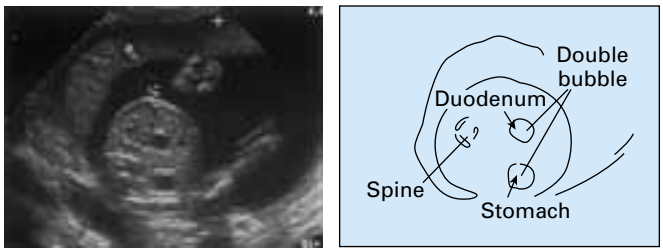


Figure 5.12 Double bubble effect of duodenal atresia, with the lower bubble in the stomach and the upper bubble in the duodenum. Normally continuity can be traced between these two bubbles

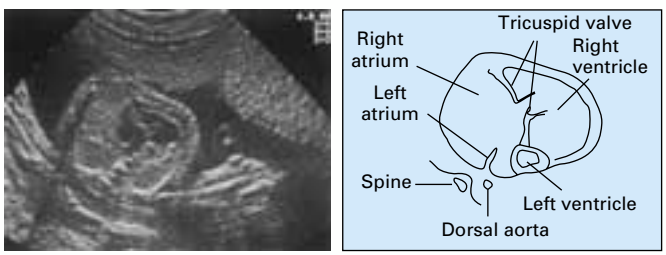


Figure 5.10 Ultrasound scan of the heart of a fetus with mitral atresia taken at 23 weeks gestation

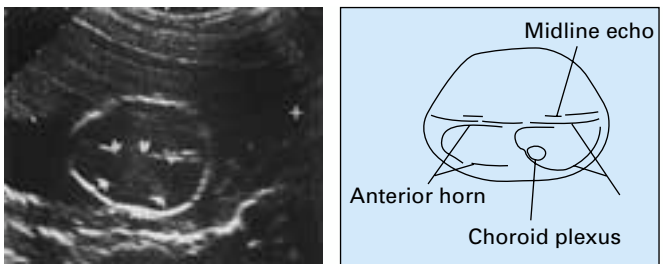


Figure 5.13 Hydrocephalus showing enlarged posterior horn of the ventricle (between single arrows) and anterior horn (between double arrows)

the great vessels can also be checked so that major cardiac abnormalities can be excluded. Limbs can be seen to exclude any shortening and, if relevant, the sex of the child may be determined by sighting the external genitalia.

Later still the kidneys may be assessed for cysts or damming back of urine, producing hydronephrosis. Blockage in the intestinal tract can be checked by the presence of bubbles of fluid in the stomach, duodenal, or large bowel area. The cerebral cortex and ventricles can also be easily visualised and measured; any persistent choroid plexus cysts can be detected. Structural abnormalities of the limbs and digits will be apparent later and some degrees of cleft lip or palate can be found.

The volume of amniotic fluid can be calculated from measurements inside the uterine cavity or more pragmatically by measuring the longest column at the maximum diameter of the largest fluid pool.

These investigations permit a thorough knowledge of the unborn child. Many of the skills are available in the ultrasound departments of district general hospitals, but there is more expert back up at the special obstetric ultrasound clinics of tertiary referral hospitals.

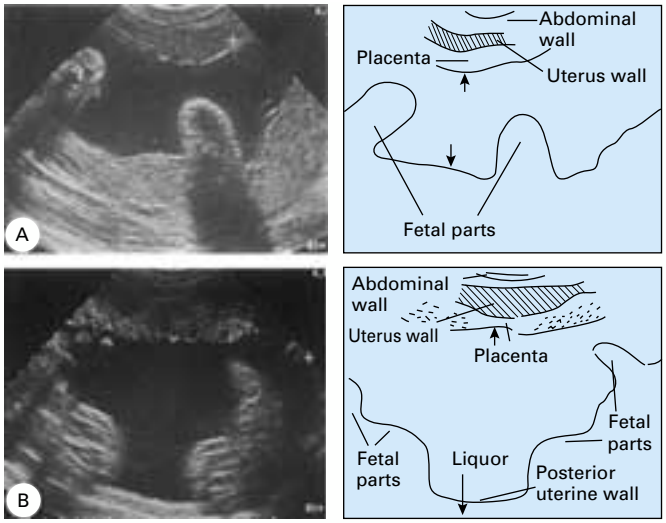


Figure 5.14 Amniotic fluid estimation. (A) The largest pool has the longest column of 5.3 cm (between the arrows); this is normal. (B) In polyhydramnios the longest pool between the arrows has a column of 9.1 cm. Generally 8.0 cm is taken as the upper limit of normal

Chromosomal abnormalities

In mid-pregnancy the chromosomal state of the fetus may be checked from cells removed at amniocentesis. Early amniocentesis (before 14 weeks) appears to be associated with significant problems including increased fetal loss, fetal talipes, and difficulty with culturing the chromosomes. New polymer chain reaction techniques have enabled preliminary results to be available within 24 hours but the full results for chromosome cultures still take some weeks.

The commonest use of amniocentesis is for the diagnosis of Down's syndrome (trisomy 21), and in most parts of England and Wales women over the age of 35 are offered this screening test. Serum screening of hCG, α fetoprotein and oestriol is used to determine those at higher risk of Down's syndrome. Amniocentesis is an invasive procedure with a small risk of spontaneous miscarriage (0.3–1.0% above background rate of miscarriage). This risk is less if the procedure is done under ultrasound guidance by an experienced obstetrician (0.3–0.8%).

Occasionally from about 20 weeks of pregnancy it is necessary to be certain that the fetus has normal chromosomes if a high risk pregnancy is to be continued under adverse circumstances. It is wise to know that the baby is normal before putting the mother through many weeks of anxiety and possibly a caesarean section. The white cells of fetal blood can be obtained at cordocentesis by penetrating the umbilical cord where the vessels are held firmest, close to the placenta or to the fetal belly wall. Chromosome examination of the white blood cells gives a result fairly speedily (two or four days).

(NORMAL ✓ NOT SEEN *NS*)

DATE ..... GESTATIONAL AGE BY EDD ..... wks			
CRANIUM		HEART – 4 chambers	
VENTRICLES		STOMACH	
CEREBELLUM		KIDNEYS	
SPINE		BLADDER	
4 LIMBS SEEN		CORD INSERTION	

Figure 5.15 A typical anomaly checklist to be completed at the 18–22 week ultrasound scan

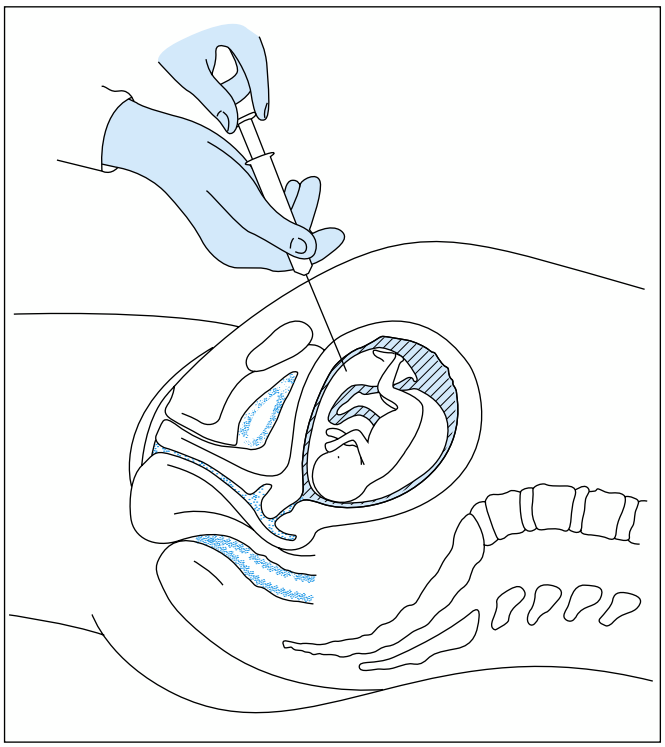


Figure 5.16 Amniocentesis under local anaesthesia. The fluid withdrawn (about 10–15 ml) is spun down and the cells are used for culture

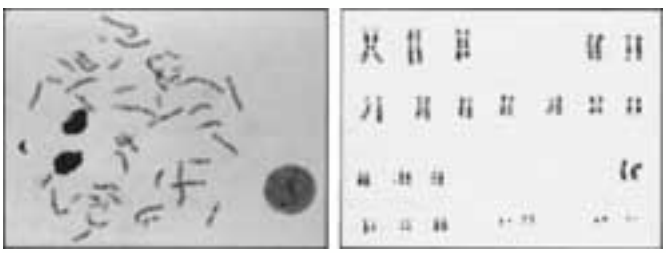


Figure 5.17 Metaphase spread of chromosome material from a nucleus after culture. The chromosomes are photographed and the print cut out and arranged in pairs to show the normal arrangement for a female, two X chromosomes at the end of the bottom grouping

Abnormalities of the central nervous system
The total number of abnormalities of the central nervous system in England and Wales has fallen since the early 1970s. Data are based on three sources:

- notification of termination of pregnancy for abnormalities of the central nervous system;
- death certification of stillbirths and neonatal deaths because of abnormalities;
- notification of abnormalities of babies who live.

Whilst rates are at 1.75 per 1000 in Wales, there is a differential in the south of Britain, where the proportional decrease is even greater. In many parts of southern England, the rate of abnormalities of the central nervous system is less than 1 per 1000. At this level a screening programme that used α fetoprotein might do more harm than good because action might be taken on false positive results. Many authorities have abandoned biochemical screening for these reasons.

Ultrasound as a screening test for anencephaly has good results, and when modern, high resolution equipment is available spina bifida can be detected. Although the special skills and equipment are currently not always available in DGH ultrasound clinics, regional centres do provide them.

Availability of tests

Biochemical screening for abnormalities of the central nervous system and for Down's syndrome is patchy and varies from one district health authority to another. The reasons lie not just in the whims of economic diktat but with variations in the interpretation of epidemiological data.

Congenital abnormalities

Down's syndrome
As explained previously, the risks of Down's syndrome are greater in women over 35, but because most babies are born to women under this age about half of the babies with the syndrome will be missed if age is used as an indicator for fetal chromosome tests. The use of serum screening with ultrasound has offered younger women the option of testing for Down's syndrome, although costs will have to be considered. To detect one affected fetus it now costs about £15 000 to screen for Down's syndrome. Some health authorities would set this against the cost of maintaining a child born with Down's syndrome for the rest of his or her life in an institution, probably between £17 000 and £30 000. The cost of diagnosis, however, comes from one year's budget, whereas the cost of maintenance is spread over many years' budgets in the future; local health authorities are forced into this philosophical financial juggling.

Many units in the UK have introduced serum screening enabling those women considered to be at high risk by age alone to be reallotted to a lower risk group if the results are favourable. With recent advances in early ultrasound, it is likely that a combination of serum screening and measurement of the nuchal fold will produce the best pick-up rate for the lowest level of false positives.

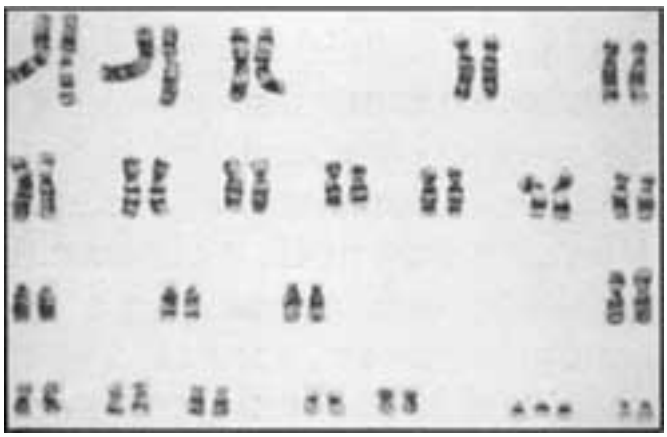


Figure 5.18 Chromosomes of a woman with trisomy 21. The last but one grouping (position 21) has three chromosomes instead of two

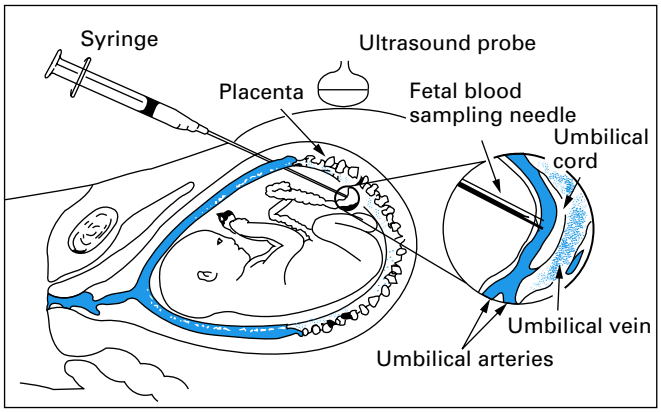


Figure 5.19 Cordocentesis

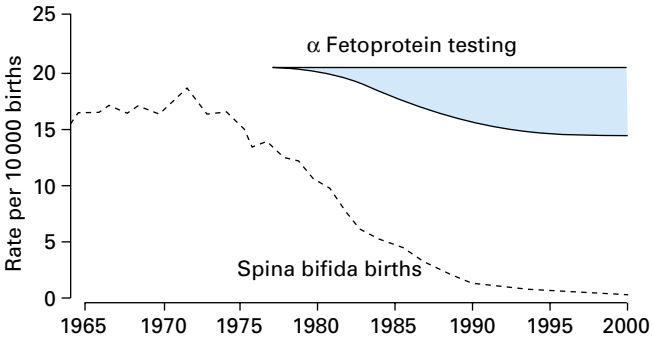


Figure 5.20 Birth prevalence of spina bifida in England and Wales. A slight reduction has occurred from the mid-1960s, becoming sharper from 1973. Testing for α fetoprotein, although described in the early 1970s, was not widespread until the 1980s and so there may be a coincident factor in this reduction as well as the effect on screening. Many think this is due to an improved diet for the women of this country

Conclusion

At first the antenatal detection of congenital abnormalities may seem to lead only to a nihilistic outcome, but the diagnosis can lead to other lines of management such as the preparation for early paediatric surgery or, in future, to genetic engineering. This is unlikely to be of any help once the embryo has started its development, but work done now on forming embryos can be extrapolated back to research on the oocyte. Here recombinant DNA technology may be used to change the affected part of a chromosome before cell development starts, thus producing a normal fetus. Such technology obviously needs to be controlled by society to help couples who previously had no chance of producing a normal baby.

Detection of fetal abnormalities in early pregnancy need not just lead to termination of pregnancy. Many results confirm normality and so reassure the mother. Even when positive, the results lead to the provision of better neonatal services when the affected baby is born.

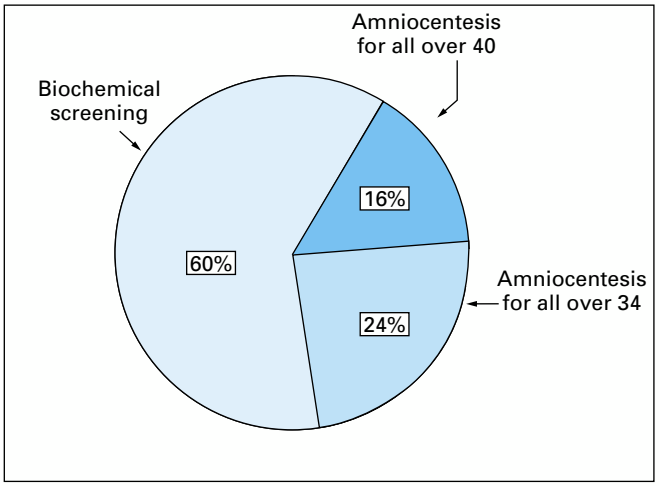


Figure 5.21 Detection rates of Down's syndrome comparing age as the only criterion with the results of triple biochemistry screening to indicate amniocentesis

References

1 Lie R, Wilcox A, Skjaerven R. Population based study on the risks of recurrent birth defects. *New Engl J Med* 1994;**331**:1–4.
2 Firth HV, Boyd TA, Chamberlain P, Mackenzie IZ, Linderbaum RH, Huson SM. Severe limb abnormalities after chorion villus sampling at 56–66 days' gestation. *Lancet* 1991;**337**:762–3.
3 Brizot M, Snijdes R, Butler J, *et al.* Maternal serum hCG and fetal nuchal translucency thickness for prediction of fetal trisomies in the first trimester of pregnancy. *Br J Obstet Gynaec* 1995;**102**:127–32.
4 Bewley S, Robers I, MacKinson A, Rodeck C. The use of first trimester measurements of fetal nuchal translucency. Problems of screening a general population. *Br J Obstet Gynaec* 1995;**102**:386–8.

Recommended reading

- Grundzinskas J, Ward R. *Screening for Down's syndrome in the first trimester.* London: RCOG Press, 1997.
- Hill L. Detection neural tube defects. In Rodeck C, Whittle M, eds. *Fetal medicine.* London: Harcourt Brace, 1999.
- Ott W. *Clinical obstetrical ultrasound.* Bristol: Willey, 1999.
- RCOG. *Amniocentesis.* Guidelines no. 8. London: RCOG, 1996.
- RCOG Working Party. *Ultrasound screening for fetal abnormalities.* London: RCOG, 1997.

6 Work in pregnancy

Both the proportions and numbers of women in the paid workforce have been increasing in England and Wales since before the second world war. In 2000 46% of the workforce were women, many in part-time posts, and this statistic has important implications for childbearing and reproduction.

Other important changes are women working longer in pregnancy and the postponement of starting a family to an older age. Three-quarters of couples need two incomes to pay the mortgage and other loans. When the woman becomes pregnant she receives maternity benefits, but these are poor compared with those in other European countries and income will be reduced. Every woman is entitled to 18 weeks of maternity leave. During the first six weeks of this she gets 90% of her average pay and for the next 12 weeks she gets standard maternity pay which is currently £62.20, going up to £75 per week in 2002 and £100 in 2003, hence the total standard maternity pay is for 18 weeks for those who have worked before pregnancy. Maternity allowance is separate and may be claimed. Currently this is £62.20 a week for women who are employed in pregnancy. There are no deductions for tax or National Insurance contributions. This is paid for 18 weeks when the woman is not working.

These and other allowances change often and practitioners would be wise to update themselves from time to time. Details can be obtained from the local Social Security Office or Factsheets from the Maternity Alliance (45 Beech Street, London EC2P 2LX) who provide up-to-date information on this and many other matters. They are most helpful to the cause of women who work in pregnancy.

In the UK the number of women over the age of 35 having babies has increased in the past 30 years because the years of reproduction are those of career advancement and each pregnancy becomes a gap in climbing the ladder of promotion.

Two-thirds of the women in the paid workforce currently continue to work longer into pregnancy than women did in the 1960s. Whereas some stop around the 28th week of pregnancy, most of them continue into the 34th or 35th week. Women are entitled to maternity leave for six weeks on 90% and 12 weeks on £62.20. This can start from 11 weeks before the expected time of delivery, as certified by a doctor or midwife on the MATB1 form. Most women, however, prefer to have as much time as possible with their newborn child after delivery and so do not leave work early.

In certain circumstances a woman leaving her job during pregnancy is entitled to return after maternity leave up to one year after delivery. The employer must, however, employ more than five people and the woman must have worked with the employer for two years in a full-time job or longer in a part-time post. If she wishes to protect her job she must give her employer 21 days' notice of her intent to stop working and she cannot leave until the 28th week of pregnancy. In return for this the employer must keep the job open for a year and, though the exact job may not be there, a job of an equivalent nature must be offered.

Types of work

It is an implicit and undiscussed assumption (by men) that any woman who works outside the home will continue to keep house as well. Hence housework must always be considered when examining work in pregnancy. All women work in the

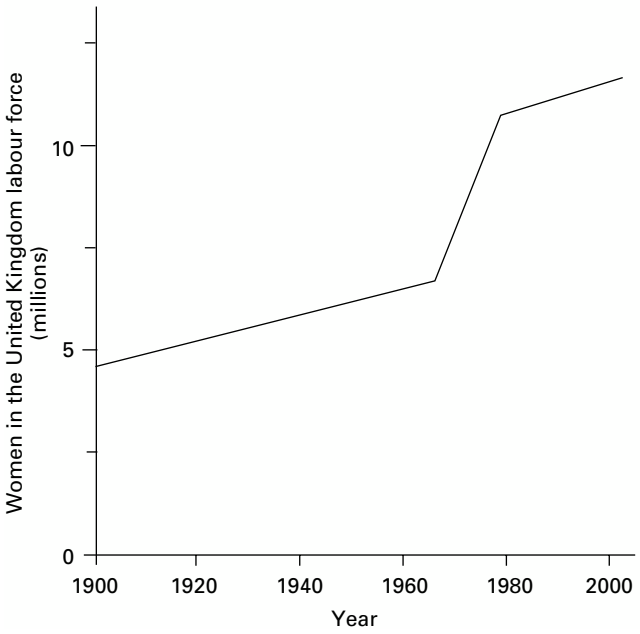


Figure 6.1 Numbers of women in the labour force in the UK

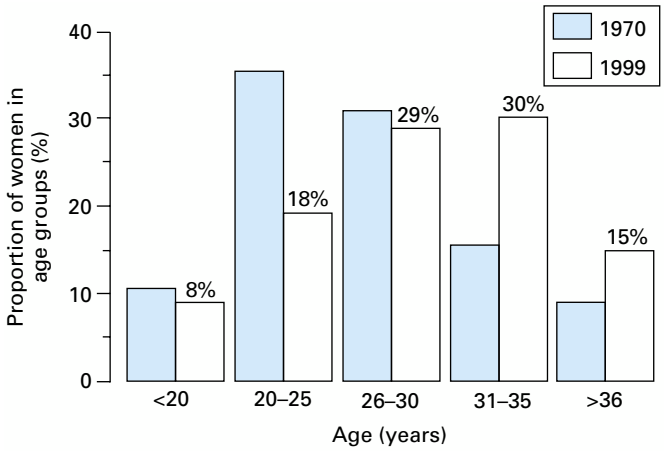


Figure 6.2 Proportions of births in England and Wales by maternal age in 1970 and 1999

Box 6.1 Current maternity benefits (April 2001)

Statutory maternity pay (from employer)
- Non-contributory
- Taxable
- Overlapping
- Paid for 18 weeks—90% of wage for first 6 weeks, £62.20 a week thereafter

Maternity allowance (from DSS)
- Contributory
- Taxable
- Paid for 18 weeks at £62.20 per week

Sure/Start Maternity grant (from DSS)
- £300

Maternity leave
- 18 weeks (see text)

house, where there is washing, cooking, cleaning, and the loads imposed by other children, a husband, and maybe parents. When a woman works at home she has no rest or meal breaks; if she works outside the home as well, housework is often done in the evenings and at weekends.

About 45% of jobs done by women are part time so, although the activity may be great, the number of hours spent away from the home are fewer.

Specific hazards at work

Outside the home three million women work in offices, two million in hotels and shops, and one million in the health service or education; another four million work in a wide range of jobs, though few women in this country do the very heavy jobs that are done by women in the United States and the former Soviet Union, for example. Indeed, in this country under the Mines Act 1889 women are not allowed to work down mines.

Most women are aware of specific hazards in their workplace. These are most important in very early pregnancy, when teratogenic influences may occur at a specific time in embryogenesis. The same stimulus acting later in pregnancy can affect growth, causing intrauterine growth restriction.

Chemical hazards

Over 30 000 individual chemicals are used in industry, with a further 3000 compounds being added each year. It is impossible to test all of them on pregnant animals, and much of the evidence about safety depends on retrospective reports of damage to humans. The number of chemicals that are proved to be teratogenic are few.

If a woman is worried about chemicals in her workplace and consults her family doctor, the doctor would do well to discuss the problem with a health and safety officer or trade union official at the woman's work. If there is no help there, the best reference source is the local or central office of the Health and Safety Executive. Any woman who thinks that she is working with a toxic hazard should discuss this well before pregnancy for it is often too late to start making enquiries in early pregnancy. There are special codes of practice for certain toxic chemicals which safeguard pregnant women and their unborn children. The employer should offer alternative work with no loss of pay or benefits. Toxic chemicals can still enter the mother's body after childbirth and be excreted in milk, so a lactating mother also should take precautions against such chemicals.

Many chemicals have been blamed at some time for affecting an early embryo. This makes big news but when, a few years later, the reports are refuted, it is not newsworthy and often not reported in newspapers.

Box 6.2 Chemical hazards in pregnancy

- Metals—for example, lead, mercury, copper
- Gases—for example, carbon monoxide
- Passive smoking
- Insecticides
- Herbicides
- Solvents—for example, carbon tetrachloride
- Drugs during their manufacture
- Disinfecting agents—for example, ethylene oxide

Physical hazards

At specific times in embryogenesis physical hazards can cause abnormalities. *X* rays are a risk in early pregnancy, particularly if a series of films of abdominal structures are exposed during

MAT B1

MATERNITY CERTIFICATE

Please fill in this form in ink

Name of patient

TO THE PATIENT
Please read the notes on the back of this form ▶

Fill in this part if you are giving the certificate before the confinement.
Do not fill this in more than 14 weeks before the week when the baby is expected
I certify that I examined you on the date given below in my opinion you can expect to have your baby in the week that includes ..... / ..... /.....
We use week to mean the 7 days starting on a Sunday and ending on a Saturday

Fill in this part if you are giving the certificate after the confinement
I certify that I attended you in connection with the birth which took place on ..... / ..... / ..... when you were delivered of a child [ ] children
In my opinion your baby was expected in the week that includes ..... / ..... / .....

Date of examination ..... / ..... / .....

Date of signing ..... / ..... / .....

Signature

Registered midwives
Please give your UKCC PIN here

Doctors
Please stamp your name and address here if the form has not been stamped by the Family Practitioner Committee.

Figure 6.3 Maternity certificate

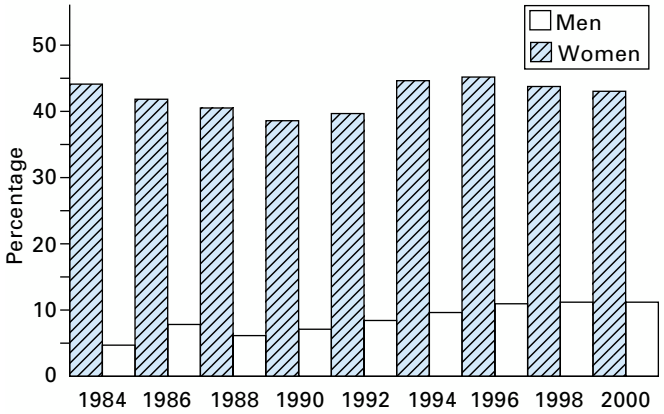


Figure 6.4 Proportion of men and women working part time in the UK

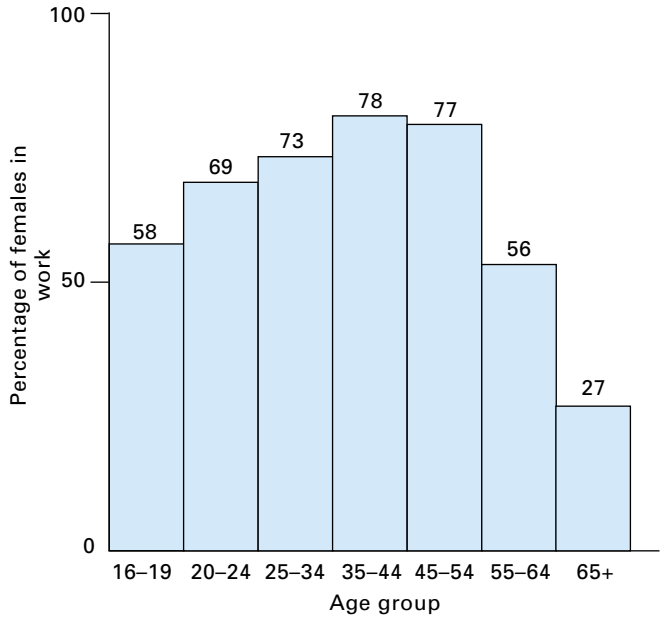


Figure 6.5 Female economic workforce by age

early pregnancy, e.g. for intravenous urography or barium studies of the intestine. It is wise always to ask about the last menstrual period, contraceptive practices, and the possibility of pregnancy specifically before any *x* ray in women of childbearing age. The 10-day rule (whereby no woman is exposed to *x* rays within 10 days of the next menstrual period) has now lapsed in most hospitals but inquiry should be made.

The risks of *x* rays to the female staff in a well managed therapeutic radiation department are probably low, but some women work with radioisotopes in laboratories. The Health and Safety Executive has laid down standards that women should follow. Less well regulated are the *x* ray machines used for security checks in many large firms. There is probably little risk to a visitor passing once through the system, but the people who work the equipment might be exposed to repeated radiation, which should be checked.

Ultrasound is used widely in industry and at the dosage used is probably safe. Certainly, diagnostic ultrasound used in medicine has low energy and is pulsatile; the risk of cell damage or vacuolation that occurs with high energy ultrasound does not exist with this common use. There is no epidemiological evidence of medical ultrasound associated abnormalities: some 60 million women have been exposed to ultrasound in early pregnancy, yet no pattern of problems has yet been shown. Nearly all pregnant women in the UK have one or two ultrasound scans but 46% report having more than two during the pregnancy.

Another physical hazard which caused a scare was the use of visual display units (VDUs) in personal computers (PCs). There are millions of PCs in the homes and offices of the UK. Some 20 years ago small groups of women working with VDUs were reported to have a high rate of pregnancy wastage. These were small clusters, and the measured outcomes were often a mixture of miscarriage, congenital abnormality, and stillbirth. More recent studies show no increased risk due to the use of such units and a wide ranging review concluded, "At present it seems reasonable to conclude that pregnancy will not be harmed by using the VDU. Statements on the contrary are not soundly based."[1]

Biological hazards

Nurses, female doctors, and others who handle body fluids, as well as women who work in microbiological laboratories may be handling toxic materials, but usage is usually well regulated for all workers in or out of pregnancy. Rules must be followed.

Animal workers may be at increased risk, and there have been reports of miscarriage after handling ewes at lambing because of the passage of ovine chlamydia, and toxoplasmosis infection may be more prevalent among those who handle domestic pets in their jobs. The position with bovine spongiform encephalopathy (BSE) for pregnant workers is unclear for too few cases have been documented. There is probably no extra risk over background for the pregnant.

Probably the most commonly transmitted infection which may affect the fetus is German measles. Epidemics occur among young children, and so teachers who are constantly in contact with them are at risk. All young women entering teaching should have their serum rubella antibody titre checked; if they are found to be seronegative they should be offered vaccination.

Non-specific hazards

As well as specified toxins, various physiological changes of pregnancy in the mother might affect the embryo deleteriously. During strenuous exercise the blood supply to the non-skeletal parts of the body are reduced, including the kidneys, intestines,

Box 6.3 Physical hazards in pregnancy

- Ionising radiation—for example, *x* rays
- Noise
- Vibration
- Heat
- Humidity
- Repetitive muscular work—for example, at visual display units
- Lifting heavy loads
- Uninterrupted standing

Figure 6.6 Use of ultrasound for screening

Figure 6.7 Millions of personal computers are used in the UK

Normal pregnant women in jobs with no toxic risk need not be deterred from working for as long as they wish into pregnancy.

Box 6.4 Biological hazards in pregnancy

- Contact in crowded places—for example, in travelling to work
- Contact with higher risk group—for example, schoolchildren
- Food preparation
- Waterborne infections
- New arrivals from abroad

and uterus; the blood supply to the leg muscles can be increased 20-fold and that to the uterus halved. Hence in hard physical work, as occurs in agriculture, there may be some diminution of uterine blood flow, but this is unlikely with ordinary work. Similarly, stress can reduce blood flow to the uterus if the degree of agitation is high enough; if a woman is working inside her own limits there probably will be no problem.

Environmental factors at work that induce boredom and fatigue were found to have effects on pregnant women in a French study.[2] Women in industrial and agricultural jobs were compared with those working in offices. Multivariant analyses of the repetitive nature of the work, the physical effort required, the boredom of the work, standing, and the effect of background noise showed an increased proportion of preterm deliveries when these factors were high, and this might be important in women who have previously had preterm labours.

Recent British and U.S. studies found no effect of work on birth weight.[3] Infants born to women in full-time employment had no significant differences from those born to women who were not in paid work. Data on hours of work, energy expenditure, and posture were collected at 17, 28, and 36 weeks, and these too had no discernible association with birth weight.

It is probable that a higher proportion of those who work are well women and those with chronic ill health do not work. (Further, those of better educational attainments report more out of home work than those with less.) Nevertheless, once the sociobiological variables are removed, work in pregnancy is still associated with a better outcome.

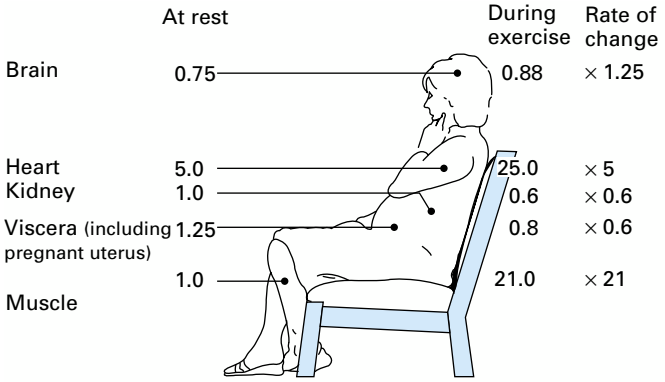


Figure 6.8 Changes in blood flow (l/min) in pregnancy, at rest and during exercise

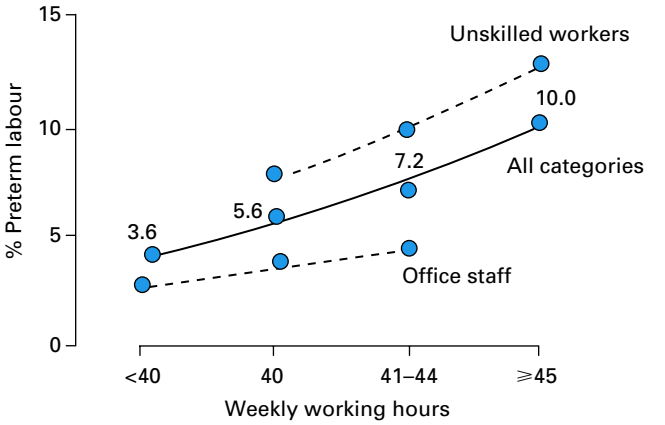


Figure 6.9 Weekly working hours and rate of preterm labour

Travel to work

If a woman has paid work outside the home, she has to get there. If travelling entails a short walk in the morning and evening it can be enjoyable, but most women live in large towns with an unpleasant 30–90 minutes of travel at the beginning and end of the day. There is noise, heat, fatigue, and, in some cases, other people's tobacco smoke. Travel is stressful in crowded, unpleasant conditions. Studies in Spain showed that the likelihood of preterm labour increases with the duration of stressful public travel the woman has to suffer.[4] It may be wise for a woman contemplating pregnancy to arrange to work flexible hours if her work is in a big city. The employer could then perhaps allow her to arrive a little before or after the rush hour, with time being made up in other ways.

Conclusion

More women work during pregnancy and want to continue for longer. Pregnancy is a normal event and, generally speaking, most jobs cause no increased hazard to the mother or baby. A woman should, however, be warned that if any complications arise she must be able to leave work easily. If there is flexibility and the job is not one entailing a high risk from toxic agents most women can continue working for as long as they wish in pregnancy.

When the effects of work on maternal and fetal outcomes are assessed, after adjusting for environmental and background social factors, work seems to have very little detrimental or beneficial influence.

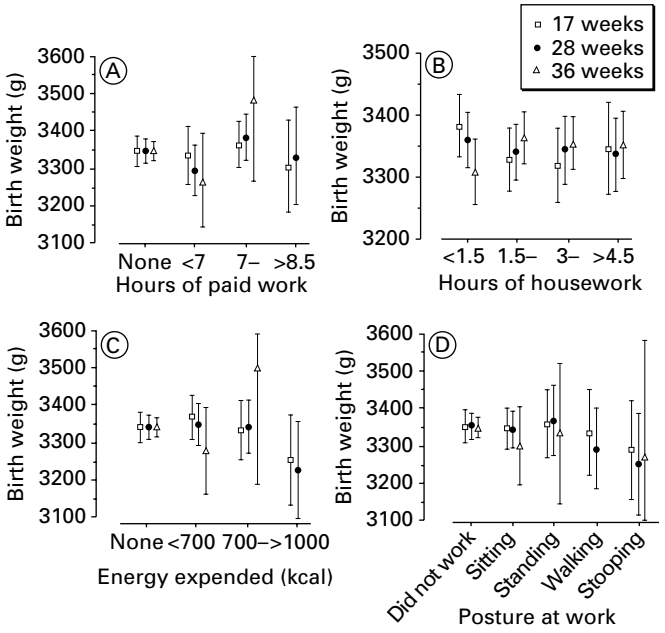


Figure 6.10 Birth weight and work. An antenatal population was sampled at 17, 28, and 36 weeks of gestation[3]

References

1 Blackwell R, Chang A. Video display terminals and pregnancy. *Br J Obstet Gynaec* 1988;**95**:446–53.
2 Mamelle N, Laumon B. Occupational fatigue and preterm birth. In: Chamberlain G, ed. *Pregnant women at work*. London: Royal Society of Medicine, 1984;105–16.
3 Rabkin CS, Anderson HR, Bland JM, Brooke OG, Peacock JL, Chamberlain G. Maternal activity and birthweight. *Am J Epidemiol* 1990;**131**:522–31.
4 Rodrigues-Escudero R, Belanstegreguria A, Gutierrez-Martinez S. Perinatal complications of work and pregnancy. *An Esp Pediatr* 1980;**13**:465–76.

Recommended reading

- Chamberlain G. Work in pregnancy. *Am J Indust Med* 1993;**23**:559–75.
- Henriksen T, Savitz D, Hedegaard M, Secher N. Employment during pregnancy in relation to risk factors and pregnancy outcome. *Br J Obstet Gynaec* 1994;**101**:858–65.
- Maternity Alliance. Factsheets on allowances and benefits in pregnancy. London: Maternity Alliance, 2001.

7 Vaginal bleeding in early pregnancy

Bleeding drives patients to their general practitioner swiftly. Vaginal bleeding early in pregnancy makes the woman think that she may be miscarrying, so this brings her even more promptly; the practitioner thence has the opportunity to diagnose the cause and start management.

Bleeding has four known causes in early pregnancy (Box 7.1). In addition, bleeding may occur for no apparent reason in a large number of cases. In early pregnancy such cases are commonly categorised as threatened miscarriage, but this is fudging the issue for in many cases the conceptus and its future placental system are not involved; doctors should be honest and say that they do not know the cause rather than mislabel it.

Box 7.1 Causes of bleeding in early pregnancy

- Miscarriage
- Ectopic pregnancy
- Trophoblast disease
- Lesions of the cervix or vagina

Box 7.2 Types of miscarriage and abortion

- Threatened miscarriage
- Inevitable miscarriage
  - complete
  - incomplete
- Silent miscarriage
- Recurrent miscarriage
- Criminal abortion
- Septic abortion
- Therapeutic abortion

Miscarriage and abortion

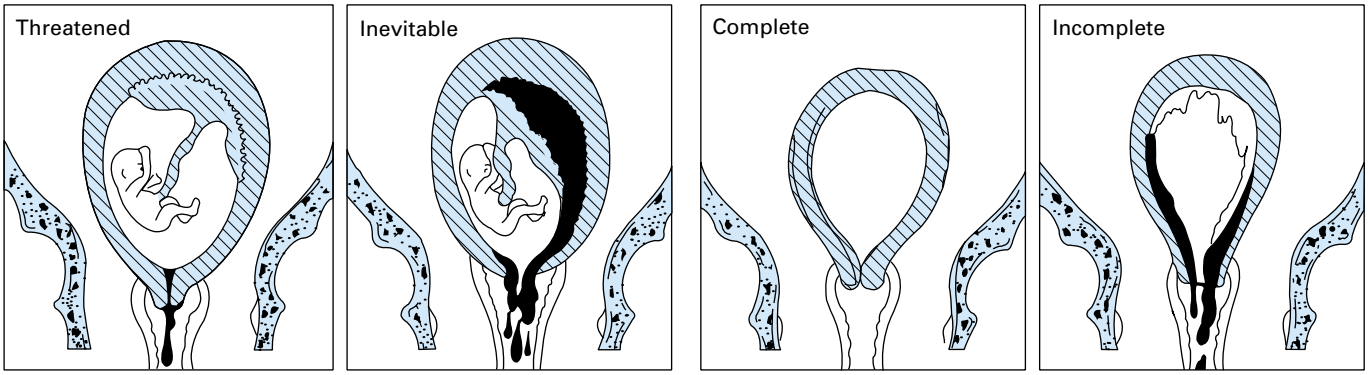


Figure 7.1 In a *threatened* miscarriage the cervix is still closed and there is not much bleeding. In an *inevitable* miscarriage the cervix has started to open and the membranes often have ruptured. There is usually more bleeding. A *complete* miscarriage means that the uterus is empty of clot and decidua. In an *incomplete* miscarriage the embryo has been passed vaginally but some part of the membrane or decidua is retained. There may also be clots

The terms miscarriage and abortion have been used synonymously but miscarriage is the word which should be associated with the spontaneous event.

- *Threatened miscarriage.* Women bleed a little from the vagina during a threatened miscarriage but there is not much abdominal pain. The uterus is enlarged and the cervix closed. Pregnancy may continue.
- *Inevitable miscarriage.* Miscarriage is inevitable if the cervical os is open. Blood loss can be great and lower abdominal cramping pains accompany the uterine contractions. Some products of conception and clots may be passed but often decidua is retained and then the miscarriage is called incomplete.
- *Complete miscarriage.* The cervical os is open and the uterus completely expels its contents. Such miscarriages are more likely after 14 weeks of pregnancy than earlier, when they are often incomplete.
- *Septic miscarriage.* This follows the ascent of organisms from the vagina into the uterus, often after an incomplete miscarriage or an induced abortion under non-sterile conditions. As well as heavy bleeding and pain, the woman commonly has a fever and may develop signs of endotoxic shock. The commonest organisms are *Escherichia coli* and *Streptococcus facecalis.*
- *Silent miscarriage.* The embryo dies and is eventually absorbed but the uterus does not expel the decidua and sac of membranes. The woman sometimes feels a dull weight in

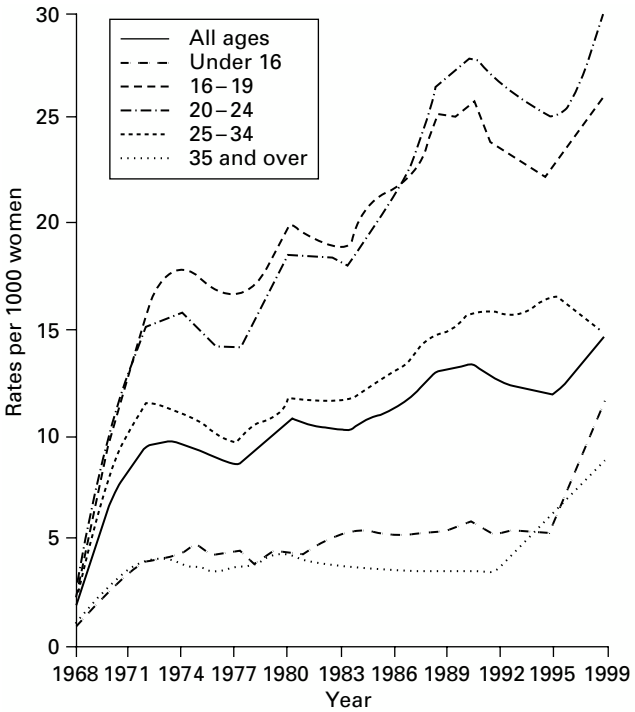


Figure 7.2 Terminations of pregnancy by age group in England and Wales, 1968–99. Many operations are performed outside the NHS, some being done through charity clinics

the pelvis, the symptoms of pregnancy regress and the uterus stops enlarging. Old blood is passed as a brown, watery discharge. This condition is diagnosed more frequently now that ultrasonography is used in very early pregnancy.

- *Recurrent miscarriage* is diagnosed when a woman has three or more consecutive spontaneous miscarriages. Such women deserve gynaecological and immunological investigation; many gynaecologists start investigations after two consecutive miscarriages in women over 35. A specific cause for the recurrence can be found in up to 40% of cases.
- *Therapeutic abortion.* This is now common in Britain, with over 180 000 women in England and Wales having such abortions each year. Usually the general practitioner knows but occasionally the woman has bypassed him, presenting only after the event with vaginal bleeding, an open cervix, and some abdominal pain. This means that decidua or blood clot is left in the uterus and needs the same attention as does an incomplete miscarriage.

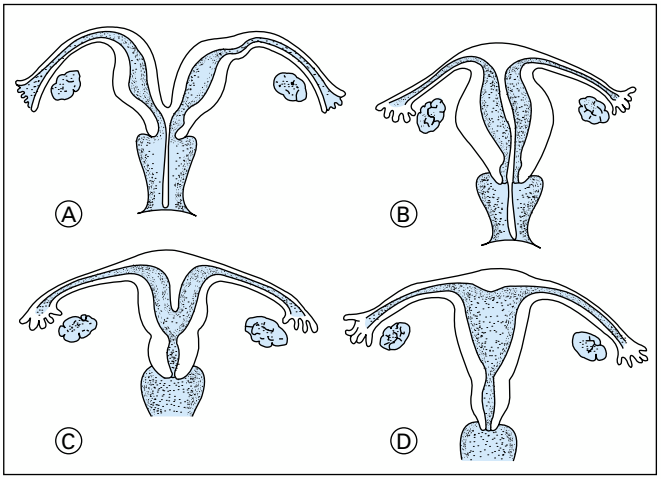


Figure 7.3 Congenital abnormalities of the uterus caused by non-absorption of the septum during fusion of the Müllerian ducts. (A) Complete double uterus, double cervix, and vaginal septum. (B) Double uterine cavity within a single body; the cervix and the vagina have a septum. (C) A subseptate uterus in which the septum does not reach down to the cervix. (D) Arcuate uterus with a dimple on top of the single uterus with a single cervix

Causes of miscarriage

Embryonic abnormalities

Chromosomal abnormalities are common, arising from a change in the nucleus of either gamete or a spontaneous mutation inside the fertilised oocyte. At the time of fertilisation splitting and rejoining of genetic material may be imperfect. Such changes are not usually recurrent, and parents should be told this.

Immunological rejection

The fetus is genetically foreign to the mother and yet most fetuses are not rejected. In many cases blocking antibodies that inhibit the cell-mediated rejection of the embryo are stimulated by antigens from the trophoblast. Antiphospholipid antibodies have been linked with recurrent early pregnancy loss as well as later placental bed insufficiency. This may act through placental thrombosis or decidual vasculopathy. Treatment with aspirin, heparin or a combination offers hope of a successful pregnancy.

Uterine abnormalities

The uterus is formed during embryonic development from two tubes fusing together to make a common cavity. Occasionally various degrees of non-absorption in the midline septum occur, leaving either two cavities or a cavity partly divided by a septum down the middle. The blood supply to this median structure is usually poor and implantation of an embryo here may be followed by miscarriage.

Cervical incompetence

The cervix may have some weakness which could be associated with a spontaneous miscarriage in the mid-trimester (13–27 weeks). This can be either congenital or acquired after overstretching at a previous dilatation and curettage or birth. The unsupported membranes bulge into the cervical canal through the internal os and rupture early, which causes the abortive process. The incompetence may be diagnosed before pregnancy by a hysterogram (a radiological examination of the uterine cavity) or in pregnancy by ultrasonography. Cervical cerclage is usually performed in pregnancy following a history of mid-trimester miscarriage, particularly if the membranes ruptured before any uterine contractions occurred (see Chapter 12).

Maternal disease

This is unlikely to be a major cause of miscarriage in the UK, but hypertension and renal disease are still associated with

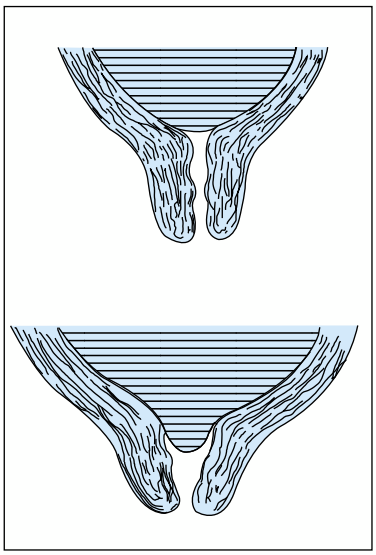


Figure 7.4 Above: normal cervix with maternal os closed protecting amniotic sac. Below: incompetent internal cervical os with membranes bulging

higher rates of miscarriage in later pregnancy. Maternal infections can affect the fetus, particularly rubella, toxoplasmosis, cytomegalic inclusion disease, and listeriosis. Severe maternal malnutrition is most unusual in this country, though it can still occur in developing countries. Deficiency of individual vitamins (such as vitamin E) is extraordinarily rare in the mixed diet of this country, and there is no evidence of it being a substantial cause of miscarriage in women. Shortage of folic acid is associated with fetuses with major central nervous system abnormalities, some of which may miscarry.

Endocrine imbalance

Diabetes and thyroid hyperfunction are associated with increased risks of spontaneous miscarriage. If diagnosed, both are now usually well controlled and the risk is reduced. Abnormalities in the ratio of luteinising hormone to follicle-stimulating hormone in a particular cycle may lead to miscarriage. An insufficiency of progesterone from the corpus luteum used to be regarded as a cause of miscarriage. This is hard to prove, and most randomised trials using progestogens in early pregnancy have failed to show an improvement. Some consider that hCG injections may help. If, however, the woman has faith in this treatment and had a previous successful pregnancy taking it, the practitioner would do well to treat the psyche as well as the soma and prescribe a progestogen or hCG.

Criminal abortion

This is now much less common in Britain but still occurs in other countries and in populations derived from those countries. Although infection has been introduced, only rarely do criminal abortionists leave signs that can be spotted in the genital tract and so the woman is often treated for an incomplete or septic miscarriage. With the reduction in illegal abortion, maternal mortality from this cause has disappeared in the UK.

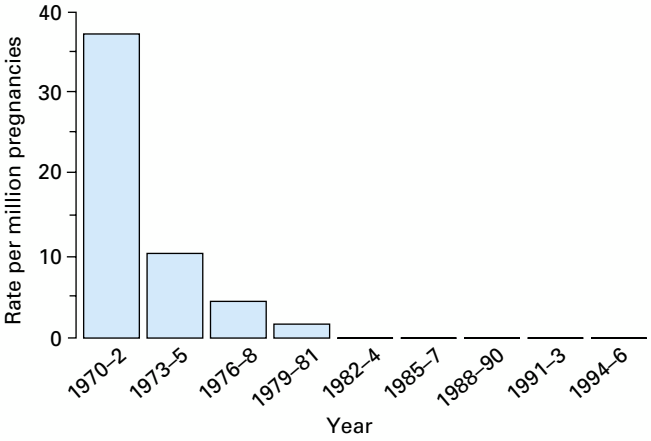


Figure 7.5 The number of deaths reported after illegal abortion is reducing rapidly in England and Wales, with none reported for the 14 years 1982–96. Death used to be mostly from sepsis or renal or hepatic failure

Box 7.3 Treatment of miscarriage

Threatened	Bedrest Avoid intercourse Reassure with ultrasound
Inevitable	Hospitalisation If heavy bleeding, use ecbolic Evacuate uterus

Presentation

A woman who is miscarrying usually presents with vaginal bleeding and may have some low abdominal pain. The bleeding is slight in a threatened miscarriage, greater amounts being present with an inevitable miscarriage. Pain with uterine contractions may be compared with dysmenorrhoea. The degree of shock usually relates to the amount of blood loss from the body and the degree of cervical dilatation.

The differential diagnosis includes ectopic pregnancy and salpingitis.

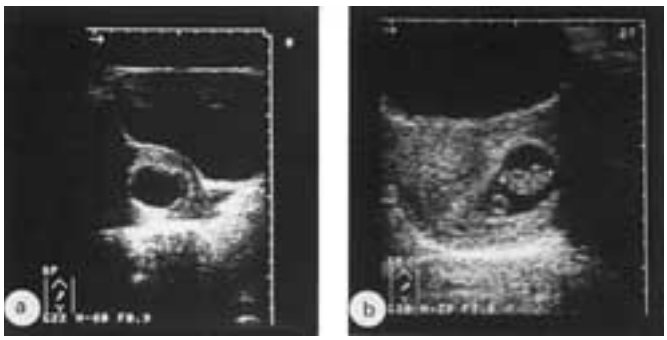


Figure 7.6 (A) Ultrasound scan of an empty sac in the uterus at seven weeks gestation. This woman had a silent miscarriage, the embryo having been resorbed. (B) Ultrasound scan of a continuing pregnancy at just over seven weeks; fetal tissue is easily seen between the crosses

Management

Threatened miscarriage

A woman with a threatened miscarriage is best removed from an active environment. If the practitioner tells her to go to bed to rest for 48 hours she may feel happier but there is no real evidence that bedrest makes any difference to the incidence of miscarriage. Some 5% of women who deliver safely report vaginal bleeding in the same pregnancy; the effectiveness of specific treatments is difficult to assess. The avoidance of sexual intercourse is probably sensible as it might act as a local stimulus.

Inevitable miscarriage

If events progress to an inevitable miscarriage the woman often needs to be assessed in hospital;[1] an ecbolic agent might be

given if the bleeding is excessive and a paramedic may be needed to cover transfer. Ultrasound is useful in determining if the miscarriage is complete. Retained products can be removed surgically under a general anaesthetic or medically using vaginal prostaglandins.

Complete miscarriage
This is more common than was once thought; practitioners may see the sac containing the embryo and feel that this is complete. They would do well to remember, however, that a large amount of decidua can be left behind and an evacuation may prevent the woman having a haemorrhage or infection a week or so later. Ultrasound in an early pregnancy unit can help to make the diagnosis.

Silent miscarriage
This is sometimes diagnosed from the woman's symptoms of a brown discharge and a heavy, dull feeling in the pelvis; the finding of no embryonic tissue inside the gestation sac on ultrasonography confirms the diagnosis. It can also appear as a complete surprise at a routine ultrasound. Management options include a conservative line with a repeat ultrasound, surgical evacuation or medical management with prostaglandins.

Septic abortion
This may require the full management of severe sepsis. Endocervical swabs should be sent to the laboratory and treatment with a broad spectrum antibiotic started immediately. Central venous pressure measurement and intravenous rehydration will be required; the urinary output should be watched carefully. Evidence of disseminated intravascular coagulopathy should be sought and the uterus evacuated once a reasonable tissue concentration of antibiotics has been achieved. Watch for renal failure.

Recurrent miscarriage
The management of recurrent abortion is outside the scope of this series but some aetiological features are given in Box 7.1. It requires sympathetic handling by both general practitioners and specialists.

Box 7.4 Treatment of severe septic abortion

- **Hypovolaemia**
  - Monitor—Blood pressure
    Central venous pressure
    Cardiac output
    Renal output
  - Treatment—Intravenous rehydration and maintenance
- **Infection**
  - Identify organisms
  - Treatment—Systemic
    Antibodies
    —Local
    Evacuate uterus (dilatation and curettage)
    Remove uterus (hysterectomy)
- **Coagulation abnormalities**
- **Renal shutdown**
  - Monitor oliguria
  - Monitor electrolytes
  - Plan early dialysis
  - Watch for renal failure
- **Respiratory system**
  - Monitor—Blood gases
  - Treatment—Oxygen
    Ventilate
- **Anaemia and white cell deficiencies**

Table 7.1 Current aetiological causes of recurrent miscarriage

Cause	Major manifestations
Anatomical	Uterine cavity anomalies, cervical incompetence
Infective	Bacterial vaginitis
Genetic	Incorrect implanting of genome
Endocrine	Defective corpus luteum, luteal phase deficiency
Autoimmune	Systemic lupus erythematosus, antiphospholipid syndrome
No obvious cause is found by current tests in 30% of all cases	

Ectopic pregnancy

An ectopic pregnancy is one that implants and develops outside the uterine cavity. The sites are shown in the figure, but most (96%) are in the fallopian tube.

Causes
Anything that slows the passage of the fertilised oocyte down the fallopian tube can cause a tubal ectopic pregnancy. Previous tubal infection, an intrauterine device in place, and late fertilisation are quoted causes, but in most ectopic pregnancies no cause is found.

Presentation
A tubal ectopic pregnancy may either rupture through the wall (more common with isthmial and cornual implantations) or leak a little blood from the lateral end of the fallopian tube (with ampullary or fimbrial implantations). Vaginal bleeding can occur.

With rupture there is a brisk peritoneal reaction and the woman may fall to the ground as though kicked in the stomach.

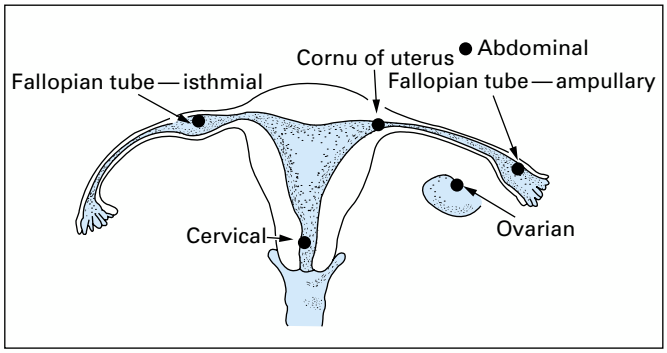


Figure 7.7 Possible sites for ectopic pregnancy

She quickly becomes very shocked because of the large volume of blood released into the peritoneal cavity and the stimulation of the peritoneum. The abdomen is tender with guarding and rebound tenderness, and vaginal examination causes intense pain on touching the cervix.

A more gradual leak from the tubal end causes irritation of the pouch of Douglas. The woman goes to her doctor complaining of vague, low abdominal pain, sometimes with vaginal bleeding occurring after the pain. The abdomen may be uncomfortable in the suprapubic area, and a very gentle vaginal assessment may show a tenderness in the pouch of Douglas or in the adnexa on one side.

The differential diagnosis includes an abortion or any other cause for a sudden release of blood into the peritoneal cavity, such as a bleeding vessel over an ovarian cyst. Inflammatory conditions such as appendicitis may mimic a leaking ectopic pregnancy. Ectopic pregnancy should always be considered in any cases of lower abdominal pain because an unruptured ectopic pregnancy, leaking a little blood over the course of some days, is hard to diagnose. A negative routine pregnancy test is not exclusive; it is usually positive.

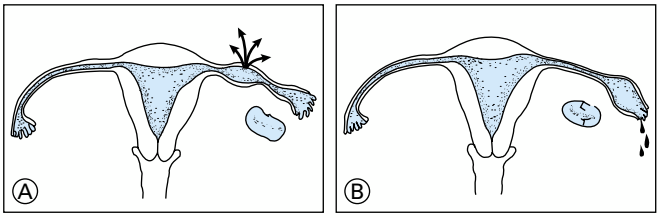


Figure 7.8 The most common ectopic pregnancies are in the fallopian tube. (A) Those at the medial end rupture. (B) Those at the lateral end leak more gradually

Table 7.2 Symptoms and signs of ectopic pregnancy

	Unruptured	Ruptured
Symptoms	Gradual onset Dull ache over days	Sudden onset Severe pain over minutes
Signs	No shock Vague suprapubic tenderness No great cervical tenderness Vague mass often felt	Commonly shocked Rigid abdomen with rebound tenderness Extreme tenderness on cervical movement Too tender to palpate

Management

The management of a woman with a ruptured ectopic pregnancy is straightforward. She should go to hospital immediately, if necessary accompanied by her general practitioner. Intravenous support may be required in the home, and in severe cases a flying squad (if available) may be required. Once in the hospital, surgery should be immediate. Most tubal ectopic pregnancies need a laparoscopy to confirm the diagnosis and maybe provide access for treatment. Those in severe shock due to a ruptured ectopic will also need a laparotomy and surgical removal. Unruptured ectopics can be managed laparoscopically by opening the tube and aspirating the gestational material. If some part of the tube can be left behind it is psychologically helpful to the woman. There is a small risk of a second ectopic pregnancy developing in the remaining tube but there is also the possibility of reparative surgery later. This is particularly important when a woman has had a previous ectopic pregnancy and one fallopian tube has already been removed.

A leaking tubal pregnancy is harder to diagnose, such cases being usually referred to outpatient departments in a more leisurely fashion. If the diagnosis is suspected, laparoscopy is the best test; ultrasonography is not exclusive, although fluid in the pouch of Douglas with no intrauterine pregnancy in a woman with 6–8 weeks' amenorrhoea and a raised hCG level is highly suggestive.[2] At laparoscopy the swollen area of the tube can usually be seen and little blood may come from the lateral end. Under ultrasound guidance, injections of fetotoxic agents such as methotrexate or potassium chloride are given. Alternatively, some, having made a firm diagnosis of ectopic pregnancy by a raised hCG, an empty uterus, and ultrasound evidence of fluid in the pouch of Douglas, will treat the woman conservatively if the ectopic is unruptured. Systemic methotrexate given as a single dose works in 90%. If hCG levels persist, a second dose a few days later usually suffices. The results of such minimally invasive management are as good as those of more conservative surgery, with the time spent in hospital and the emotional effects on the woman's life being much reduced.

Any woman treated for an ectopic pregnancy should be warned of the increased chances of a recurrence. The increased risk is said to be seven times above background.

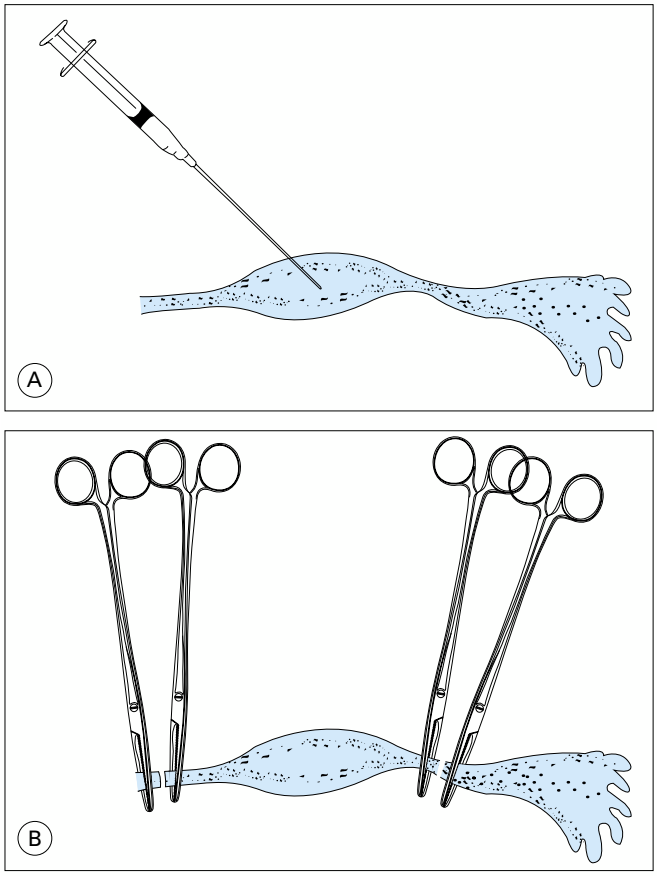


Figure 7.9 Tubal ectopic pregnancy. (A) Laparoscopic injection of methotrexate, a fetocidal substance. (B) Clamping and cutting out of the affected part of the tube at open surgery

Gestational trophoblastic disease

Causes

Chromosomal changes in the fertilised oocyte lead to degeneration of the stem blood vessels in the villi in very early pregnancy, so producing a vast overgrowth of swollen villae (vesicles) inside the uterus. This is a hydatidiform mole, and commonly no embryo is found. It is usually benign but in less than 10% of cases it develops into an invasive mole or even a gestational choriocarcinoma.

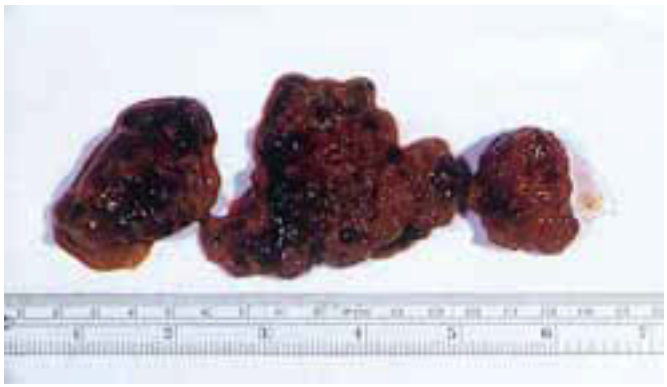
Although rare in the UK (0·6 per 1000 pregnancies), hydatidiform moles and their malignant sequelae seem to be reported more commonly in other parts of the world such as in the Pacific region (2·0 per 1000 pregnancies).

Presentation

A woman with a mole will bleed, sometimes heavily, after 6–8 weeks of gestation. She is often unwell with signs of anaemia and excessive vomiting. Proteinuric hypertension can occur as early as eight weeks. After 12 weeks of gestation the uterus often feels much bigger than expected for dates but no fetal parts can be felt or fetal heart heard. Occasionally the woman may pass vesicles through the vagina; this is diagnostic but rarely occurs.

Moles are diagnosed from this clinical presentation backed up by either an excessively high estimation of human chorionic gonadotrophin in the urine or by ultrasonography, when a characteristic picture is seen.

The differential diagnosis must include twins with a threatened miscarriage, but ultrasonography, which should be readily available to most general practitioners, gives the answer immediately.



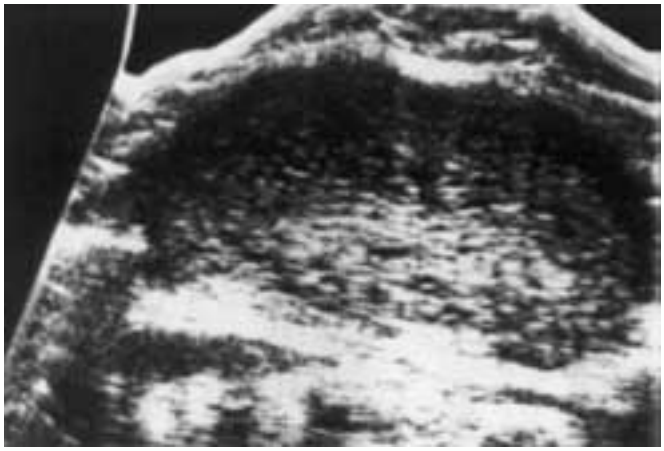


Figure 7.10 Above: Hydatidiform mole. The bunch of vesicles rapidly expands the uterine cavity. Below: A hydatidiform mole may be diagnosed readily on ultrasonography, the sound waves being reflected off the vesicles to give a picture of soap bubble foam. With early ultrasound equipment, however, hydatidiform moles looked like a snowstorm and so this term came into use. Do not expect snow if you use a B scan machine, the now commonly used apparatus

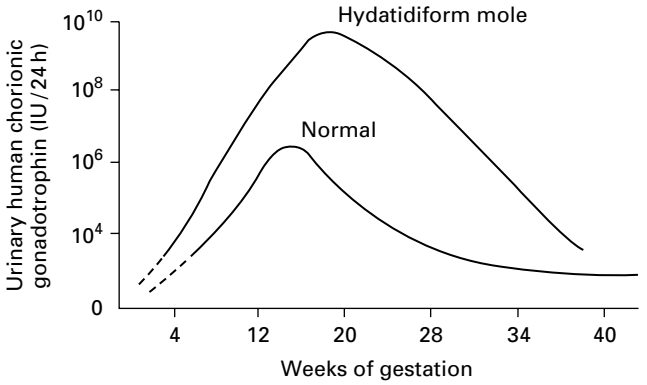


Figure 7.11 Human chorionic gonadotrophin values are much higher in women with a hydatidiform mole than in women with a normal pregnancy (note the log scale on the *y*-axis)

Management

Once diagnosed a mole should be evacuated quickly. The woman should be admitted to hospital and a suction curettage performed under anaesthesia with the protection of an oxytocin drip. All tissue is sent to the laboratory for examination of its neoplastic potential.

After an evacuation all women should be registered for follow up at one of the supraregional trophoblast disease centres, where human chorionic gonadotrophin concentrations in urine or blood can be measured. If these are high at six weeks chemotherapy is recommended to prevent subsequent malignancy.

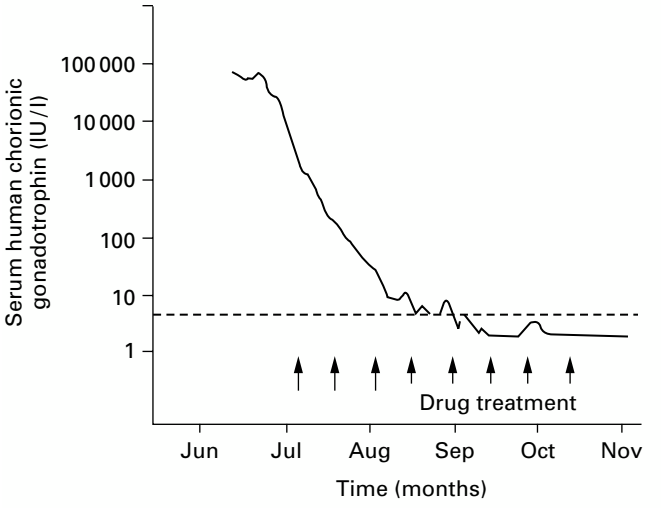


Figure 7.12 After evacuation of a hydatidiform mole the human chorionic gonadotrophin concentration remains high. Methotrexate and folinic acid were given on nine occasions, the treatment being associated with a reduction in hormone concentration

Other causes of vaginal bleeding

Bleeding may come from local problems in the vagina or cervix.

- Cervical ectropion is common in pregnancy; bleeding is not profuse.
- Vaginal or cervical infections can cause mild bleeding.
- Adenomas and polyps of the cervix become more pronounced during pregnancy. They may bleed on stimulation.
- Carcinoma of the cervix is rare but important in women of childbearing age. It may cause bleeding on stimulation and examination with a speculum reveals the cause. If there is any doubt a biopsy must be performed under anaesthesia even when a woman is pregnant.
- A general maternal disease such as blood dysplasia, von Willebrand's disease, or leukaemia may cause symptoms in rare cases.

Lesions of the cervix or vagina may cause bleeding in early pregnancy.

References

1 Westergaard J, Teisner B, Sinosich M *et al.* Ultrasound and biochemical tests in the prediction of early pregnancy failure. *Br J Obstet Gynaec* 1985;**92**:77–83.
2 Cacciatore B, Stenman U, Yostalo P. Diagnosis of ectopic pregnancy. *Br J Obstet Gynaec* 1990;**97**:904–8.

Recommended reading

- Aukum W. Diagnosing suspected ectopic pregnancy. *Br Med J* 2000;**321**:1235–60.
- Hejenius P. Interventions for tubal ectopic pregnancy. Cochrane Database of Systematic Reviews. Oxford: Update Software, 2000.
- Kulteh W. Recurrent pregnancy loss—an update. *Curr Opin Obstet Gynaec* 1999;**11**:904–8.
- Rai R, Regan L. Obstetric complications of antiphospholipid antibodies syndrome. *Curr Opin Obstet Gynaec* 1997;**9**:387–390.
- RCOG. *Management of gestational trophoblast disease.* Guidelines no. 18. London: RCOG, 1999.

8 Antenatal medical and surgical problems

Pregnant women are usually young and fit. They rarely have chronic medical conditions but when they do, those in charge of antenatal care need to consider how the disease might affect pregnancy and how pregnancy might affect the disease.

Box 8.1 Problem diseases in pregnancy

- Heart disease
- Diabetes
- Thyroid disease
- Epilepsy
- Jaundice
- Anaemia
- Haemoglobinopathies
- Urinary tract infection
- HIV infection
- Psychiatric changes and diseases

Heart disease

Most heart disease in women of childbearing age is rheumatic in origin despite the recent great reduction in the prevalence of rheumatic fever. Better living conditions in the UK and the more prompt treatment of streptococcal sore throats with antibiotics in childhood have reduced rheumatic damage to the heart valves and myocardium. An increasing proportion of pregnant women have congenital heart lesions that have been treated previously.

Pregnancy puts an increased load on the cardiovascular system. More blood has to be circulated so that cardiac output increases by up to 40% by mid-pregnancy, staying steady until labour, when it increases further. This increased cardiac work cannot be done as effectively by a damaged heart; if the heart is compromised a woman would be wise to avoid other increased loads that might precipitate cardiac failure. The most frequently encountered are:

- Household work
- Paid work outside the home
- Care of other family members
- Pre-eclampsia
- Anaemia
- Recrudescence of rheumatic fever
- Respiratory infection
- Urinary infection
- Bacterial endocarditis

Care should be taken just after delivery: with the uterine retraction up to a litre of blood can be swiftly shunted from the uterine veins into the general venous system.

Rheumatic heart disease

The commonest single cardiac lesion found in women of this age group is rheumatic mitral stenosis, sometimes accompanied by the after effects of rheumatic myocarditis. The commonest complication of overload is pulmonary oedema in late pregnancy or immediately after delivery. Right-sided cardiac failure may occur but is less common.

Cardiomyopathy of pregnancy occurs mostly post partum but occasionally in late pregnancy. There is no obvious predisposing cause; the heart is greatly distorted, leading to right-sided cardiac failure.

Congenital lesions

The most serious of the congenital lesions in pregnancy are those accompanied by shunts.

- Women with Eisenmenger's syndrome do particularly badly in pregnancy, especially those with severe pulmonary hypertension, which leads to a right to left shunt.
- Tetralogy of Fallot has a lower risk of cardiac failure because there is less resistance at the pulmonary valve regulating right ventricular outflow.
- Artificial heart valves are now present in an increasing number of women who become pregnant. Commonly they are man-made replacements of the mitral or aortic valve; affected women continue anticoagulant treatment with warfarin despite the theoretical risk of teratogenesis in early

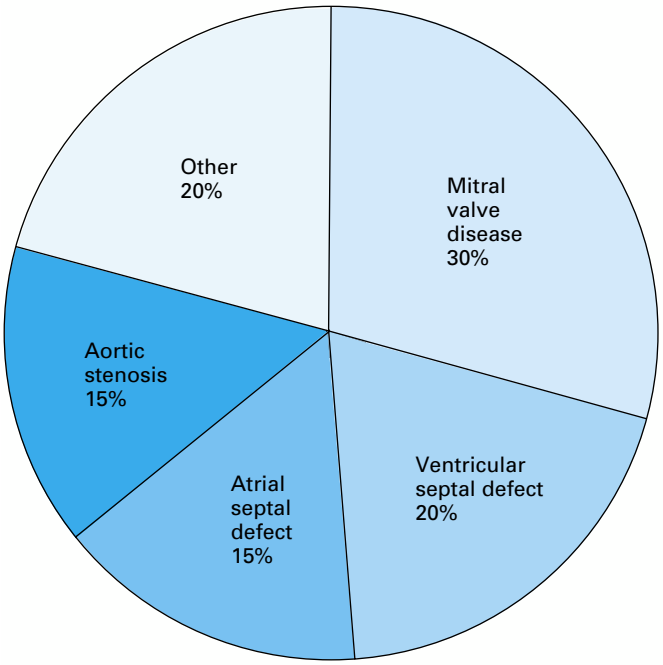


Figure 8.1 Main structural causes of heart disease in pregnancy. Other causes of heart disease include thyrotoxicosis and coronary artery disease

Table 8.1 Modified New York Heart Association's classification of exercise tolerance

	Symptoms of cardiac insufficiency	Limitation of activities
I	None	None
II	Only after exercise	With moderate exercise
III	After any activity	With ordinary activities
IV	At rest	Unable to perform any physical activities

pregnancy and fetal bleeding later. It is still widely used and may be replaced two or three weeks before the expected date of delivery by heparin.

Management

Most women with heart disease who are of childbearing age are known to their family practitioner. He or she should ensure that they go for antenatal care at a centre where a cardiologist works alongside an obstetrician, ideally at a combined cardiac antenatal clinic if there are enough cases.

Early assessment should be made of the severity of the disease, paying attention to the features that may worsen the prognosis: the woman's age, the severity of the lesion, the type of lesion, and the degree of decompensation (exercise tolerance). Rest should be encouraged during pregnancy and extra physical loads avoided. Labour should be booked at a consultant unit with an interested cardiologist involved. The ward may need the extra drugs and equipment to be available if a woman with a heart condition is admitted. Delivery should be planned at a unit with ready access to a cardiac centre and availability of cardiologists and cardiac anaesthetists.

Care should be taken to avoid the development of acute bacterial endocarditis by ensuring that the woman is given antibiotics when she has any infection or is at potential risk of developing an infection—for example, at a tooth extraction or labour. This precaution is more important for congenital lesions of the heart than for rheumatic lesions.

The prognosis for a woman with heart disease in pregnancy is now greatly improved. It used inevitably to be associated with deterioration of the heart condition, but now, with proper care, this is not so.

Diabetes

Diabetes is a metabolic disease found in about 1% of women of childbearing age. In addition, another 1–2% of women will develop gestational diabetes during the course of their pregnancy; the incidence is higher in older than younger women. Glycosuria (checked by dipstick testing) is even more common than this, occurring at some time in pregnancy in up to 15% of women and is no longer a screening test for diabetes in pregnancy. Instead finger-prick or venous blood samples should be checked for blood sugar levels.

Established insulin dependent diabetes

Four fifths of women with diabetes are known to the practitioner before they become pregnant. All diabetic women of reproductive age should be using effective contraception and be encouraged to attend a prepregnancy clinic so that pregnancy is planned. Good control of diabetes before and in early pregnancy reduces the incidence of congenital anomalies and miscarriage.

Antenatal care is best performed by an obstetrician and a diabetic physician at a combined diabetic antenatal clinic. The general practitioner must be kept well informed of changes in management of the diabetes during pregnancy, because between antenatal clinic visits the woman may depend on her family practitioner for continuity of care. Detailed ultrasonography to exclude congenital abnormalities and to monitor growth is vital.

Pregnancy makes the control of diabetes more difficult; close monitoring is the key to a successful outcome. Women are encouraged to eat enough carbohydrate to satisfy them without

Box 8.2 Drugs which may be needed when a woman with severe heart disease is admitted in pregnancy or labour

- Oxygen
- Digoxin
- Frusemide
- Aminophylline

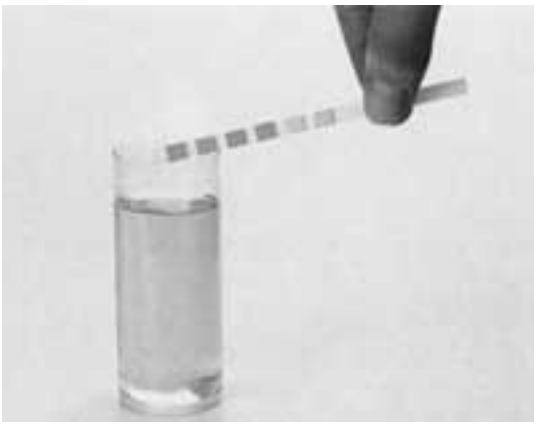


Figure 8.2 Dipstick testing of urine

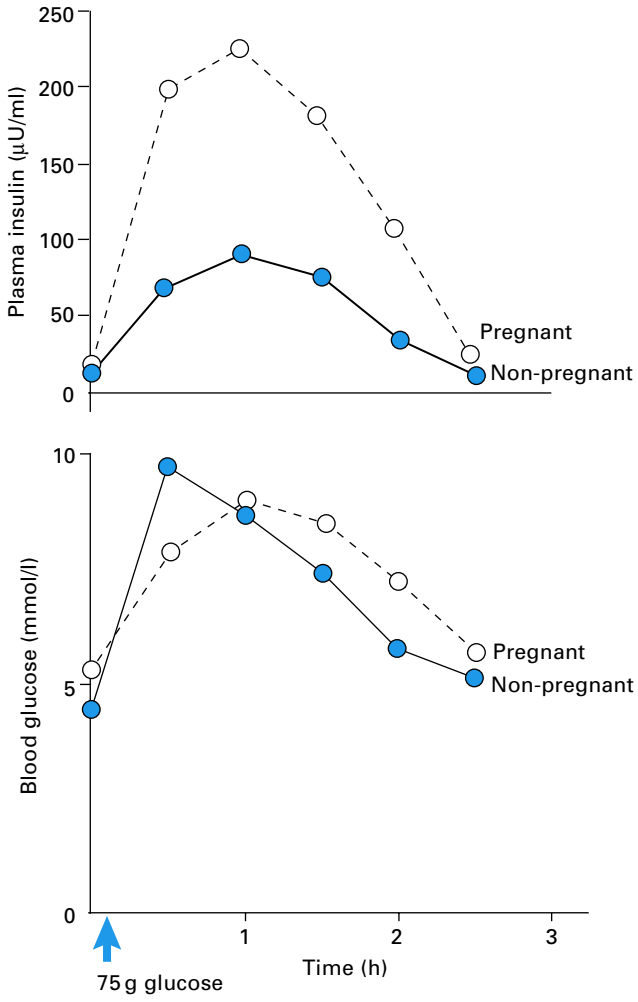


Figure 8.3 Plasma insulin and blood glucose response to oral glucose (75 g) in pregnant and non-pregnant women

restriction and should take regular snacks between meals. Most women who have attended a prepregnancy clinic will have already been converted to a basal bolus insulin regime. This consists of three short-acting doses during the day and one long-acting insulin dose at night. This regime enables good glucose control to be achieved and is started in early pregnancy, if not before.

Glucose concentrations in blood are measured by the woman as frequently as four times a day with her own glucose meter at home. Virtually all diabetic women require an increase in their insulin dosage during pregnancy. Frequent clinic visits are necessary to facilitate this and the careful monitoring of the fetus.

Diabetes controlled by oral hypoglycaemia agents

Oral hypoglycaemic agents are not advised in pregnancy and conversion to the basal insulin regime is best done before conception, if possible. Such women are then monitored in the same way as women with established insulin dependent diabetes.

Gestational diabetes

Gestational diabetes is diagnosed when a woman develops abnormal glucose tolerance for the first time in pregnancy; a small number of such women will remain diabetic after the pregnancy. Currently, many hospitals will perform a random blood glucose test during the antenatal course, interpreting the result in relation to the timing of the last meal. Women with high values will then have a glucose tolerance test or have blood glucose concentrations measured serially (preprandial and postprandial tests three times a day) to determine whether they are glucose intolerant.

Women with gestational diabetes do not have an increased rate of babies with congenital abnormalities but the babies are at risk of being large. There is no consensus on treatment, which ranges from controlling dietary intake to insulin treatment and dietary control. Such women usually have labour induced at term and are at risk of having long labours and babies with shoulder dystocia.

After delivery insulin should be stopped; all affected women should have a glucose tolerance test at six weeks. About 40–60% of such women will develop non-insulin dependent diabetes (type II) in later life but this proportion rises to 70% among those who are obese.

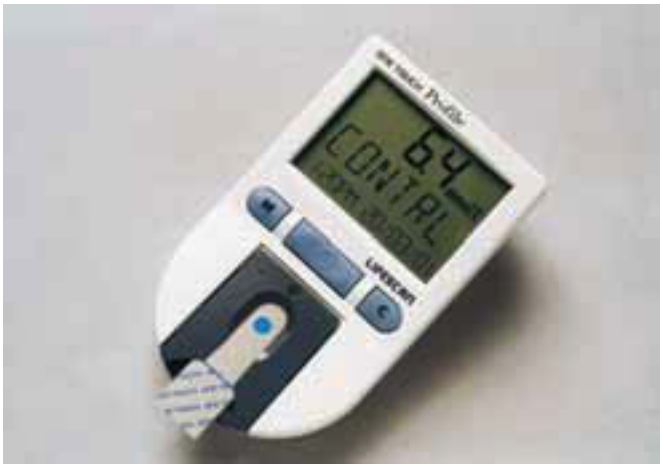


Figure 8.4 Blood glucose concentration meter for home use

Box 8.3 Vaginal delivery in diabetic mothers

- *Good prognostic features*
  - Primigravida < 30 years
  - Multigravida with good obstetric history
  - Estimated fetal weight < 3500 g
  - Well engaged cephalic presentation
  - Stable diabetic control
- *Bad prognostic features*
  - Primigravida >30 years
  - Multigravida with poor obstetric history
  - Large fetus (>3500 g)
  - Non-engageable head or breech presentation
  - Unstable diabetes

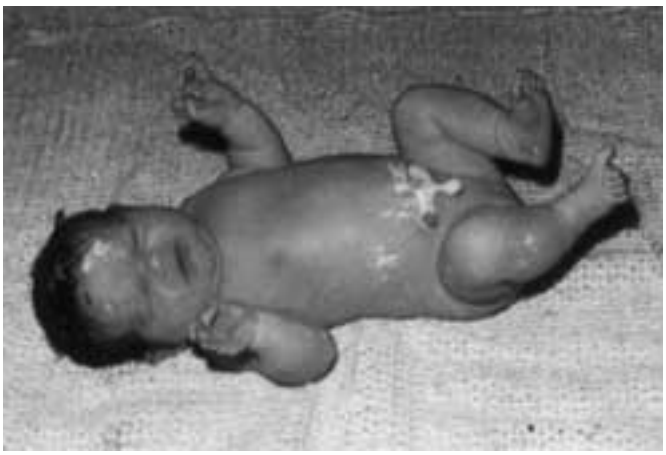


Figure 8.5 A typically large baby born to a diabetic mother

Thyroid disease

Hyperthyroidism

Women who are already hyperthyroid are usually receiving treatment, which may have to be continued throughout pregnancy. The most commonly used drugs are carbimazole and propyl-thiouracil; the former is in more common use but the latter is often chosen in pregnancy as it is less often associated with congenital abnormalities of the scalp. The minimum dose should be prescribed to alleviate any symptoms and to suppress free thyroxine concentration to the normal range. However, some of these women find that their hyperthyroidism ameliorates in the last weeks of pregnancy. In such cases withdrawal of antithyroid drugs may reduce the severity of any fetal goitre.

These women should be tested for the presence of IgG thyroid antibodies (long-acting thyroid stimulator and thyroid receptor antibodies) as these cross the placenta and cause neonatal thyrotoxicosis when present in high titres. Thyroid

Table 8.2 Effect of thyrotoxicosis and pregnancy on some thyroid tests

	Thyrotoxicosis	Pregnancy
Tri-iodothyronine:		
free	Increased	No change
protein bound	Increased	Increased
Thyroxine:		
free	Increased	No change
protein bound	Increased	Increased
Thyroxine binding globulin	No change	Increased

crises (storm crises) are now rare in pregnancy and the immediate puerperium. They are best treated with iodine, which works quicker than β blockade and carbimazole. Operation on the thyroid is rarely indicated in pregnancy but is safest in the middle trimester.

Hypothyroidism

Hypothyroid women are commonly anovular. If they are receiving adequate replacement treatment, however, they ovulate as normal. Such treatment should be continued and may need to be increased during pregnancy.

Epilepsy

An epileptic woman will often consult before becoming pregnant as she may have heard of the potential hazards of antiepileptic drugs. Most antiepileptic drugs have teratogenic properties to a varying extent, but it must be emphasised that epileptic women have an inbuilt increased risk of having babies with malformations even without treatment. This risk should be carefully balanced against the risks to the embryo if the woman has a series of convulsions when anticonvulsant treatment is withdrawn in early pregnancy.

Generally, the woman may stop or modify treatment after full consultation when she has not had a recent fit. However, if the epilepsy is well controlled, there is little point in changing antiepileptics in pregnancy. If she needs treatment the same dose must be continued; phenytoin treatment may be associated with a slightly lower risk of fetal neural tube defects and might be substituted instead of valproate or carbamazepine.

Seizure frequency seems to be the same in pregnancy as outside pregnancy for most epileptic women; if the rate of fitting worsens, blood concentrations of all anticonvulsants should be checked as overdose as well as underdose may be responsible for loss of seizure control.

Prophylactic folic acid (5 mg/day) should be given before and during pregnancy as folate absorption is changed by the antiepileptic drugs. Vitamin K should be given to all the newborn infants of such mothers for similar reasons.

Status epilepticus is unusual in a pregnant woman unless she is known to be a severe epileptic. Diazepam is the best drug to use.

Table 8.3 Therapeutic concentrations of anticonvulsants in blood

	mg/l
Phenytoin	10–12
Phenobarbitone	15–40
Carbamazepine	4–12
Primidone	5–12
Ethosuximide	5–12
Valproate	4–100

Box 8.4 Potential effects of epilepsy on the fetus

- Increased risk of epilepsy in the baby:
  - if mother alone affected 4%
  - if both parents affected 15%
  - if another child affected 10%
- Increased risk of congenital abnormalities:
  - if either parent affected
  - if mother takes more than one anti-epileptic drug
- Isolated maternal fits do not usually affect fetus. Status epilepticus does

For most epileptic women the frequency of seizures is not affected by pregnancy.

Box 8.5 Transmission of HIV

Transmission of HIV from mother to fetus may be:

- across the placenta in pregnancy
- due to exposure to blood during vaginal delivery
- by breast feeding

The most frequent cause is vaginal transmission which can be reduced by bypassing the vagina (i.e. CS)

HIV infection

The human immune suppression retrovirus (HIV) attacks CD 4 lymphocytes leading to their suppression and hence increasing susceptibility to infection. The acquired immune deficiency syndrome (AIDS) is the end stage of such a process and develops some years after the initial HIV infection. Transplacental transmission of the virus antenatally from mother to fetus or breast feeding after delivery can lead to an infected baby.

HIV infection is found more commonly in the big towns such as London where 1 in 600 antenatal attenders is HIV positive. In the country generally it is nearer 1 in 10 000. It is probable that pregnancy does not increase the progression of the disease in the mother.

The baby will be infected in 15–20% of cases.[1] There is a possibility that elective caesarean section would reduce this risk by eliminating fetal exposure to the secretions of the genital tract. The European Study, considering 1000 mother/baby pairs, considered that caesarean section halved the risk of infection[1] although subsequent analyses have shown only a

20% reduction due to caesarean section.[2] At present the best prospect of management is to prevent women becoming HIV infected. In pregnancy, the established infected women should be detected by antenatal screening for HIV with proper counselling and offered treatment with anti-retroviral agents, the current product being zidovudine.

It is worth diagnosing HIV in pregnancy for now there is a reasonable treatment which reduces the rate of transmission of HIV to the fetus from 25% in a control group compared with 7% in a zidovudine group.

All infants of HIV positive mothers should be commenced on zidovudine for six weeks and tested at one month and four months for antibodies. Breast feeding is contraindicated in the UK but may be the only method of contraception available in developing countries; the extra risks of HIV transmission should be weighed against further unwanted pregnancies. Folate supplements are especially recommended for the prepregnancy period and the first trimester for all women with HIV infection, to prevent neural tube defects. Infected women who have a high viral load or who have not had any antenatal treatment may be better delivered by caesarean section to reduce the transmission to infants.

Jaundice

The commonest causes of jaundice in pregnancy are the various forms of hepatitis and drugs that affect the liver. Gall stones and severe pre-eclampsia may be responsible, but in the UK gall stones are rare in the age group concerned. Cholestasis in the last trimester may occur spontaneously or follow the use of steroids; fatty degeneration of the liver in the last weeks of pregnancy is very rare but can lead to liver failure as can severe autoimmune disease.

The results of the conventional liver function tests are not as helpful during pregnancy, and the early participation of liver experts in the care of a woman with jaundice during pregnancy is essential.

Box 8.6 Some causes of jaundice in pregnancy

- *Pregnancy associated*
  - Cholestasis
  - Acute fatty liver of pregnancy
  - Disseminated intravascular coagulopathy
  - Severe pre-eclampsia and HELLP syndrome
  - Excessive vomiting (hyperemesis)
  - Severe septicaemia in late pregnancy
- *Unrelated to pregnancy*
  - Viral hepatitis
  - Drugs
    - chlorpromazine
    - tetracycline
    - steroids
  - Chronic liver disease
  - Gall stones
  - Chronic haemolysis

Table 8.4 Normal haematological values in pregnancy

	Range
Total blood volume (ml)	4000–6000
Red cell volume (ml)	1500–1800
Red cell count (10^{12}/l)	4–5
White cell count (10^9/l)	10–15
Haemoglobin (g/dl)	11.0–13.5
Erythrocyte sedimentation rate (mm/hr)	10–60
Mean corpuscular volume (μm^3)	80–95
Mean corpuscular haemoglobin (pg)	27–32
Serum iron (μmol/l)	11–25
Total iron binding capacity (μmol/l)	40–70
Serum ferritin (μg/l)	10–200
Serum folate (μg/l)	6–9

Anaemia

In pregnancy, anaemia might be due to:

- lack of haemoglobin from a low intake of iron (microcytic anaemia) or of folate (megaloblastic anaemia)
- haemorrhagic anaemia following chronic blood loss
- haemolytic anaemia in those with abnormalities of the genes of the haemoglobin molecule or of the envelope of the red cell.

Box 8.7 Indices of iron deficiency anaemia

- *Blood film: red cells*
  - normal size or microcytic
  - hypochromic
  - anisocytosis
  - poikilocytosis
- *Haematological values*
  - haemolobin ↓
  - mean corpuscular volume ↓
  - mean corpuscular haemoglobin ↓
  - serum iron ↓
  - serum ferritin ↓

Iron deficiency anaemia

This is the most common form of anaemia in the UK. The daily need for iron rises from 2 mg per day to 4 mg in pregnancy. This can be provided by improved diet or more practically by taking regular prophylactic tablets containing 60 mg per day of elemental iron. This supplement is given to most pregnant women in the UK. If they cannot take iron tablets, a liquid preparation or intramuscular iron should be provided.

Folate deficiency anaemia

This is less common than iron deficiency anaemia in the UK. Folate needs are increased because of increased maternal demands from growth of the uterus and breasts as well as the increased tissues laid down in the growing fetus.

The woman may produce symptoms of anaemia with breathlessness and pallor; the blood film may show a low

haemoglobin concentration, maybe with macrocytes. The latter may be missing and a bone marrow sample from the iliac crest may be required to show megaloblastic changes.

The condition is treated by oral folate; the diet can be improved and should contain dark green leaf vegetables and yeast extracts. However, in Britain, usually folate is given prophylactically, often combined with iron, to prevent folate deficiency. Those with twins and women taking antibiotics require extra folate. These needs are in addition to the folate used before pregnancy and in early gestation to prevent the formation of central nervous system abnormalities.

Haemorrhagic anaemia

Haemorrhagic anaemia is rare in the UK among women of childbearing age, but chronic bleeding from peptic ulceration, aspirin ingestion, or piles may occur. In other countries tapeworms or hookworms may cause a constant chronic blood loss. Treatment is that of the causative condition.

Haemolytic anaemia

Hereditary haemolytic anaemia is also a rare disease in the white population of the United Kingdom, but other races may show a variety of haemolytic anaemias.

Haemoglobinopathies

Women liable to haemoglobinopathies and their antecedents usually come from Mediterranean countries or Asia and are often known to the family doctor beforehand. All such women should have a blood film examined and their blood checked by electrophoresis at the booking clinic. If they are found to be carriers, their partner's blood should be checked. If they too are carriers, fetal diagnosis is available from early chorionic villus sampling and from fetal blood sampling in later pregnancy. Such women are best managed at special combined antenatal-haematological units and should be sent to such hospitals early in pregnancy so that plans can be made to cover all eventualities. If not, as luck would have it, the crisis will always come on Saturday night at 11.30 pm.

Sickle cell disease

Most women in the UK have haemoglobin A. Defective genes can alter the amino acid sequence of haemoglobin, which may produce symptoms. Haemoglobin S originated in the Middle East but is now found in Africa and the West Indies. Those with haemoglobin C come from West Africa. The partner's blood should be tested and antenatal diagnosis of the fetus is available by direct gene probe from a chorionic villus sample if both partners carry the trait.

In pregnancy a woman with sickle cell disease is at high risk of complications; she deserves special antenatal supervision. Even in experienced hands the perinatal mortality rate can be four times that in a normal population and maternal mortality is also greatly increased. In extreme cases sickling produces crises, leading to sudden pain in the bones, chest, or abdomen after small vessel infarction. Rates of severe pre-eclampsia are higher, as are the incidences of chest and urinary infections. Intrauterine growth retardation and fetal death occur because of placental infarction.

If a crisis occurs then both haemoglobin concentration and red cell volume should be checked every few hours. Hospital treatment with intravenous hydration, partial exchange transfusion or packed red cell transfusions, and antibiotics may be required. Women with haemoglobin concentrations below 6.0 g/dl should have exchange transfusions before elective

Table 8.5 Dose and ferrous iron content of commonly prescribed iron tablets

Iron tablets	Dose (mg)	Ferrous iron content (mg)
Ferrous sulphate (dried)	200	60
Ferrous sulphate	300	60
Ferrous fumarate	200	65
Ferrous gluconate	300	35
Ferrous succinate	100	35

Box 8.8 Indices of folate deficiency anaemia

- *Blood film*
  - red cells
    - normal size or macrocytic
    - normochromic
    - anisocytosis
    - poikilocytosis
    - sometimes nuclear material
  - white cells
    - leucopenia
    - hypersegmentation
  - platelets
    - sometimes thrombocytopenia
- *Haematological values*
  - haemoglobin ↓
  - mean corpuscular volume ↓ or =
  - mean corpuscular haemoglobin ↑
  - serum iron ↑
  - red cell folate ↓
  - marrow megaloblastosis

Box 8.9 Indices of sickle cell anaemia

- *Blood film*
  - red cells
    - polychromasia
    - sickle cells
    - Howell-Jolly bodies
  - white cells
    - leucocytosis
  - platelets
    - thrombocytosis
- *Check*
  - haemoglobin electrophoresis
  - test partner

Box 8.10 Treatment of sickle cell crisis

- Pethidine for pain
- Antibiotic only if infection also
- Oxygen
- Intravenous fluids to maintain hydration
- ? Intravenous bicarbonates for acidaemia
- ? Exchange transfusion

delivery. Babies of high risk couples should be tested and followed up if they have sickle cell disease.

Thalassaemia

In thalassaemia, the life of a red cell is shorter than the usual 120 days and so anaemia follows because there is a more rapid breakdown than production of cells. Haemoglobin concentration is low but the serum iron concentration is high.

Again, iron may not be needed if stores are adequate but many such women need extra iron as iron deficiency anaemia may accompany thalassaemia. The stress of hypoxia or acidaemia should be avoided as both increase the breakdown rate of red cells.

Box 8.11 Indices of thalassaemia

- *Blood film*
  - red cells
    - ? polychromasia
    - microcytosis
    - hypochromia
    - sometimes anisocytosis
    - sometimes poikilocytosis
    - target cells present
- *Haematological values*
  - haemoglobin ↓
  - serum iron ↓
  - mean corpuscular volume ↓
  - mean corpuscular haemoglobin ↓
- *Check*
  - haemoglobin electrophoresis
  - test partner

Urinary tract infection

Acute urinary infection occurs in about 2% of women during pregnancy. Infection of the urethra and trigone of the bladder is signalled by dysuria and increased frequency of micturition, whereas infection of the upper tract affecting the ureters or kidney produces loin pain and spikes of fever.

A midstream urine specimen should be checked for the presence of cells and bacteria (with bacterial sensitivity to antibiotics) before any treatment is started. The woman should drink much more and take a wide spectrum antibiotic such as amoxycillin until the results of the test are known. Antibiotic treatment may have to be changed according to the sensitivity results but usually amoxycillin suffices. (Alkalination of the urine may be performed, though this is unpleasant and entails taking potassium citrate mixture.)

After 7–10 days, a second midstream specimen of urine should be sent to the laboratory. If bacteria are still detected, continuous low dose antibotic prophylaxis using trimethoprim (second and third trimesters only) or amoxycillin should be considered. Cranberry juice may be useful in preventing recurrent infection.

Box 8.12 Acute urinary infection in pregnancy

- Check MSSU for organisms and sensitivity
- Use as first line drug
  - amoxycillin or
  - ampicillin or
  - cephalosporin or
  - augmentin
- Be prepared to change if sensitivity tests indicate
- Use with caution if sensitivity demands
  - sulphonamides (beware kernicterus in baby)
  - trimethoprim (beware of folic acid antagonism)
  - nitrofurantoin (because of G6PD deficiency in baby)

Asymptomatic bacteriuria

Infection may be low grade and asymptomatic. About 4% of pregnant women have evidence of bacterial infection of the urine; its significance level is arbitrarily set at more than 100 000 bacteria per ml of urine.

If all women are screened early in pregnancy and asymptomatic bacteriuria is detected it is probably wise to treat, as the risk of developing acute pyelonephritis in pregnancy is about 30%. Treatment is for five days with an antibiotic to which the bacteria are sensitive. A urine sample should be recultured 14 days later. If bacteria are still present continuous antibiotic prophylaxis should be considered.

Any woman with persistent asymptomatic bacteriuria through pregnancy should have her urinary tract checked after delivery. About 20% of this subgroup will be found to have a structural abnormality of the kidneys, ureters, or bladder.

Chronic renal disease

Most women with chronic renal disease are well known to their general practitioner and have usually been counselled by a renal physician about the risks of pregnancy and the precautions required. In brief, renal function usually improves in pregnancy, and there is no evidence that pregnancy adversely affects the long-term prognosis from the renal disease. The outlook in pregnancy is favourable if the patient is not hypertensive and does not have proteinuria before pregnancy. Pregnancy should be carefully supervised by the obstetric and renal team.

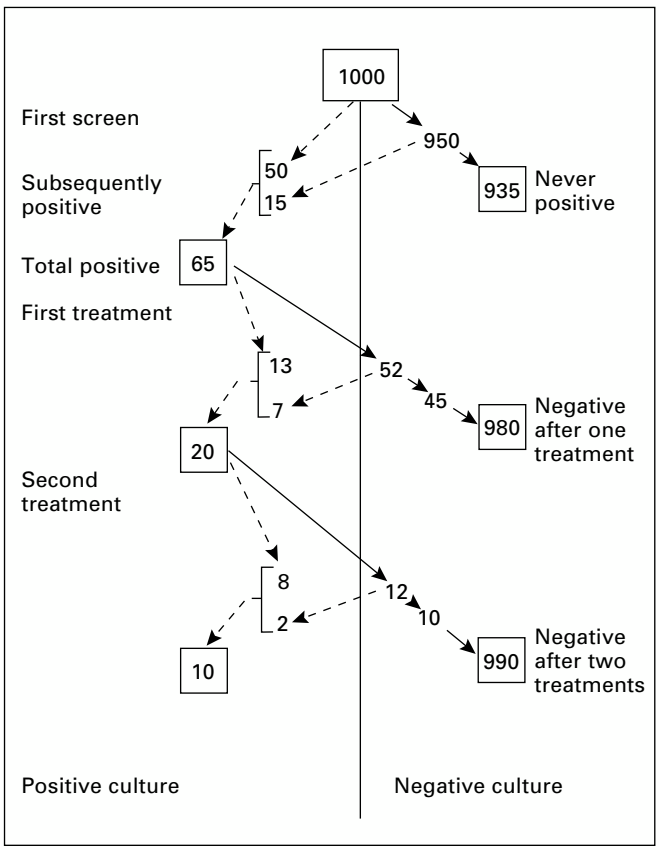


Figure 8.6 Progress of 1000 women with asymptomatic bacteriuria during pregnancy

Transplant recipients have normal fertility. There is little evidence that the commonly used immunosuppressive agents cause an excess of fetal abnormalities. Episodes of rejection are not more common in childbirth, but if they occur they usually do so in the puerperium. If the transplanted kidney is in the pelvis a caesarean section may be necessary for mechanical reasons.

Box 8.13 Considerations for pregnancy in chronic renal disease

- Type of disease
  - beware scleroderma, periarteritis nodosa
- Blood pressure
  - diastolic pressure <90 mm Hg
- Renal function
  - plasma creatinine <250 μmol/l
  - plasma urea <10 mmol/l
  - no proteinuria
- Review essential drug treatment

Abdominal pain in early pregnancy

From the uterus

Miscarriage

One of the commonest causes of pain in early pregnancy is spontaneous miscarriage. This subject is dealt with in Chapter 7.

Retroverted uterus

Retroversion is a common position for a normal uterus. In pregnancy the uterus expands into the abdomen. If adhesions are present, however, this cannot occur; by 10–12 weeks the enlarging uterus fills the pelvis and pain is associated with retention of urine. The urethra is stretched by the uterine bulk and the bladder pushed to the abdomen so that urine cannot pass. These findings can be confirmed by ultrasonography.

Management includes draining the urine with an indwelling catheter. The cure eventually comes when the uterus grows into the general abdominal cavity by anterior sacculation, so relieving the urethral stretch.

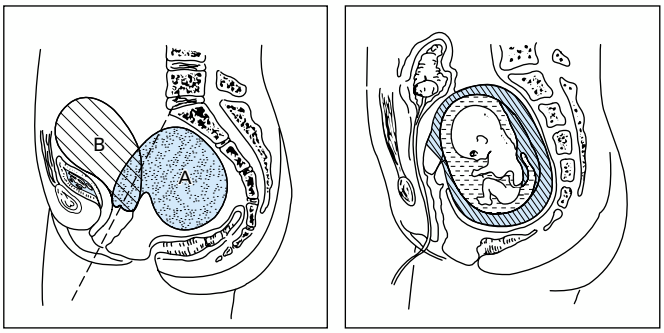


Figure 8.7 Left: Retroverted uterus (A) and anteverted uterus (B) in early pregnancy. Right: Management of impacted retroverted uterus during pregnancy (catheterisation)

Fibroids

Fibroids are found in older pregnant women (those aged 30–40), particularly among Afro-Caribbean women. In pregnancy fibroids can undergo torsion if they are subserous; this is more common in the puerperium. Red degeneration is commonest at 12–18 weeks of pregnancy but can occur throughout, with resulting necrobiosis in the fibroid. The woman presents with tenderness over the mass accompanied by vomiting and mild fever.

Red degeneration is self limiting; if the diagnosis is firm, management is bedrest with analgesia and intravenous correction of any dehydration. Ultrasound may help to confirm the presence of fibroids, although necrobiosis may not show clearly. In truly doubtful cases, as in a low-right sided fibroid that mimics appendicitis, a laparotomy should be performed to exclude surgically correctable conditions. If red degeneration is diagnosed the surgeon would do well not to remove the fibroid at this stage but to close the abdomen and continue conservative management.

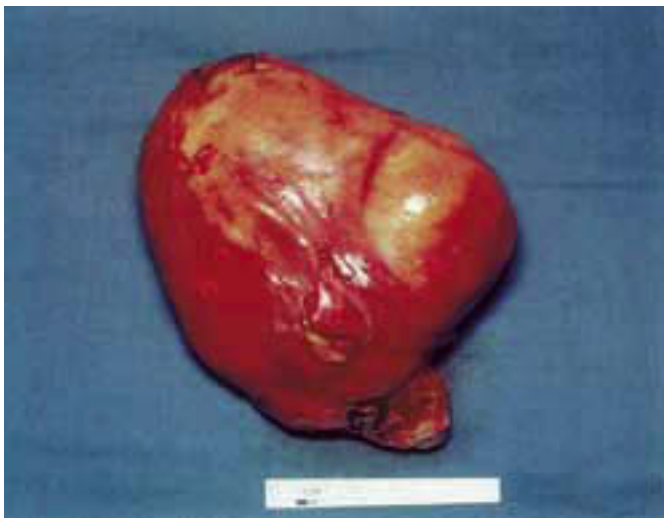


Figure 8.8 Fibroids are benign quiescent tumours consisting of whorls of fibres and few cells

From the fallopian tube

Ectopic pregnancy

Unruptured ectopic pregnancy causes chronic symptoms and needs to be managed in hospital whereas ruptured ectopic pregnancy produces acute symptoms and collapse and needs urgent hospital management. The condition is dealt with in Chapter 7.

If you do not think of an ectopic pregnancy you will not diagnose one. Always consider unruptured ectopic pregnancy in any young woman having sexual intercourse who has lower abdominal pain.

Torsion

Torsion is uncommon and occurs mainly in younger women during early pregnancy when a long tube may twist on its pedicle accompanied by torsion of the ovary, especially if the latter has a cyst in it.

The woman has non-specific hypogastric pain and a constant area of tenderness suprapubically on the lateral edge

Box 8.14 Fibroids in pregnancy

- Usually increase in size but become hypovascular
- Necrobiosis (red degeneration) is painful but treat conservatively
- Torsion of subserous fibroid is acutely painful and needs surgical removal

of the rectus abdominis muscle. Ultrasound does not help but diagnostic laparoscopy in early pregnancy is useful. A laparotomy is required; if the lateral end of the fallopian tube is non-viable it must be resected; in rare cases the ovary is also ischaemic and requires removal.

From the pelvic ligaments

Round ligament

These stretch as the uterus rises in the abdomen and pulls on the uterine round ligaments like an inflating hot air balloon tugging its guyropes. Usually the ligaments stretch easily, but if the pull is too strong small haematomas occur. This commonly starts at 16–20 weeks' gestation.

On examination tenderness is localized over the round ligament and often radiates down to the pubic tubercle alongside the symphysis pubis.

Treatment is bedrest, analgesia, and local warmth.

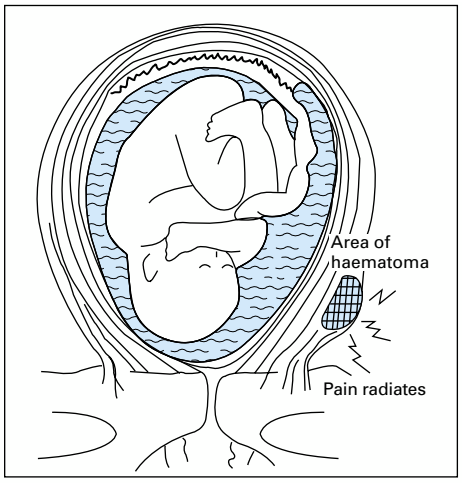


Figure 8.9 Haematoma of round ligament

From the ovary

Ovarian tumours

In early pregnancy an ovarian cystic tumour may rupture to release the contents of the cyst, irritating the parietal peritoneum. Bleeding may occur into a corpus luteal cyst. An ultrasound scan may confirm the diagnosis, and a laparotomy is indicated if the clinical situation does not settle. At laparotomy, only that part of the ovary containing the cyst should be removed. If it is a luteal cyst, conservation is necessary as the corpus luteum is probably the major source of progesterone in the first trimester and some of this metabolism continues into later gestation.

Box 8.15 Ovarian pain in pregnancy

- Tortion of pedicle of ovary with lateral end of tube
- Stretch of capsule of a cyst
- Bleeding into cavity of cyst (corpus luteum)
- Rupture of cyst with release of contents

Extrapelvic causes

Vomiting

Though many women who vomit in pregnancy have little upset, vomiting or retching may be sufficiently severe to cause muscle ache from stretch. The upper abdominal wall is tender and no specific masses can be felt. If a woman is vomiting this much it is probably wise to admit her to hospital for intravenous fluids, antiemetic treatment, and sedation to allow her intestinal tract some peace. The pain usually settles down as the vomiting decreases.

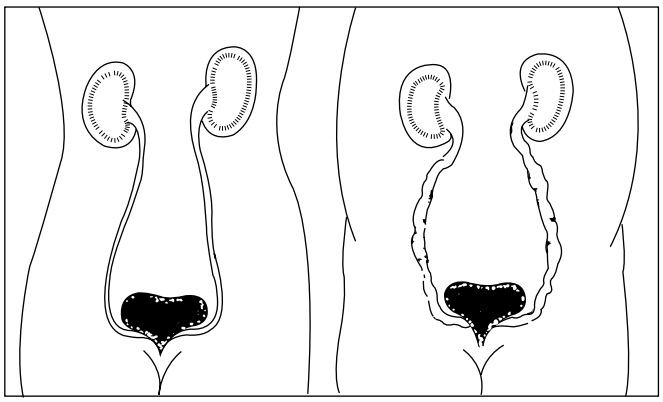


Figure 8.10 During pregnancy the ureters lengthen and become more tortuous and dilated

Pyelonephritis

Stasis in the urinary tract associated with ascending urinary infection often follows dilatation of the ureters (due to raised progesterone concentrations) and the pressure of the increasing uterus on the bladder. It is most likely in mid-pregnancy, when the woman presents with vomiting, symptoms of fever, and low hypogastric or loin pain.

Appendicitis

Appendicitis and pregnancy both occur in young women and therefore may occur concurrently by chance. The incidence of appendicitis in pregnancy is not increased but its diagnosis may be more difficult. For this reason and because of a reluctance to operate, appendicitis used to have a high mortality and morbidity in pregnancy.

As it grows, the uterus displaces the caecum from the right iliac fossa upwards and sideways, so the inflamed appendix may present with symptoms and signs in unexpected places. No longer tucked into the right iliac fossa, the appendix is now in the general abdomen and is less easy to wall off by omentum and gut when it becomes inflamed; generalized peritonitis is commoner in pregnant than non-pregnant women.

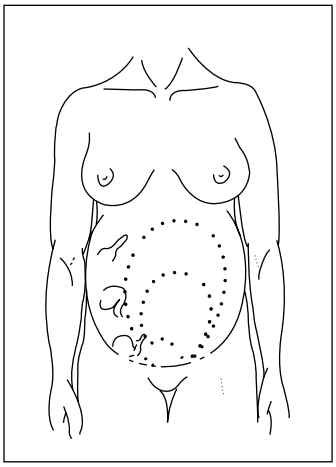


Figure 8.11 The site of the appendix changes as pregnancy advances

A history may elicit the characteristic pain shift, although it is not always localised to the right iliac fossa. Nausea and anorexia occur, sometimes confused by the symptoms of pregnancy. The tenderness over the appendix will shift higher as pregnancy continues. The treatment is operation, the incision being placed over the point of maximum tenderness marked by the surgeon before anaesthesia. Occasionally the results of a rectal examination can be falsely reassuring if the appendix has migrated from the area reached by an examining finger.

The previous reluctance to operate must be overcome; anyone suspected of having appendicitis in pregnancy should have a laparotomy by an experienced surgeon. Even in late pregnancy, caesarean section is not necessary at the same time unless the woman is in labour; women can have a normal vaginal delivery within a few days of an appendicectomy.

Other causes

Cholecystitis is commoner among women who live in or originate from countries whose residents characteristically have high cholesterol diets such as Australia and New Zealand. The pain is usually upper right abdominal with tenderness centred on the eighth or ninth rib tip. Treatment in the absence of jaundice is conservative with antibiotics or removal, depending on the surgical need.

Volvulus of large bowel can occur in pregnancy, though it presents more characteristically in the puerperium.

Small bowel colic may follow an attack of gastroenteritis.

Urinary lithiasis occurs in the same frequency in pregnancy as in non-pregnant women.

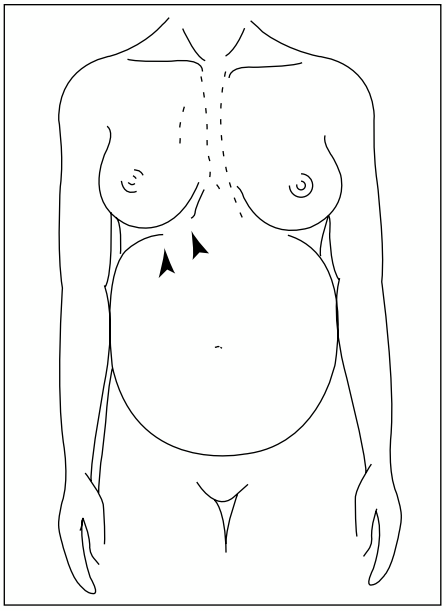


Figure 8.12 Pain in cholecystitis

Abdominal pain in late pregnancy

From the uterus

Uterine contractions

All pregnancies end in labour, which may occur well before term. Premature labour can present with abdominal pain, taking the woman and sometimes her general practitioner by surprise. Usually the pain is intermittent and recurrent and the uterus can be felt contracting coincidentally with the pain. There may be a loss of mucus or a little blood from the vagina, on vaginal examination the cervix is soft, thin, taken up, and sometimes dilated. When labour is very preterm (26–32 weeks) the woman should be transferred to a hospital with an expert neonatal unit rather than necessarily to the one where she has booked (see Chapter 12).

Placental abruption

Separation of the placenta from its bed before the third stage of labour is painful and results in shock (see Chapter 10).

Extraperitoneal causes

Pregnancy-induced hypertension

In severe fulminating pregnancy-induced hypertension a woman may complain of epigastric pain associated with vomiting. She will probably have raised blood pressure and proteinuria with oedema and be known to be hypertensive. There may also be visual symptoms (outlined in Chapter 9).

Rectus haematoma

Very rarely the rectus muscle may dehisce and the inferior epigastric veins behind the muscle rupture. As the anterior

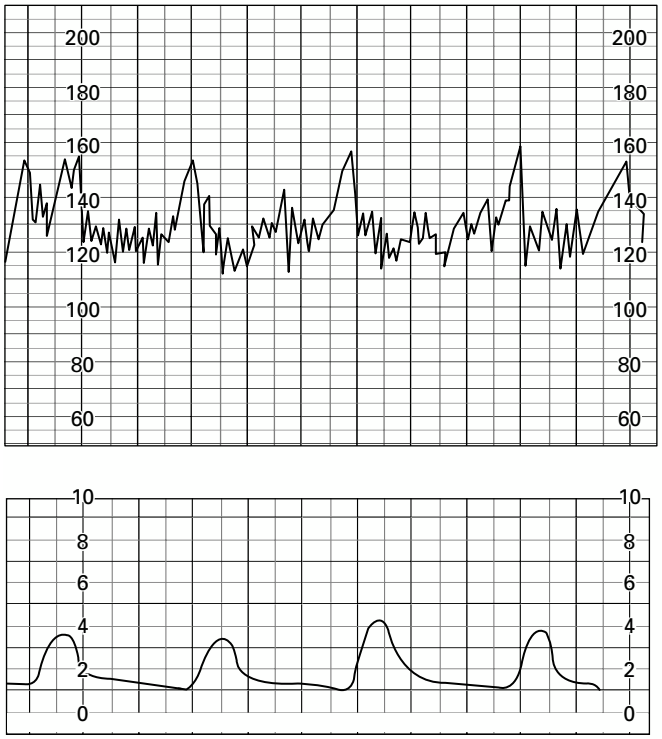


Figure 8.13 A cardiotocograph in early labour showing the fetal heart rate (above) and the regular uterine contractions every three minutes (below)

abdominal wall is greatly overstretched by the uterus, a fit of sneezing could cause this. Pain is severe and usually localised to one segment of the muscle. Blood loss is slight with the haematoma but increases if the veins rupture. Rectus haematoma is diagnosed from the fact that pain and tenderness worsen when the woman contracts the rectus muscles by raising her head. Ultrasound is helpful.

If the diagnosis is firm, management is conservative, but in doubtful cases a laparotomy should be performed, and haematoma behind the rectus muscle confirms the diagnosis.

Pelvic arthropathy

Relaxation of the ligaments guarding the pelvic joints follows the secretion of the hormone relaxin. This allows appreciable separation of the symphysis pubis, giving abdominal pain that is much aggravated by walking. In extreme cases weight bearing is impossible and the woman has to retire to bed completely. Treatment is rest; binders are of little help. Vaginal delivery should be anticipated. This condition may take up to two months to resolve after delivery, but it usually does slowly get better. Severe cases may last for up to a year, and long-term follow-up is wise.

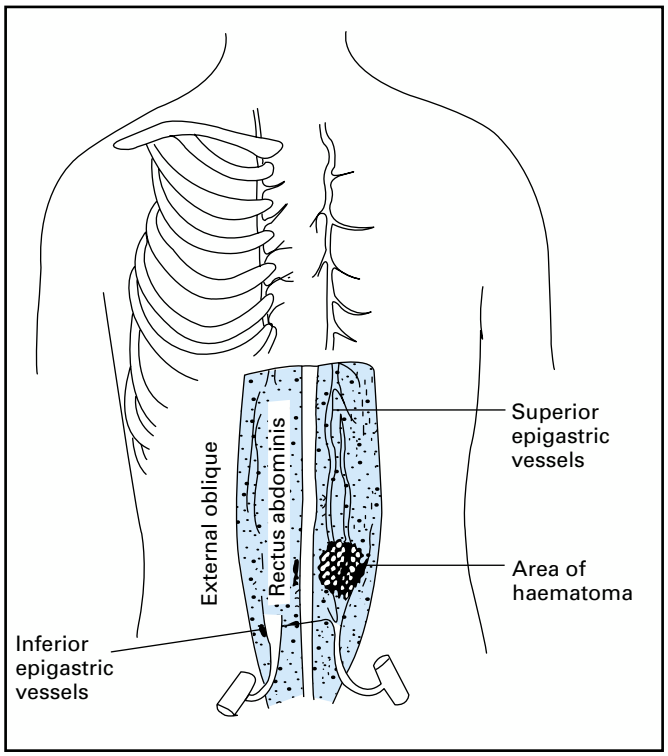


Figure 8.14 A rectus haematoma usually arises from the inferior epigastric vessels deep in the rectus muscle

Conclusion

Most women who present with abdominal pain in pregnancy may have nothing serious the matter. Pain can, however, lead the doctor to diagnose a serious condition, when action needs to be taken. As investigations play a small part in many of these diagnoses, experienced general practitioners can often diagnose its cause and continue the management of many women at home, but if there is any doubt the local obstetric department ought to be consulted.

All general medical conditions are modified by pregnancy; diagnosis may be clouded and treatment may have to be changed. Early abdominal examination will usually help differentiate serious from lesser conditions. If the condition is thought to be serious consult an obstetrician early rather than send to a general surgeon.

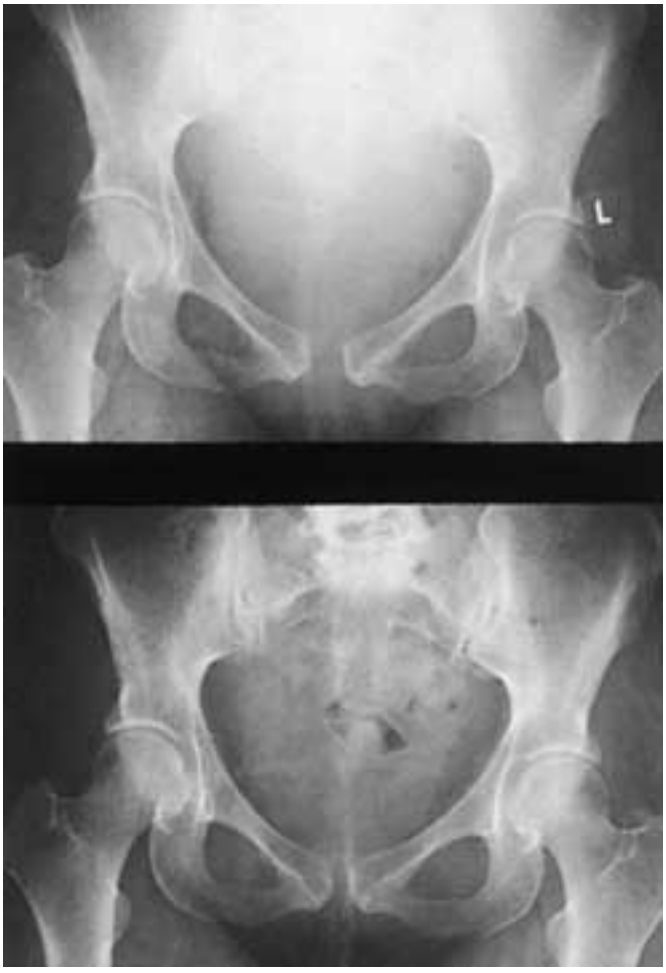


Figure 8.15 Above: Pelvis immediately after delivery showing dehiscence of pubic symphysis. Below: Same pelvis six weeks later. Imaging by ultrasonography reduces the risks of irradiation in a young woman

References

1 European Collaborative Study. Caesarian section and the risk of vertical transmission of HIV-1 infection. *Lancet* 1994;**343**:1464–7.
2 Dunn D, Newell M, Mayaux M *et al.* Mode of delivery and vertical transmission of HIV-1. *J AIDS* 1994;**7**:1064–6.

Recommended reading

- Johnstone F. *HIV and pregnancy*. Year of Obstetrics and Gynaecology, Volume 8. London: RCOG Press, 2000.
- Nelson-Piercy C. *A handbook of obstetric medicine*. Oxford: Isis Medical Media, 2000.
- Rubin P. *Prescribing in pregnancy*, 2nd edn. London: BMJ Publishing Group, 1995.
- Sbarouni E, Oakley C. Outcome of pregnancy in women with valve prosthesis. *Br Heart J* 1994; **71**:176–201.

The table showing therapeutic concentration of anticonvulsants is based on that by J Donaldson in *Critical care of the obstetric patient*, edited by R Berkowitz, and is reproduced by permission of Churchill Livingstone. The photographs of the glucose testing equipment are reproduced by permission of Boehringer Mannheim (United Kingdom).

9 Raised blood pressure in pregnancy

One of the original aims of promaternity (antenatal) care in 1901 was the prevention of fits and convulsions due to eclampsia, which was often associated with pre-eclampsia. The term pre-eclampsia has been refined in later years as eclampsia now occurs rarely.

Raised blood pressure affects the fetus as well as the mother. In the later weeks of pregnancy it may fall into one of several categories.

- Chronic hypertension is present before the 20th week and has causes outside pregnancy.
- Pregnancy-induced hypertension develops after the 20th week of pregnancy and usually resolves within 10 days of delivery.
- Pregnancy-induced hypertension with proteinuria now is called pre-eclampsia and occurs mostly in primigravidas.
- Pregnancy-induced hypertension with or without proteinuria may be superimposed on chronic hypertension and this is a most dangerous combination, the effects of pregnancy being added to those of chronic hypertension.
- Eclampsia is a convulsive condition usually associated with proteinuric hypertension.

Box 9.1 Some accepted definitions of raised blood pressure

- Hypertension
  - Mild—diastolic blood pressure >90 mm Hg
  - Severe—diastolic blood pressure >110 mm Hg
- Pregnancy-induced hypertension
  - Mild—diastolic blood pressure >90 mm Hg after the 20th week of pregnancy with no raised blood pressure beforehand and no proteinuria
  - Moderate—diastolic blood pressure >100 mm Hg after the 20th week of pregnancy with no raised blood pressure beforehand and no proteinuria
  - Severe—diastolic blood pressure >90 mm Hg after the 20th week of pregnancy with no raised blood pressure beforehand but with any degree of proteinuria

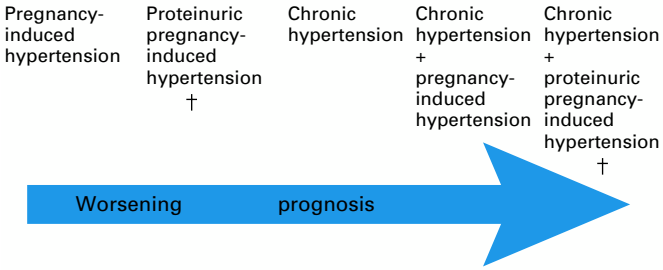


Figure 9.1 Permutations of hypertensive disease in pregnant and non-pregnant women. †These are designated as pre-eclampsia

Causes

The mechanism of pregnancy-induced hypertension is now almost completely understood, with reasonable educated guesses being possible in unknown cases. The primary defect is failure of the second wave of trophoblastic invasion into the decidua. Usually the trophoblast invades the entire length of the spiral arteries by 22 weeks of gestation. This leads to an appreciable fall in peripheral resistance and therefore a fall in blood pressure. In addition, as the trophoblast usually removes all the muscle coat of the spiral arteries, blood flows unimpeded into the intervillous space, gushing like a fountain over the villous tree that contains the fetal vessels. This ensures adequate time for exchange of oxygen, nutrients, and the waste products of metabolism.

If the second wave of trophoblastic invasion fails, the peripheral resistance does not fall and the haemodynamic mechanisms are not reset for the increased vascular space of pregnancy. Furthermore, the muscle coats retained by the spiral arterioles are sensitive to circulating pressor agents, particularly angiotension II. Most of the hypertensive changes are due to hormonal rather than sympathetic nervous system influence. At the spiral arterioles, the reduced volume of trophoblast leads to an imbalance in the prostacyclin–thromboxane system. The comparative overproduction of thromboxane encourages vasospasm of the spiral arteries and also local platelet aggregation. The lower concentrations of prostacyclin remove the protection that pregnancy offers against angiotension II.

The damaged muscle coating and intima of the spiral arteries undergoes acute atherosis, an accelerated form of arteriosclerosis that further narrows and then occludes the arterioles. A further increase in blood pressure follows, and the decrease in perfusion of the intervillous space leads commonly to intrauterine growth retardation.

Low dose aspirin may reduce the severity of pregnancy-induced hypertension in patients at risk, moderating the disease once established. The mode of action is irreversible

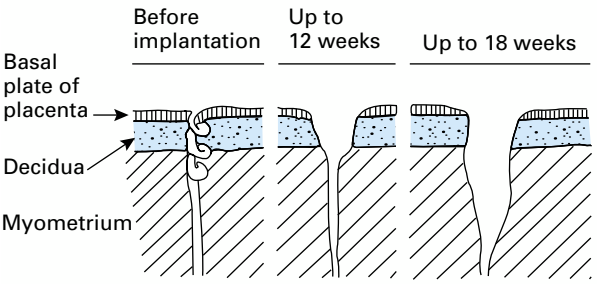


Figure 9.2 The invasion of spiral arteries by the trophoblast converts them into deltas and so improves blood flow

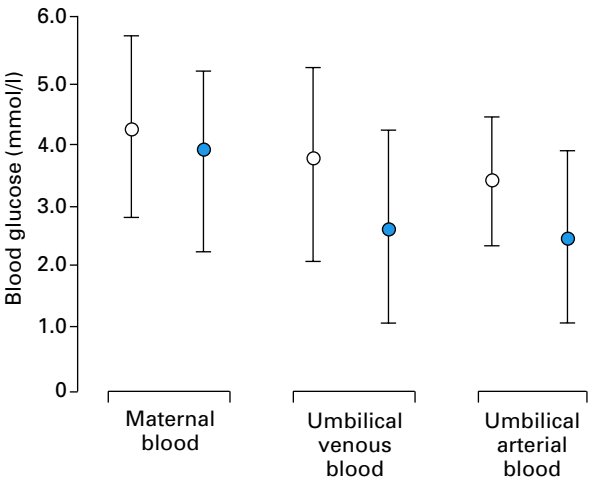


Figure 9.3 Transfer of glucose from mother to fetus in babies who show normal growth (o) and in those who are small for gestational age

poisoning of platelet cyclo-oxygenase. This probably prevents or delays clotting in the spiral arterioles.

The effects of pregnancy-induced hypertension on organs other than the placenta are mediated by the effects of hypertension or by activation of the complement system. This causes immune complexes to be deposited on the basement membrane of the kidney and allows protein to leak into the urine. In severe disease platelets are both consumed and activated so that coagulopathy may follow.

Management

Though pregnancy-induced hypertension develops out of the blue, particularly in first pregnancies, many women who already have hypertension will wonder about becoming pregnant and the effects that the pregnancy may have on their underlying hypertension. This matter should be considered carefully before a woman becomes pregnant, and if necessary the woman should be referred to a local prepregnancy advisory service. Since tobacco is associated with increased risks of cardiovascular disease in general, one would expect smoking mothers to have a higher rate of pre-eclampsia. This is not so and many studies have shown that smoking is associated with lower rates of pre-eclampsia. However, if it does occur it is often more severe in the smoker.

Generally speaking, if the blood pressure is not very high, or it can be kept low with antihypertensive drugs, and if there is no concomitant proteinuria before pregnancy, most women will have a successful pregnancy. They should continue their antihypertensive treatment in pregnancy.

Women with renal damage already leading to proteinuria and those who have diastolic pressures above 100 mm Hg despite adequate antihypertensive treatment should be investigated more thoroughly. Such women have a three to seven times increased risk above background of developing pregnancy-induced hypertension on top of their disease and the prognosis is worse for both mother and baby.

The ideal start to the management of pregnancy-induced hypertension, with or without proteinuria, is to detect it early. Each visit to the antenatal clinic includes a blood pressure recording. Recently, women likely to develop pregnancy-induced hypertension have been detected before this happens at 24 weeks by the use of Doppler measurements of blood flow velocity of uterine arteries, from which a measure of placental vascular resistance is derived. Doppler investigation may become available as a screening test in the next few years, providing, for example, an indicator of which women would benefit from low dose aspirin. Once prostaglandin was shown to be involved, an obvious antidote seemed to be aspirin and for a while this was in favour. Unfortunately the randomised CLASP study showed that in 9264 women there was only a 12% reduction in the incidence of proteinuria pre-eclampsia which was not significant.[1] Another possible organic cause of proteinuric hypertension has been the reduction of nitric oxide. This has led to the use of glyceryl trinitrate patches but this is still in the realms of research.

Once raised blood pressure is established, rest is usually central to primary management. Without accompanying proteinuria, the woman may be treated at home, where rest must take priority over everything else, including work at home or outside and care of other members of the family. Those with other children find it difficult to follow this regime and probably a third of women do not rest when so advised. If the hypertension increases despite proper bedrest, or proteinuria follows, admission to hospital is required.

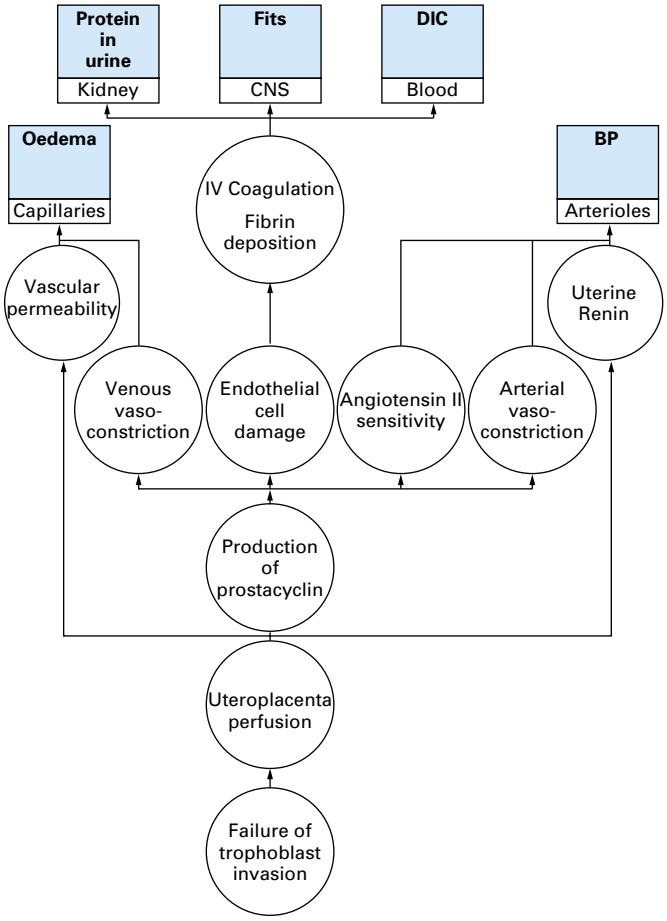


Figure 9.4 The suggested pathways (–O–) of pregnancy-induced hypertension changes related to their outcomes (■)

Table 9.1 Risk factors for the development of pregnancy-induced hypertension

Risk factors	Ratio
Nulliparity	3:1
Age above 40 years	3:1
Chronic hypertension	10:1
Chronic renal disease	20:1
Twins	2:1

Figure 9.5 Blood pressure measurement is a simple and useful screening test when performed repeatedly by standardised techniques. All doctors and midwives in a unit should use the same criterion for diastolic pressure—probably the loss of phase V Korotkoff sound

In hospital rest will be reinforced and the condition will be monitored by using ultrasound measurements of the growth of the fetus, Doppler measurements of blood velocity in the umbilical arteries and some would measure flow in the uterine arteries. Cardiotocographic measurements of variations in the fetal heart rate may also be used. Plasma urate concentrations and an increase in the liver enzyme aspartate transferase are useful biochemical indicators of deterioration, and a fall in the platelet count reflects severe disease. (The HELLP Syndrome – Haemolgia Elevated Liver Enzymes, Low Platelets). The management of severe hypertension now no longer includes treatment with sedatives or diuretics; sedatives tend merely to reduce the mother's level of consciousness and cross the placenta, causing depression of the fetal central and peripheral nervous systems. Similarly, diuretics are of little use, except for the relief of acutely painful oedema. They may even be harmful by reducing plasma volume and therefore perfusion of the placental bed.

Antihypertensive drugs are useful in protecting the mother's circulation, mostly against the risk of a stroke. They have no effect on the progression of the pregnancy-induced hypertension or on fetal growth but they help to maintain the pregnancy longer, so allowing the fetus to become more mature. These drugs tend to be kept for women whose hypertension increases despite bedrest. Methyldopa is still the commonest oral drug used in the short term. Hydralazine is given intravenously as first aid in acutely deteriorating hypertension. Combined α and β blockers, such as labetalol, are gaining in popularity because they give better control.

Calcium channel blockers such as nifedipine are being used more widely for they are effective in the control of acute hypertension. No serious fetal side effects occur although maternal side effects of flushing and headache may demand discontinuation.

The final and ultimate treatment of pregnancy-induced hypertension is delivery. Induction of labour or caesarean section should be reserved until the fetus is mature enough for the neonatal facilities available, but it must be used when the condition deteriorates. Two changes in managing pregnancy-induced hypertension have considerably altered the outlook for mother and fetus.

- Firstly, use of antihypertensive drugs to allow the fetus to spend longer in the uterus has spread rapidly and widely. Formerly, such drugs were thought to reduce placental bed perfusion and so affect the fetus deleteriously; their use in pregnancy was restricted. Now most obstetricians use them, and by reducing maternal risk, pregnancy is prolonged by a few more weeks so that the child is more mature.
- Secondly, the obstetrician's reluctance to perform a caesarean section earlier in pregnancy has diminished. With improved intensive neonatal care, caesarean section as early as 28 weeks gives a reasonable chance of fetal survival. The worst effects of prolonged renal and cerebral damage are reduced for the mother and the fetus is delivered before being affected by serious chronic hypoxia *in utero*.

The treatment of women with severe pregnancy-induced hypertension is best performed in special regional hypertension units, where neonatal and obstetric care is planned together. The Confidential Enquiries into Maternal Deaths have urged for years that each Health Authority should have one or more such designated units. A woman with or at risk of severe pregnancy-induced hypertension should be admitted to such a unit to obtain the best concentrated and coordinated obstetric and neonatal care.

The future management of pregnancy-induced hypertension may lie in the reduction of platelet agglutination during early

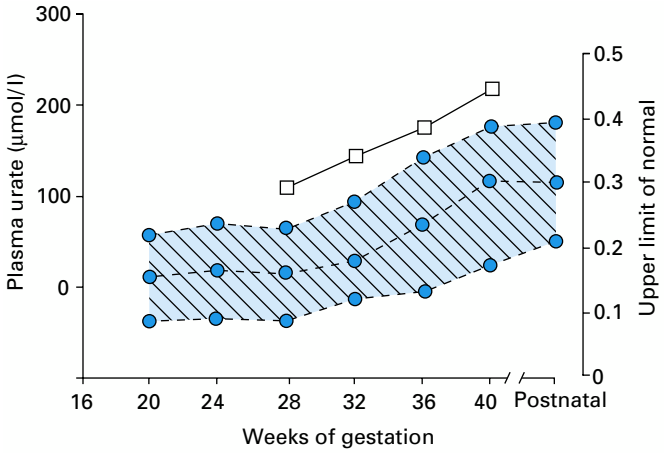


Figure 9.6 Changes in plasma urate concentration from 16 weeks of gestation showing 10th, 50th, and 90th centiles and the accepted upper limit of normal values. ❑—❑ shows the levels in a woman with severe pre-eclampsia

Table 9.2 Drugs and dosages used in treatment of pregnancy-induced hypertension

Drug	Route	Dosage	Comment
Centrally acting drugs			
Clonidine	Oral	500–100 μg three times a day	
Methyldopa	Oral	250–1000 mg daily	Safe to use
Vasodilators			
Sodium nitroprusside	Intravenous	0.3–1.0 μg/kg/min	Only for short-term use
Hydralazine	Intravenous	5–20 mg over 20 minutes	Drug of choice in emergency
β Adrenoceptor blockers			
Propranolol	Oral	80–160 mg daily	Used to be thought to reduce placental perfusion
α and β Adrenoceptor blockers			
Labetalol	Intravenous	50 mg over a minute	Water soluble and so crosses placenta; may not be effective in acute problem
	Oral	100–200 mg daily	

The ultimate treatment of pregnancy-induced hypertension is delivery.

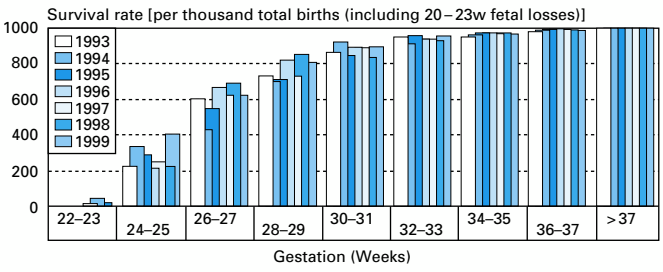


Figure 9.7 Survival by gestational age, Wales 1993–9

pregnancy, so preventing damage to the placental bed. This might halt the whole cascade of problems. Aspirin in early pregnancy might block the cyclo-oxygenase enzymes of the platelets so that they would not be able to produce thromboxane. It was thought that low dose aspirin (75 mg a day) may be helpful in mitigating the worst effects of pregnancy-induced hypertension with proteinuria but the published results of the CLASP study do not substantiate this.[1]

Eclampsia

Imminent eclampsia

The old term fulminating pre-eclampsia is less often used, but semantics are not as important as the recognition of this severe, acute change in a woman's condition. Having had moderate or even severe but symptom-free pregnancy-induced hypertension with proteinuria, the woman suddenly starts to produce symptoms. She may have frontal headaches and visual symptoms with jagged, angular flashes at the periphery of her visual fields and loss of vision in areas, both symptoms being due to cerebral oedema. She often has epigastric pain due to stretch of the peritoneum over the oedematous liver. In addition, some women have a curious itch confined to the mask region of the face. On examination her blood pressure may be much raised above previous readings or proteinuria may increase sharply; she may have increased and brisk reflex responses at knee and clonus. This woman needs urgent hypotensive and anticonvulsant treatment. If she is at home she should be admitted, with intravenous diazepam and, if necessary, hydralazine running continuously. Diazepam prevents fits and hydralazine reduces blood pressure but magnesium sulphate does both.[2]

Box 9.2 Symptoms and signs of imminent eclampsia

- Upper abdominal pain
- Itching on the face
- Flashes of light
- Headache
- Rapidly increasing blood pressure
- Increasing proteinuria
- Increased knee jerks—hyper-reflexia

Eclampsia

Convulsions associated with pregnancy-induced hypertension are termed eclampsia; they are very similar in form to those of epilepsy. Occasionally women in the beginning of the third trimester have eclamptic fits, having had perfectly normal blood pressure readings and urine test results within the previous few weeks at the routine visits to the antenatal clinic. Most women with eclampsia, however, give prodromal signs of pregnancy-induced hypertension with proteinuria in pregnancy; most are preterm (<37 weeks) while a fifth are before 32 weeks. The fits may develop in labour or the puerperium, the first day after birth having the highest risk.

The general practitioner's first move is to control the fits and prevent them causing damage to the woman. She should be laid on her side and an airway established. Intravenous diazepam is given to stop the fits, usually about 20–40 mg. This is followed in hospital by intravenous infusion of magnesium sulphate. This drug has been used for more than 60 years in the USA to prevent and treat eclamptic convulsions but has only recently found favour in the UK. It is thought to have central anticonvulsant activity. Clinical experience and research support its use in the prevention of subsequent eclamptic fits. It is usually given for at least 24 hours following the fit. Care must be taken as respiratory depression and loss of patellar reflexes may indicate toxicity.

Should the blood pressure be steeply raised, intravenous hydralazine is also given, either in a 5 mg bolus over 20 minute intervals or given intravenously as 25 mg in 500 ml of Hartmann's solution, with the drip rate titrated against the woman's blood pressure. This is best administered through a separate drip set so that magnesium sulphate and antihypertension treatments can be given at different rates

Box 9.3 Treatment of eclampsia

- Lie the woman on her side in the recovery position
- Keep airway clear
- Prevent trauma during fits
- Give diazepam immediately
- Give IV hydralazine if blood pressure is raised
- Give IV magnesium sulphate
- Use epidural anaesthesia if the woman is in labour or a caesarean section is planned

Box 9.4 Mode of delivery after control of eclampsia

- Factors favouring vaginal delivery
  - Multiparous mother
  - Stable blood pressure and diminished cerebral irritability
  - Ripe cervix
  - Mature fetus (>1500 g estimated weight)
  - Cephalic presentation
  - Normally grown fetus
  - Fetus in good state to stand uterine contractions
- Factors favouring caesarean section
  - Primiparous mother
  - Unstable blood pressure control or cerebral irritability
  - Unripe cervix
  - Immature fetus (<1500 g estimated weight)
  - Breech presentation
  - Intrauterine growth restriction
  - Poor prognosis of fetal state from Doppler blood flow rates or cardiotocography

according to clinical needs. If the woman is in labour or induction is considered, an epidural anaesthetic may be helpful, both to lower the blood pressure and to reduce the tendency to fit by removing the pain of intrauterine contractions. Any tendency of the woman to have disordered blood clotting should be excluded before insertion of a regional anaesthetic.

The ultimate treatment of eclampsia is delivery. Should eclampsia occur at home the woman must be transported to hospital immediately. Although rare, eclampsia still occurs in this country and the triennium 1994–96 was associated with 8 maternal deaths in the UK.

The ultimate treatment of eclampsia is delivery.

Maternal and fetal factors must be considered to find the best time for delivery of the fetus.

Timing of delivery

It must be emphasised that the ultimate cure of pregnancy-induced hypertension and eclampsia is delivery. The obstetrician must weigh the answers to two often conflicting questions:

- When would it be safer for the mother to be delivered?
- When would it be safer for the baby to be outside the uterus rather than on the wrong side of a failing placental exchange system?

Maternal considerations may be judged by the speed of deterioration of the condition (blood pressure and proteinuria) and the expected proximity of severe complications such as eclampsia. Fetal state is best evaluated by assessing the circulation supplying the fetus both in the spiral arteries with Doppler ultrasound measurements coming to the placental bed and in the umbilical vessels (discussed in Chapter 4). If there is time, serial ultrasound measurements of fetal growth are useful. If these data are available a rational decision can be made about the timing of the removal of the fetus from the hostile environment in a hospital with a neonatal intensive care unit. Women should be transferred early to regional centres for hypertension in pregnancy when it is obvious that the pregnancy-induced hypertension is not going to settle with bedrest and mild or moderate drug treatments. There is little place for heroic management in peripheral hospitals of a greatly compromised baby and mother.

Once it has been decided that it would be safer for the mother and the baby that delivery should occur the method and route of that delivery should be considered. If it is thought unsafe for the baby to undergo the contractions of labour, or if the baby is immature or has an inappropriate presentation, a caesarean section is indicated. If the mother's condition is deteriorating rapidly, again, the abdominal route would be swifter. An unripe cervix or an unsatisfactory presentation would also be grounds for a caesarean section. If, however, the woman has a ripe cervix, the hypertensive state is not worsening rapidly, and the fetus is in an acceptable position and of reasonable maturity, induction of labour should be performed with prostaglandin pessaries or membrane rupture, depending on the usage in the individual labour ward.

Intrauterine growth restriction is associated with pregnancy-induced hypertension. The two go together and share common causes. Narrowing of the placental bed vessels reduces nutrition to the fetus in pregnancy just as it reduces available oxygen during labour. Many fetuses born to women with unmanaged pregnancy-induced hypertension are small for their gestational age. Unfortunately so are many fetuses born to women who are very well managed; the fetal growth restriction therefore probably starts long before conventional management of the mother.

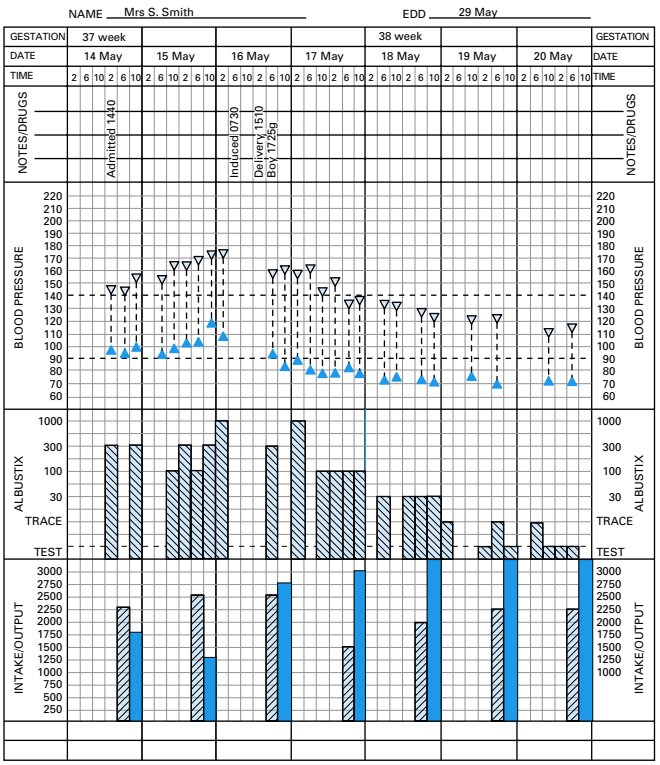


Figure 9.8 Partogram of a woman with severe pregnancy-induced hypertension before and after delivery

Box 9.5 Method of delivery (%) after various onsets of labour in women with pregnancies complicated by hypertension

- Spontaneous onset
  - Normal delivery – 3%
  - Vaginal operative delivery – 5%
  - Caesarean section – 10%
- Induced labour
  - Normal delivery – 17%
  - Vaginal operative delivery – 23%
  - Caesarean section – 22%
- Elective caesarean section – 11%

Conclusion

Pregnancy-induced hypertension is still a major problem in antenatal medicine but many of its worst effects can be mitigated by early diagnosis from blood pressure readings at clinic visits. The future includes predictive Doppler measurements of blood flow and preventive treatment, which may include aspirin, although the results of the CLASP trial in the United Kingdom are disappointing. If the condition is severe the mother's and baby's prognoses will be greatly improved if a regional hypertension in pregnancy unit is used.

Early diagnosis can modify some effects of pregnancy-induced hypertension.

References

1 CLASP. A randomised trial of low dose aspirin for the prevention and treatment of pre-eclampsia. *Lancet* 1994;**343**:619–29.
2 Eclampsia Trial Collaborative Group. Which anticonvulsant for women with eclampsia? *Lancet* 1995;**345**:1455–63.

Recommended reading

- Broughton Pipkin F. The hypertensive disorders of pregnancy. *Br Med J* 1995;**311**:609–13.
- Duley L. Anticonvulsants for the treatment of eclampsia. In: *Yearbook of obstetrics and gynaecology*, vol 5. London: RCOG Press, 1997.
- RCOG. *Management of eclampsia*. Guidelines no. 10. London: RCOG, 1999.

The figure showing transfer of glucose is reproduced by permission of Blackwell Scientific Publications from *Modern antenatal care of the fetus* edited by G Chamberlain and that showing change in plasma urate concentrations by permission of Churchill Livingstone from *Turnbull's obstetrics* edited by G Chamberlain.

10 Antepartum haemorrhage

Antepartum haemorrhage is bleeding from the genital tract between 24 completed weeks of pregnancy and the onset of labour. Some of the causes exist before this time and can produce bleeding. Although strictly speaking such bleeding is not an antepartum haemorrhage, the old fashioned definition is not appropriate for modern neonatal management.

The placental bed is the commonest site of antepartum haemorrhage; but in a few cases bleeding is from local causes in the genital tract. In a substantial remainder the bleeding may have no obvious cause but is probably still from the placental bed.

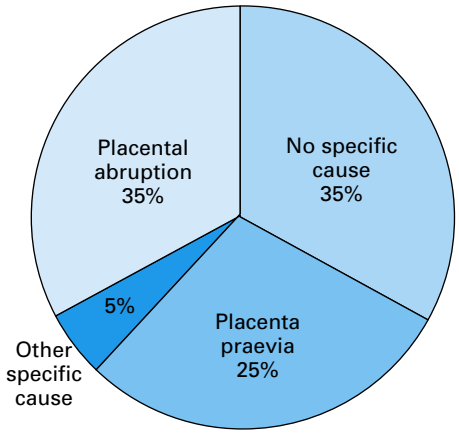


Figure 10.1 Causes of antepartum haemorrhage

Placental abruption

If the placenta separates before delivery, the denuded placental bed bleeds. If the placenta is implanted in the upper segment of the uterus the bleeding is termed an abruption; if a part of the placenta is in the lower uterine segment it is designated a placenta praevia.

Placental abruption may entail only a small area of placental separation. The clot remains between placenta and placental bed but little or no blood escapes through the cervix (concealed abruption). Further separation causes further loss of blood, which oozes between the membranes and decidua, passing down through the cervix to appear at the vulva (revealed abruption).

In addition, the vessels around the side of the placenta may tear (marginal vein bleeding), which is clinically indistinguishable from placental abruption. The differentiation between revealed and concealed abruption is not very useful. The important factor is the amount of placenta separated from its bed and the coincident spasm in the surrounding placental bed vessels. If the area of separation and the proportion of placental bed vessels driven into spasm is sufficient, it will lead to fetal death.

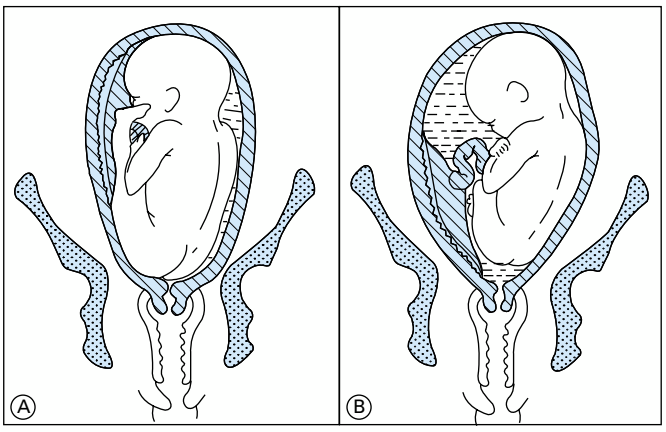


Figure 10.2 Placenta sited in (A) upper and (B) lower segment

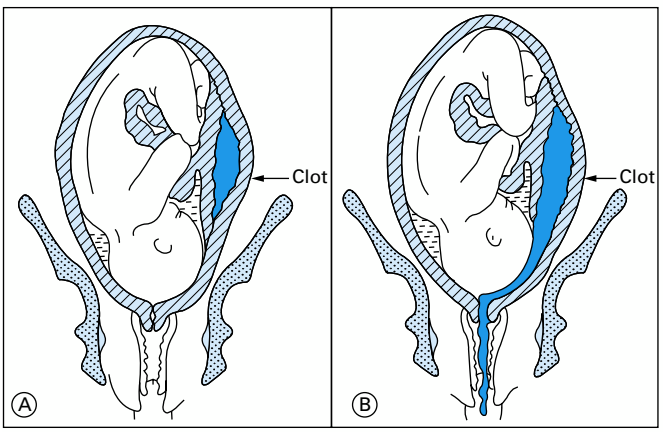


Figure 10.3 (A) Concealed and (B) revealed abruption from a normally sited placental bed

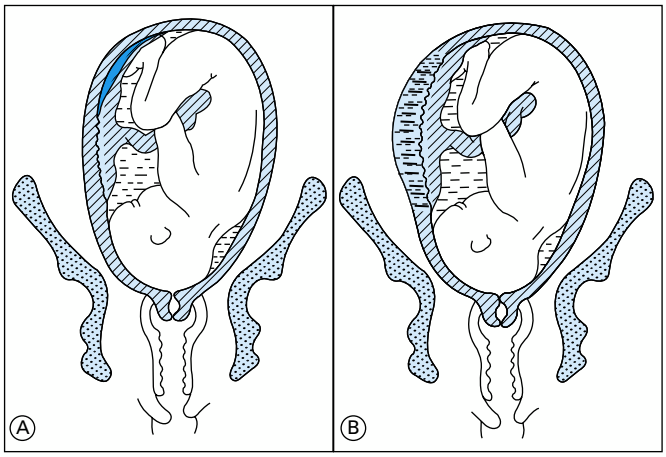


Figure 10.4 The degree of fetal effect depends on the amount of separation and spasm of placental bed vessels (A), while the maternal effect depends on the amount of tissue damage to the myometrium (B)

Pathology

Bleeding between the placenta and its bed causes separation; as more blood is forced between the layers, detachment becomes wider. Blood also tracks between the myometrial fibres, sometimes reaching the peritoneal surface. The mother's pain and shock depend on the amount of tissue damage rather than on the volume of bleeding. The fetal state depends on both the

amount of separation and the spasm of the more peripheral blood vessels in the placental bed.

Sometimes amniotic fluid or trophoblast tissue is forced into the maternal circulation after a placental abruption. Thromboplastins start disseminated intravascular coagulation, which in a mild case is coped with by the maternal fibrinolytic system, but if an amniotic fluid embolus is large, maternal plasma fibrinogen concentration is depleted. Uterine bleeding continues with activation of the maternal fibrinolytic system; widespread deprivation of fibrin and fibrinogen follows, producing a vicious circle of more bleeding.

The cause of placental abruption is unknown. It happens more commonly in association with a uterine abnormality and there is a 10% risk of recurrence if it has occurred previously. Conditions of uterine overstretch such as twin pregnancy are associated with higher rates of abruption if amniotic fluid is released suddenly at the rupture of the membranes. Abdominal trauma is a less common association.

Diagnosis

The woman presents with poorly localised abdominal pain over the uterus; there may be some dark red vaginal bleeding or clots. Depending on the degree of placental separation, uterine spasm, and the loss of circulating blood into the tissue space, clinical shock may also be present. If the abruption is severe the uterus contracts tonically so that fetal parts cannot be felt; the fetus may be dead with no fetal heart detectable. Ultrasonography may show the retroplacental clot but gives no measure of the extent of functional disorder.

The differential diagnosis is from:

- Placenta praevia, which is not usually accompanied by pain, often results in brighter red bleeding as the blood is fresher and rarely results in so much shock.
- Rupture of the uterus, which may present with a similar picture to that of placental abruption.
- Red degeneration of a uterine fibroid at 24–30 weeks' gestation.
- Bleeding from a ruptured vessel on the surface of the pregnant uterus, which is rare.

The diagnosis of abruption is finally confirmed after delivery by finding organized clot firmly adherent to the placenta.

Management

A woman with an abruption is in a potentially dangerous condition and requires all the facilities the emergency services can provide. She must be admitted to hospital quickly. Group O rhesus negative blood may rarely be required urgently in the home but even if not, supportive intravenous treatment should be established. Hartmann's solution or saline may be used at first followed by a commercial plasma expander such as Haemaccel. Pain may be relieved by morphine, and the woman must be transferred to hospital, escorted by her GP, trained paramedic staff or the Flying Squad, when her condition is stable.

In hospital the antishock measures will be continued and blood given. At least six units of blood must be crossmatched, irrespective of the scant external blood loss; fresh frozen plasma and platelets should be available. Central venous pressures are a guide to the amount of blood required to prevent undertransfusion before delivery or overtransfusion afterwards. Once the condition is stabilised delivery should take place immediately. If the fetus is still alive, this could mean a caesarean section. This can be a difficult operation needing a

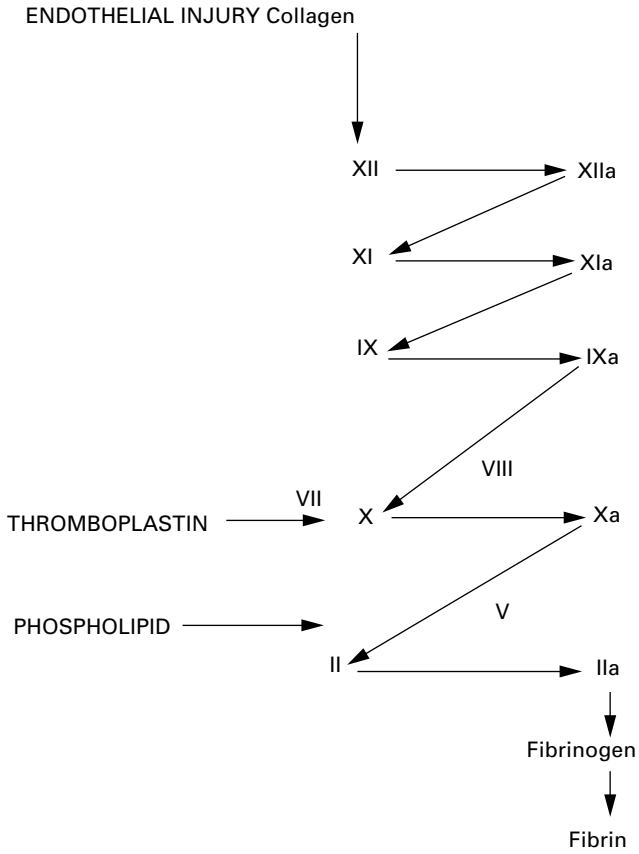


Figure 10.5 Points in the clotting cascade at which the sequelae of a placental abruption can intervene and so lead to disseminated intravascular coagulopathy

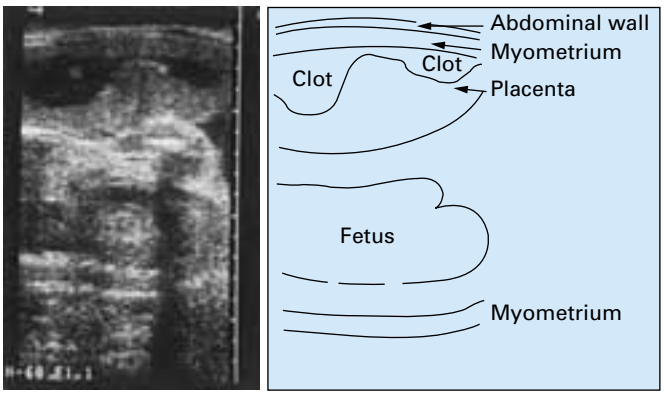


Figure 10.6 Ultrasound scan of placental abruption

Box 10.1 Management of placental abruption

- Get the woman to hospital urgently
- Replace volume of blood estimated lost from circulation rather than that seen at external loss
- Monitor central venous pressure
- Check for disseminated intravascular coagulopathy
- Check renal function and urinary output
- If fetus alive and mature, Caesarean section
- If fetus dead, induce (artificial rupture of the membranes)

senior obstetrician. If the fetus is dead, induction by rupture of the membranes usually leads to a rapid labour.

After a mild abruption and if the fetus is immature and lives the woman may continue the pregnancy under controlled conditions. She should stay in hospital with antenatal monitoring until the fetus is mature enough for delivery. In cases occurring very early in gestation the woman may have to be transferred for delivery to a regional unit with intensive neonatal facilities available.

Severe abruption may lead to severely disordered blood clotting which must be managed with the help of a haematologist. After delivery fluid balance should be carefully managed and urine output must be recorded hourly. Oliguria following reduced plasma volume is usually the result of acute tubular necrosis, though in rare cases acute cortical necrosis may occur. The help of anaesthetists trained in intensive care and of a renal physician will be needed.

Placenta praevia

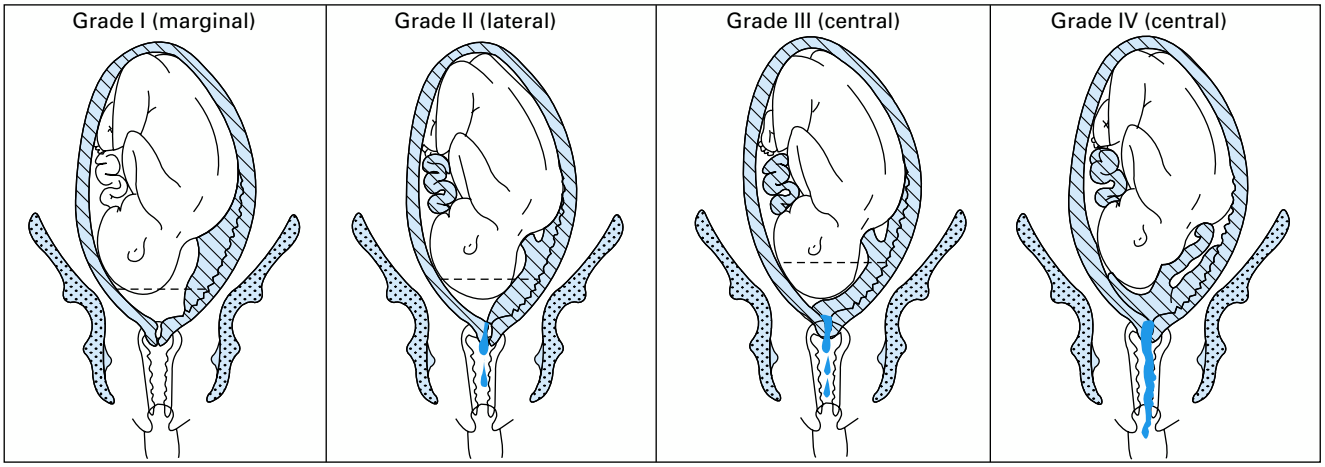


Figure 10.7 The older grades of placenta praevia were 1–4. They are now described in three grades: marginal, lateral, and central

The blastocyst usually implants in the thicker, receptive endometrium of the upper uterus, but occasionally it glissades to the endometrium of the isthmus or over a previous lower segment uterine scar. Then invasion by the trophoblast secures the embryo and when the uterus grows to form a lower segment later in pregnancy some part of the placenta is implanted there.

About a quarter of all antepartum haemorrhages are due to placenta praevia, the proportion increasing with more thorough investigative ultrasonography. In the last weeks of pregnancy the lower segment stretches whereas the placenta is comparatively inelastic. In consequence, the placenta which has implanted in the lower segment is peeled off the uterine wall with bleeding from the placental bed. A placenta praevia may be detected by ultrasonography in the mid-trimester but usually little bleeding occurs until the lower segment is formed after the 30th week.

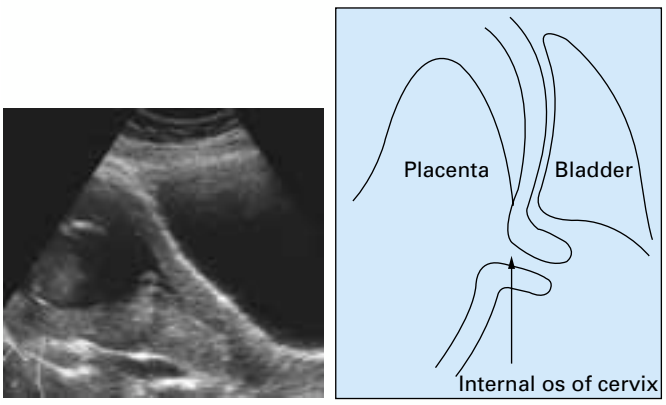


Figure 10.8 Ultrasound scan of placenta praevia

Diagnosis

A woman with placenta praevia may have bright red, painless vaginal bleeding. It comes unexpectedly, blood often being found on waking in the morning. The woman is in no way shocked and may wish to ignore the symptom as she feels normal.

A few women present with a persistent transverse lie or breech presentation in late pregnancy. The possibility of

placenta praevia should always be considered in such a case and an ultrasound scan requested urgently. The result may lead little the woman's admission to hospital, even if she has had little bleeding.

In a third group of women a placenta praevia is diagnosed incidentally on ultrasound examination. This finding is common in the middle weeks of pregnancy. A low lying placenta diagnosed at 22 weeks' gestation is often normally sited by 32 weeks. About 5% of women present with a low lying placenta at 24 weeks but only 1% of them have a placenta praevia at term. The upper segment of the uterus grows and the placental site moves with it as the lower segment is formed. If not, such women should be treated in the same way as others diagnosed clinically because the risk of bleeding in late pregnancy is as great.

The uterine spasm of placental abruption does not occur in placenta praevia and the fetus can be felt easily. The fetus is usually alive with a good heart beat. The woman's degree of shock will vary directly with the amount of blood lost. If shock is moderate the woman needs admission to hospital. If blood loss is slight she can go to the hospital conventionally but she needs to be warned of the probable diagnosis.

No vaginal examinations should be performed on any woman who bleeds in late pregnancy until a placenta praevia has been excluded by ultrasonography. If this principle is broached, further separation of the placenta may occur with very heavy, and sometimes fatal, haemorrhage. Any woman who presents to a general practitioner with vaginal bleeding in late pregnancy should be considered to have a placenta praevia until the diagnosis is disproved. She must be referred to a hospital for an urgent appointment that day. If necessary, she should be admitted if ultrasound investigations cannot be performed straight away.

In hospital blood is crossmatched and the placental site demonstrated by ultrasonography. The older diagnostic radioisotope studies and soft tissue *x* ray examinations now have no place in the UK.

Once placenta praevia is diagnosed, the aim of treatment is to maintain the pregnancy until the fetus is mature enough to be delivered; at 38 weeks an elective caesarean section will probably be performed unless the placenta praevia is a minor one with the fetal presenting part below it. Should the placenta be anterior, the descending fetal head may compress it against the back of the symphysis pubis, so allowing a vaginal delivery, but this is uncommon. The Caesarian operation may be difficult with much blood loss and should be performed by a senior obstetrician.

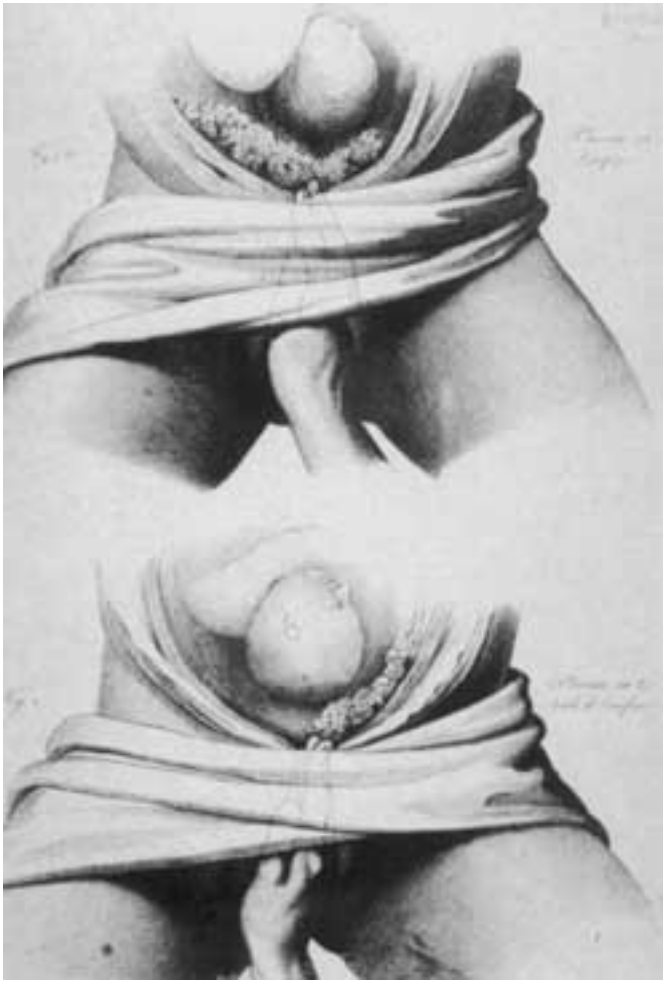


Figure 10.9 These old steel engravings show what a vaginal examination could do to a placenta praevia (central (above) and lateral (below)). NEVER DO A VAGINAL EXAMINATION UNLESS PLACENTA PRAEVIA HAS BEEN EXCLUDED

Other specific causes of bleeding

General

Few haemorrhagic diseases occur in young women but vaginal bleeding may occur in von Willebrand's disease, Hodgkin's disease, and leukaemia. All are probably known about beforehand, and the diagnosis is confirmed from the results of haematological studies.

Local

Lesions of the cervix and vagina cause slight bleeding, often only a smear of blood and mucus. Moderate bleeding may occur with a carcinoma of the cervix—unusual in women of childbearing age—or varicose veins of the vulva and lower vagina. Lesser bleeding is more likely from a polyp or an erosion of the cervix. Monilia infection may be accompanied by spotting as plaques of fungoid tissue are separated from the vaginal walls.

Table 10.1 Causes of antepartum bleeding from the lower genital tract

Cause	Characteristic bleeding
Cervical ectropion	Smear of blood loss often with mucous loss
Cervical polyp	Spotting of blood
Cervical cancer	Smear of blood on touch (rare, but diagnosis is important)
	May bleed heavily
Vaginal infection	Spotting of blood with white or pink discharge
Vaginal varicose veins	Occasionally heavy bleeding

All these causes can be diagnosed by using a speculum, but this procedure must be done in hospital after the woman has been assessed and ultrasound examination has excluded placenta praevia. If the haemorrhage is due to a benign local lesion it will be managed appropriately.

Fetal

A most unusual cause of bleeding is from fetal blood vessels. There may be a succenturiate lobe or the umbilical cord may be inserted into the membranes over the internals so that the arteries and veins pass unsupported to reach the edge of the placenta. If by chance the placenta is also low lying, the umbilical blood vessels pass over the internal os of the cervix (vasa praevia); when the membranes rupture the fetal vessels may tear and bleed. The blood is fetal and a small loss can lead to severe hypovolaemia of the fetus.

The presence of vasa praevia is difficult to diagnose but sometimes they can be suspected with colour Doppler ultrasonography. More usually the fetal heart rate may alter abruptly after membrane rupture accompanied by a very slight blood loss. Bedside tests exist to differentiate fetal from maternal haemoglobin but are rarely used. The treatment must be a rapid caesarean section as the fetus cannot stand such blood loss for long.

Bleeding of unknown origin

The real cause of antepartum haemorrhage is unknown in a large number of women. They may have bled from separation of the lower part of a normally sited placental bed or the membranes may have sheared with tearing of very small blood vessels. Some placentas bleed early from their edge.

If the cause of antepartum haemorrhage cannot be diagnosed precisely, the woman should not be dismissed lightly. The risk to her baby at subsequent labour is higher than background, although the risk to the mother does not seem to be great. It is good practice to keep such women in hospital for some days, allowing them to return home if no further vaginal bleeding occurs. This rule of thumb seems to cover most eventualities and so many women do not stay in hospital for long. Fetal growth should be monitored by ultrasonography. In labour, however, the fetus should be monitored for hypoxia: for there is a higher risk than in fetuses whose mothers have not bled.

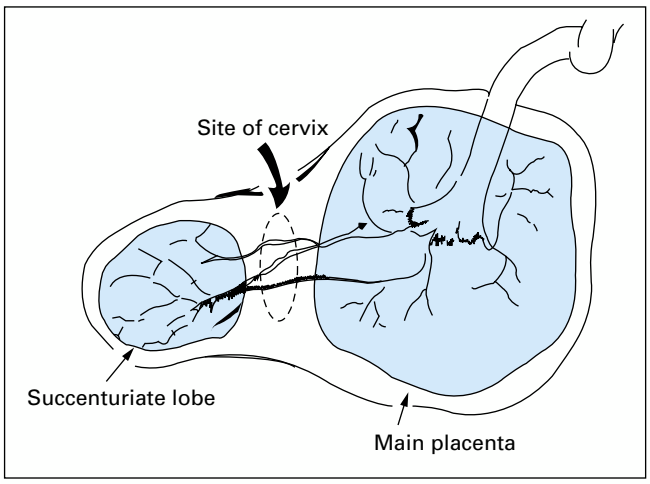


Figure 10.10 Vasa praevia. A succenturiate lobe is separated from the main body of the placenta. Should the vessels run over the cervix, when the cervix dilates they may be torn so that fetal blood is lost

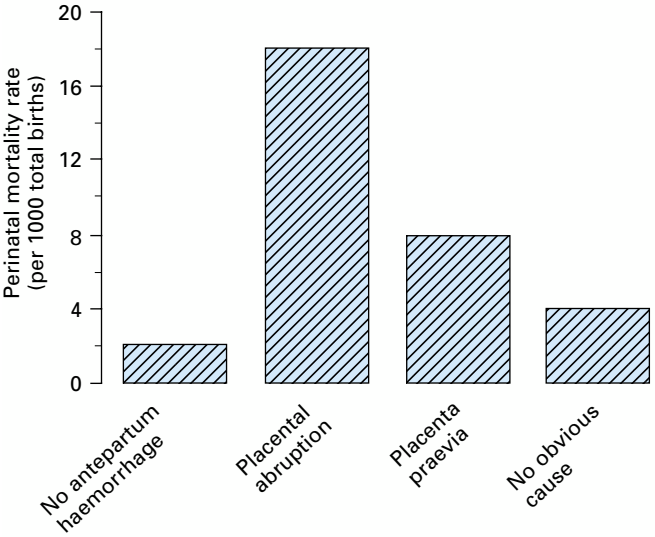


Figure 10.11 The relative risks of increased perinatal mortality from antepartum haemorrhage compared with those in pregnancies with no such haemorrhage

Recommended reading

- Barron F, Hill W. Placenta praevia, placental abruption. *Clin Obstet Gynaecol* 1998;**41**:527–32.
- Bonner J. Massive obstetric haemorrhage. *Best Pract Clin Obstet Gynaecol* 2000;**14**:1–16.
- RCOG. *Placenta praevia: diagnosis and management. Guidelines no. 27.* London: RCOG, 2001.

11 Small for gestational age

The problems of small babies and preterm labour often go together and are now the major causes of perinatal mortality and morbidity in the UK. Furthermore, they use up large amounts of facilities, manpower, and finance. Preterm labour and premature rupture of the membranes are considered in the next chapter and the antenatal care of fetuses that are small for gestational age and of their mothers in this one.

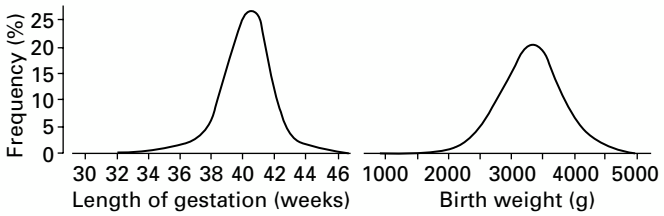


Figure 11.1 Distribution of length of gestation and birth weight (singletons, last menstrual period certain)

The phrase "intrauterine retardation" is no longer used in current obstetrics. It has been replaced by "intrauterine growth restriction" because the former phrase implied that there was some retardation of the child, particularly cerebral, and some parents found this difficult to accept.

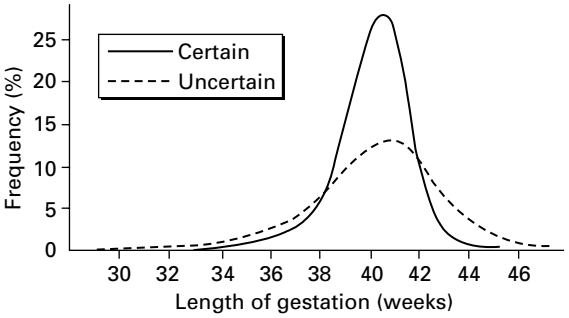


Figure 11.2 Distribution of length of gestation by knowledge of last menstrual period (singletons)

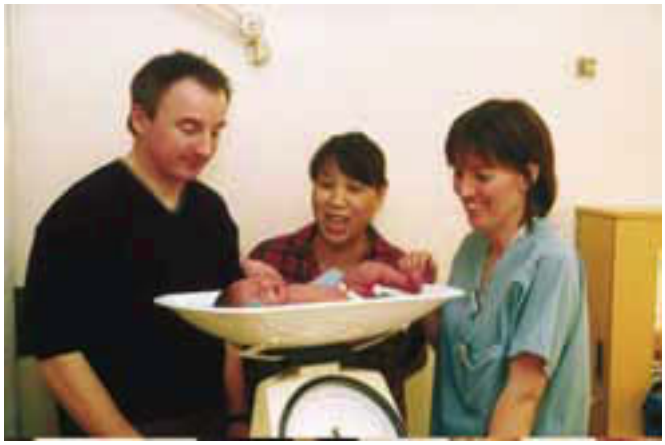


Figure 11.3 Weighing a newborn

The diagnosis of a small fetus is made more specific by examining the ratio of birth weight (or estimated birth weight) to gestational age. Both these measures have inherent problems.

Obstetricians estimate fetal weight either clinically or from measuring ultrasound determined diameters of the fetus *in utero*. Gestational age is derived from the mother's menstrual dates, which are usually confirmed by an ultrasound scan measuring the biparietal diameter performed before 20 weeks. In most parts of the UK, about 80% of women are sure of their dates. The figure shows the distribution of length of gestation for women according to whether they were sure of their dates. The frequency of heavier babies was increased among those uncertain of the date of their last menstrual period. All women in the UK with unsure dates should have gestational age established by ultrasound, as should those in whom there is a discrepancy between the dates derived from the last menstrual period and fetal size in early pregnancy. Obstetricians consider a baby to be small for gestational age when abdominal circumference readings fall below the second standard deviation below the mean; this is approximately the second centile on serial ultrasonography.

After birth paediatricians can weigh the baby and so have a precise measure, although even this varies slightly with the conditions of weighing and when it is done. Gestational age is obtained from the obstetrician by one of the previously mentioned measures or from Dubowitz scoring. The data are plotted on a specific centile chart; various groups of paediatricians take small for gestational age as being below the 10th, the fifth, or the third centile. It is very important when examining data to know which of these measures was used. The 10th centile is rather crude and will include many normal babies at the lower end of the normal birthweight distribution curves whose growth has not actually been affected by placental bed disease.

Much simpler was the old measure of prematurity, taking a cut off point of a birth weight of less than 2500 g. Unfortunately, this includes small babies whose birth weight is appropriate for their gestational age and those who are small for their gestational age, two very different groups in clinical medicine. For example, babies born with a birth weight below

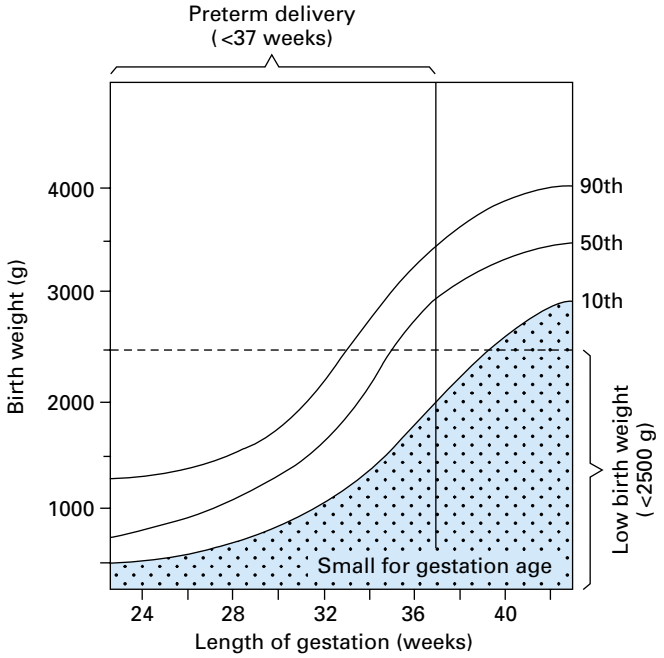


Figure 11.4 The relation between preterm and low birth weight babies. Babies who are small for gestational age fall under the 10th centile

2500 g make up about 7% of the newborn population in the UK, about 3% in Sweden, almost 11% in Hungary and a much higher proportion in many parts of the Eastern hemisphere. Such mixed data would make a nonsense of studying the influences on fetal growth and so the definition of small for gestational age relating birth weight to length of intrauterine life stands at the moment.

Causes

Genetic abnormalities

Genetic abnormalities are an identifiable but not very common factor causing growth restriction. Trisomy 21 is the commonest example, though osteogenesis imperfecta, Potter's syndrome, and anencephaly may all be associated with intrauterine growth restriction. Other congenital malformations not yet proved to have a genetic component are commonly found in fetuses that are small for gestational age; among them are gastrointestinal abnormalities such as atresia of the duodenum, gastroschisis, and omphalocele.

Maternal nutrition

In the UK the effect of maternal nutrition on low birth weight is probably small. Extremes of starvation associated with small babies are rare in Britain. During a pregnancy about 80 000 kilocalories (335 MJ) of extra energy is required, of which 36 000 kilocalories (150 MJ) is for maintenance metabolism.[1] Much of this can come from an everyday diet, and among well nourished women requirements change little for the first 10 weeks of pregnancy. Thence requirements gradually increase, but ordinary variations in food intake are unlikely to affect events. It is unwise to recommend that a mother eat for two in order to produce a larger baby. As well as the nutritional value of the food consumed, there are other factors of appetite, maternal obesity, and heartburn which must be remembered when making recommendations.

Intrauterine infection

Most intrauterine infections are viral or bacterial. Some 60% of babies with congenital rubella are born below the 10th centile of weight for gestation. Cytomegalovirus and toxoplasmosis (much less common in this country than in mainland Europe) are associated with growth restriction in about 40% of infected infants. Malaria, ubiquitous in many tropical countries, causes a massive accumulation of monocytes in the intravillus space, which is associated with a fetus being small for gestational age.

Drugs

Drugs may be a cause of babies being small for gestational age. The commonest cases in the UK are the results of tobacco fumes being absorbed during cigarette smoking. The association between smoking and small for gestational age babies is well documented. The number of affected babies whose growth drops below the 10th centile increases during the last weeks of gestation.

The effect of alcohol is difficult to sort out. At the extreme end of the range, i.e. women drinking more than 45 units of alcohol a week, some babies are born with the fetal alcohol syndrome and a distinctly reduced birth weight. At lower intakes of alcohol covariables come into play; a deficient maternal diet and increased cigarette smoking are often associated with the alcohol habit. In some studies multivariant analyses show that the main causal factor associated with low birth weight is not alcohol intake but cigarette smoking. The whole lifestyle is probably the important factor. Some doctors

In the UK most of the energy required by a pregnant woman can come from an ordinary diet, with little need for supplementation.

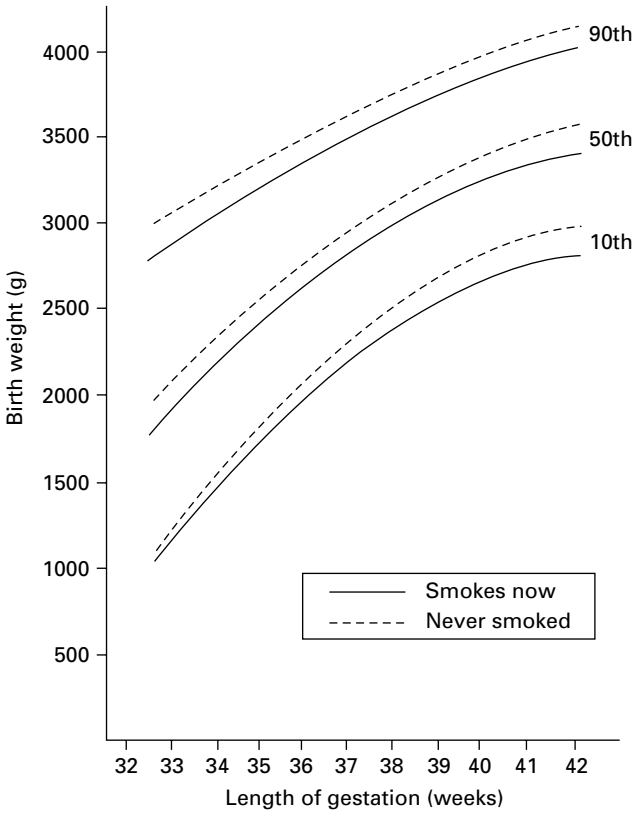


Figure 11.5 Centiles of birth weight by length of gestation and mother's smoking habit (singletons, last menstrual period certain)

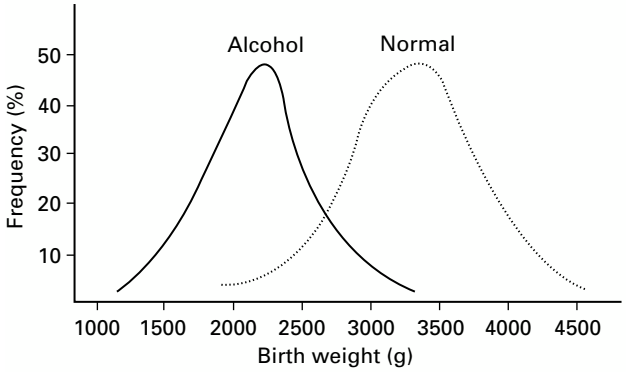


Figure 11.6 Distribution of birth weight in a normal population of women and in one consisting of women who drank more than 45 units of alcohol a week (heavy drinking)

consider that smoking in pregnancy is the most important single cause of low birth weight, the greatest single factor associated with death and illness in the first weeks of life.

The use of narcotic drugs is commonly associated with low birth weight but, again, the total lifestyle of the woman may be the real factor. For this reason, perhaps, an increased incidence of small for gestational age babies persists with methadone users.

Therapeutic drugs such as carbamazepine and the valproates have been associated with an increased incidence of small for gestational age babies, as have the more powerful antiviral drugs such as azathioprine. Such powerful drugs are not given in pregnancy unless they are needed to treat a serious maternal medical condition, which in itself may affect nutrition or metabolism of the mother and therefore growth of the fetus.

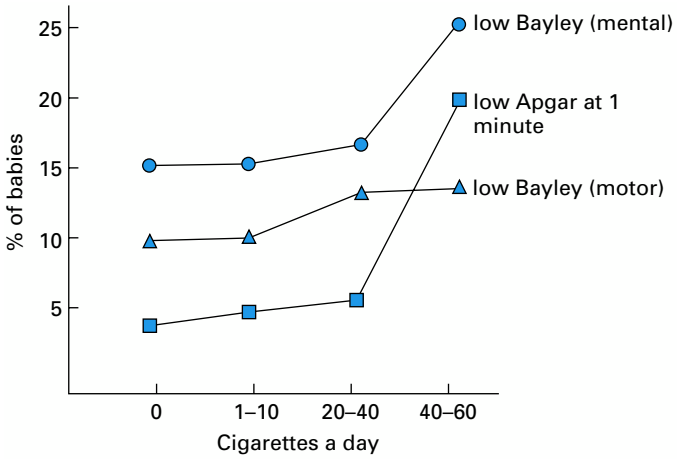


Figure 11.7 Relationship between amount of cigarette smoking in black American women and various non-weight related indices at birth and seven months later

Hypertension

One of the major current causes of babies being small for gestational age in the UK is hypertension in the mother, either pregnancy induced or pre-existing. After other features have been taken into account such types of hypertension are associated with about a third of all cases of intrauterine growth restriction. The effects of hypertension are made worse when raised blood pressure is associated with proteinuria, implying a greater reduction of the maternal perfusion of the placental bed. The duration of the condition also has an effect; for example, 80% of mothers who have proteinuric pregnancy-induced hypertension before the 34th week of pregnancy have infants with a birth weight below the 10th centile.

Other factors

The maternal body habitus is not a major factor in babies being small for gestational age, but big women do produce larger children. The father's influence is less important, classically shown in the 1938 study of Walton and Hammond on Shire horses and Shetland ponies.[2]

The altitude at which a woman lives in pregnancy has a negative effect on fetal growth, particularly if she is not used to high altitudes.

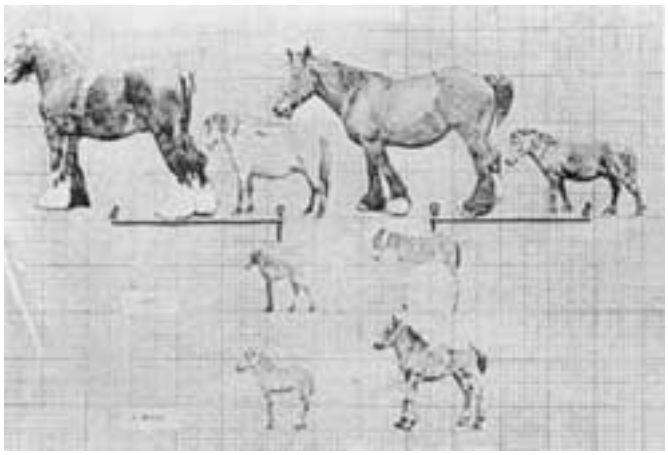


Figure 11.8 The effects of sire and mare on the size of offspring are shown in this 1938 experiment in which Shire horses and Shetland ponies were mated. The maternal influence predominates

Diagnosis

Extreme examples of fetuses that are severely small for gestational age can sometimes be diagnosed by palpation. This is most likely if the same midwife or doctor sees the woman at each antenatal visit and uses the written records of previous visits longitudinally. In several control studies false positive rates as high as 50% and low predictive values have been found in the clinical estimation of intrauterine growth restriction.

The use of symphysio-fundal height measurements is probably of more use in detecting the large baby or polyhydramnios than the small baby or oligohydramnios. A randomised trial of symphysio-fundal height mesurements was able to detect fewer small for gestational age fetuses by this method, 28% compared with 48% in the palpation group without measuring fundal height.[3]

Sometimes the lack of amniotic fluid is diagnosed more readily; oligohydramnios accompanies fetuses that are small for gestational age and therefore may lead to ultrasound investigation more swiftly than when fetal size has been estimated clinically.

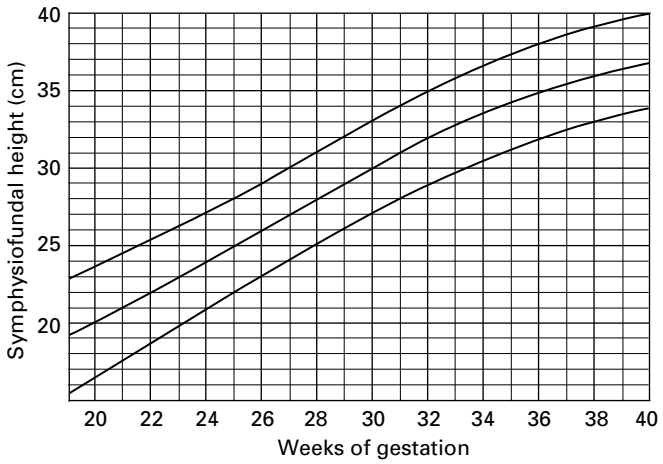


Figure 11.9 Mean (± SD) of symphysio-fundal height by weeks of gestation. Note the wide range of readings for any given week of gestation and the even wider range of expected gestation weeks for any given reading

Most fetuses that are small for gestational age are diagnosed in this country by ultrasound. When a good estimate of gestational age in early pregnancy has been obtained and fetal abnormalities have been excluded, ultrasound scans can give valuable measures of fetal growth. Scans of the abdominal circumference at the level of the umbilical vessels give a measure of liver growth. Another measure of somatic growth is femur length.

Fetuses with small abdominal circumferences can have their head circumference measured and the ratio of head to abdominal circumference derived. A small for gestational age fetus with a normal ratio of head to abdominal circumference tends to be a perfect miniature (bonsai baby) and is usually normal, representing the lower end of biological variation. Such fetuses, however, may also be associated with chromosomal anomalies, drugs, infection, and malnutrition.

Fetuses suffering from placental bed malperfusion tend to preserve growth of the head at the expense of the body because a protective mechanism shunts blood to the brain. Measuring the ratio of head circumference to abdominal circumferences can sometimes differentiate those that are just normally small (normal ratio) and those that are growth restricted (increased ratio).

Occasionally there may be a chromosomal reason for the poor growth picture and an ultrasound assessment may help determine if a karyotype is indicated. Structural anomalies such as cardiac defects, dilated renal pelves, or abnormal head shapes may be suggestive. Alternatively a history of maternal infection or increased viral antibodies may point to an infective cause.

All small babies require close assessment. Estimating fetal weight should include serial ultrasound including measurements, liquor volume, and Doppler studies of the umbilical artery. Cardiotography is used to give reassurance especially if there is doubt about fetal movements.

Small for gestational age fetuses may be screened by using early ultrasound to confirm gestational age and later to confirm growth. Finer tuning is possible by Doppler measurement of the afferent blood supply to the placental bed, with later changes in blood velocity along the umbilical vessels giving a more precise warning of fetal state. Should the umbilical artery end diastolic frequencies be lost, delivery should be considered very soon, provided pregnancy is far enough advanced that the neonatal unit of the hospital concerned is happy to deal with a child of that gestation.

In the past, most babies had ultrasound reading of the biparietal diameter at 16–20 weeks to confirm gestation and a second scan of abdominal circumference at 32–36 weeks to check growth. The later scan is now more commonly done only on suspicion of poor fetal growth and has been dispensed with in most UK obstetric departments.

Mothers whose fetuses are at greater risk of intrauterine growth restriction often have several ultrasound readings performed in later pregnancy. Such women include those with a history of perinatal death and of intrauterine growth restriction previously as well as those in whom the fetus is exposed to some of the aetiological factors already considered and where oligohydramnios may give a clue.

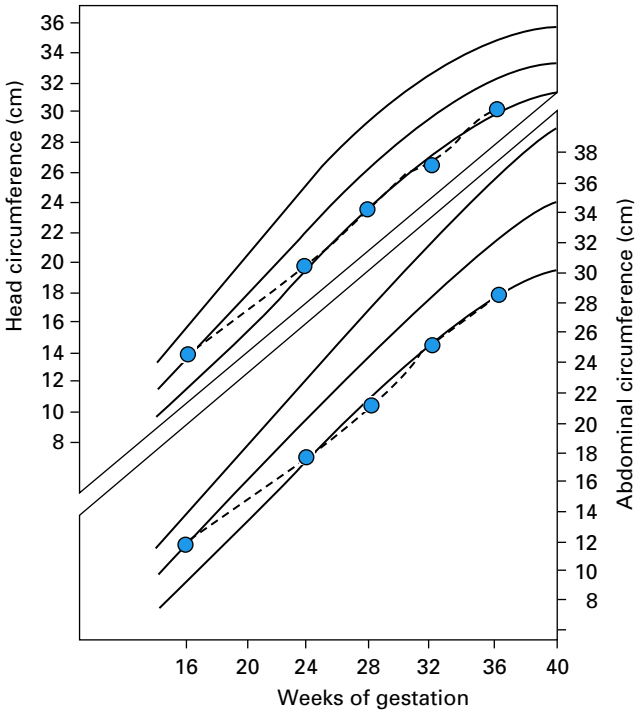


Figure 11.10 Ultrasound measures of the head and abdominal circumference. Although growth rates are diminished, they fall at the same rate—symmetrical growth restriction

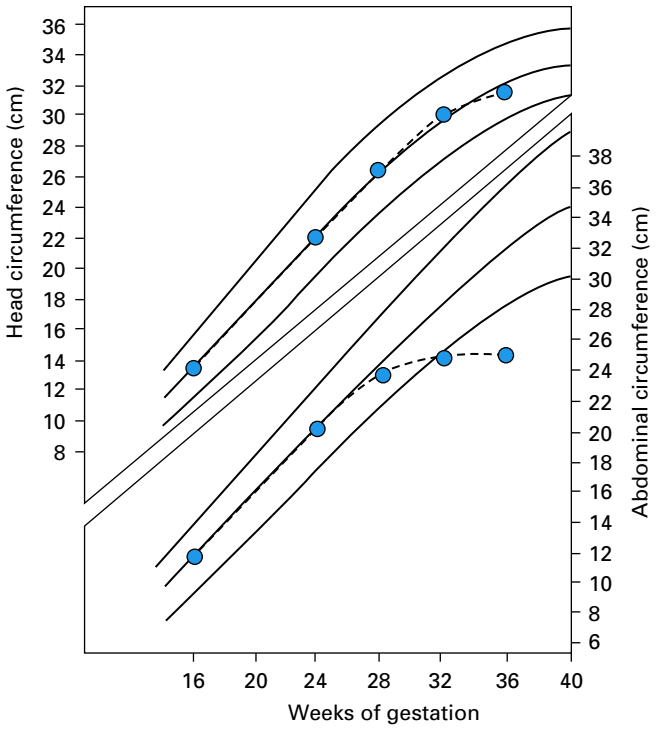


Figure 11.11 Ultrasound measures of head and abdominal circumference. Abdominal growth slows more than head growth—asymmetrical growth restriction

Treatment

The ultimate treatment of a fetus with impaired growth associated with an abnormal placental bed is delivery. Diagnosis encapsulates the fact that a baby getting insufficient nutrition

for normal growth will be in greater danger of oxygen deprivation in labour. Removal from the hostile environment would be the ultimate answer, but this might not be wise in earlier gestation (24–28 weeks); efforts are made to improve the blood supply to the placental bed.

Rest, particularly with the woman lying on her left side for some hours a day, should theoretically improve placental perfusion, but Doppler studies show little evidence for its effectiveness. Measures to restore the plasma volume and to give adequate hydration may be useful theoretically as they should decrease viscosity and lead to an improvement of intrauterine blood flow. Again, theory is not matched by practice.

In fetuses that are small for gestational age, correction and reversal of some of the causal factors might have helped, but it is too late to do this when the fetus is detectably small for gestational age. For example, curtailment of cigarette smoking should happen in early pregnancy. Such reduction in the first 16 weeks allows fetuses to follow a normal growth pattern rather than that of growth restricted babies of smoking mothers.

The mother of a fetus that is small for gestational age should attend a hospital with the capacity for more precise diagnosis and where special ultrasound and Doppler measurements are available. Many tertiary referral centres have a fetal assessment unit run on a day care basis. Women who live near large hospitals with such facilities can still be outpatients while having full surveillance. If they live away from the centre, however, they may have to be transferred and become inpatients; this is the keystone of the *in utero* transfer system widespread in the UK. Probably a third of the women admitted as *in utero* transfers have fetuses that are small for gestational age as their indication for admission.

The ultrasound surveillance of fetal growth, liquor volume, and umbilical vessel blood flow allows more precise fetal prognosis. Prospective frequent and regular consultations with the neonatal paediatrician who will be involved are essential. This will help to prepare for a premature delivery. The mother is also given steroids to reduce the risk of respiratory distress syndrome in her baby.

The fetus must be delivered at the most appropriate time by the most appropriate method. The time depends on weighing up the risks of keeping the fetus inside the uterus, that is, those of diminished placental bed perfusion, against the risks of being outside, that is, the risks of immaturity and survival in a good intensive care neonatal unit. The critical gestational age for these decisions is being pushed back all the time; now the worrying time for most obstetricians and neonatal paediatricians is 24–28 weeks. Once a pregnancy passes 28 weeks the concern is much less, although the respiratory distress syndrome can still cause morbidity and even death after delivery, especially in those small for gestational age.

The next time

Studies of pregnancies subsequent to one producing a small for gestational age baby showed that growth restriction only recurs in 20% and when it did it was less severe. In consequence, although this gives a better prognosis, it makes any management plans hard to assess for four-fifths of women will not get the problem anyway in the next pregnancy, prevention measures used prospectively thus may not have been needed. Even harder is research; use of paired studies with controls or randomised controlled trials is essential.

Figure 11.12 Woman lying in the left lateral position

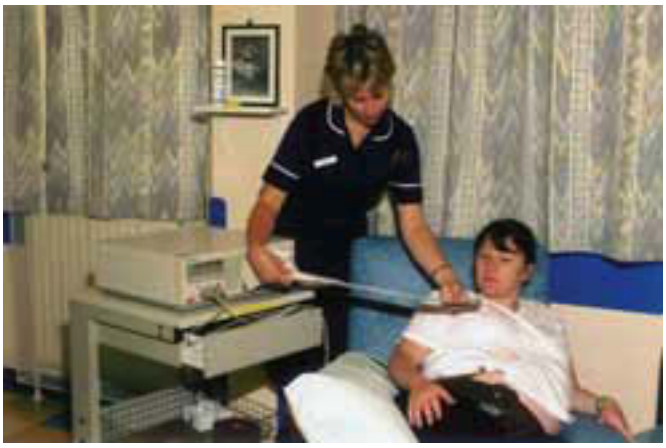


Figure 11.13 Corner of a fetal assessment unit

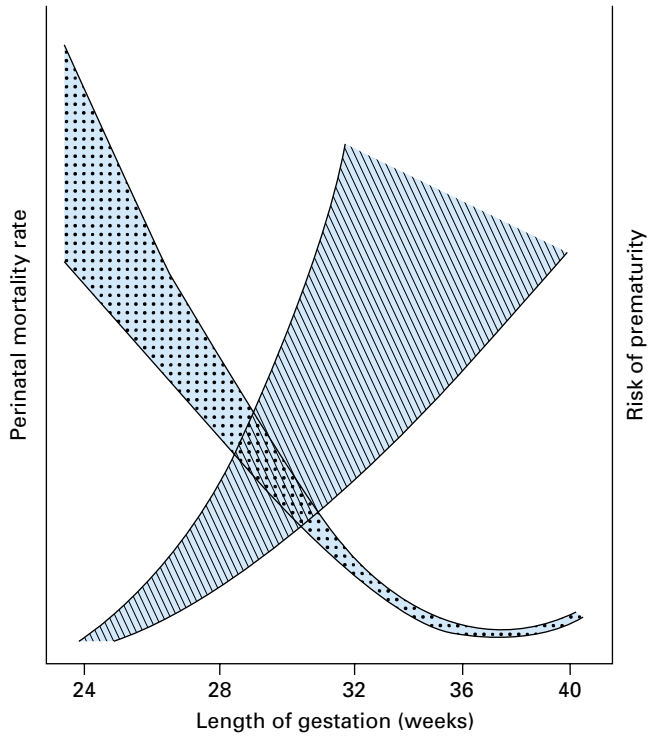


Figure 11.14 The relative risks to a fetus of staying in the uterus on the wrong side of a poor placental bed perfusion system ⊡ compared with the risks of being delivered too soon ▨

Conclusion

It must be remembered that the definitions of small for gestational age are used imprecisely and much that was thought to be known about its causation depended on data that were not mutually comparable. Until Doppler measurement, the measures of fetal wellbeing were also inexact; even Doppler ultrasound is not the last word on the subject. The ultimate management depends on avoiding trouble. Maybe we are overprotective of fetuses that are small for gestational age, but it is the best that we can do in 2001.

The diagnosis, causes, and management of small for gestational age fetuses are all still uncertain. The best management is prevention.

References

1 Hytten FE. Nutrition. In: Hytten FE, Chamberlain G, eds. *Clinical physiology in obstetrics*. 2nd edn. Oxford: Blackwell Scientific Publications, 1990:150–72.
2 Walton A, Hammond J. The maternal effects on growth in Shire horses and Shetland pony crosses. *Proc Roy Soc London [Biol]* 1938;**125**:311–35.
3 Lindhard A, Nielson P, Mooritsen L, Zachariassen A, Sørensen H, Rosenø A. Implications of introducing symphysio-fundal height measurements. *Br J Obstet Gynaec* 1990;**97**:675.

Recommended reading

- Nathanidez P, Smith G. Regulation of the myometrium in term and preterm labour. In O'Brien P, ed. *Yearbook of obstetrics and gynaecology*, no. 8. London: RCOG, 2000.
- Robinson J, Hok F, Decker G. Intrauterine growth restriction. In O'Brien P, ed. *Yearbook of obstetrics and gynaecology*, no. 8. London: RCOG, 2000.

12 Preterm labour

Preterm labour may result in the birth of an immature infant. Together with intrauterine growth restriction it is the main problem of obstetric care in the UK. The conventional definition of preterm labour includes women delivering before 37 completed weeks of gestation, but in practice in the UK problems arise mostly with births before 34 weeks. Babies more mature than this can be cared for successfully in many district general hospitals without intensive care facilities; most problems arise in babies weighing less than 1500 g (3·5 lb).

Perinatal mortality rates relate sharply to maturity and birth weight; similarly, neonatal mortality rates relate to weight at birth. Probably some 6% of babies in the United Kingdom are born before 37 weeks and 2% before the 32nd week of pregnancy.

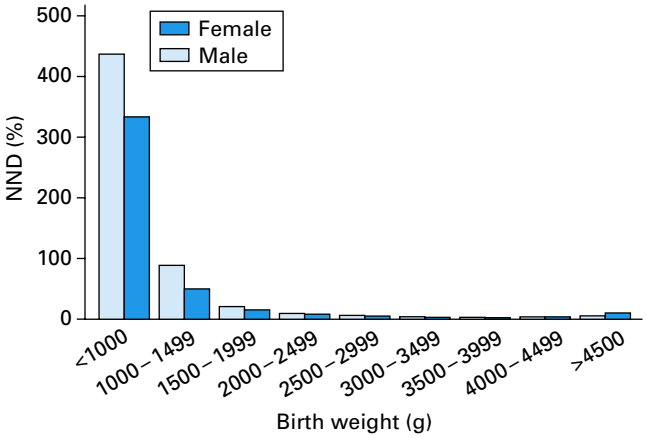


Figure 12.2 Neonatal deaths per 1000 live births by sex and birth weight, England and Wales 1995

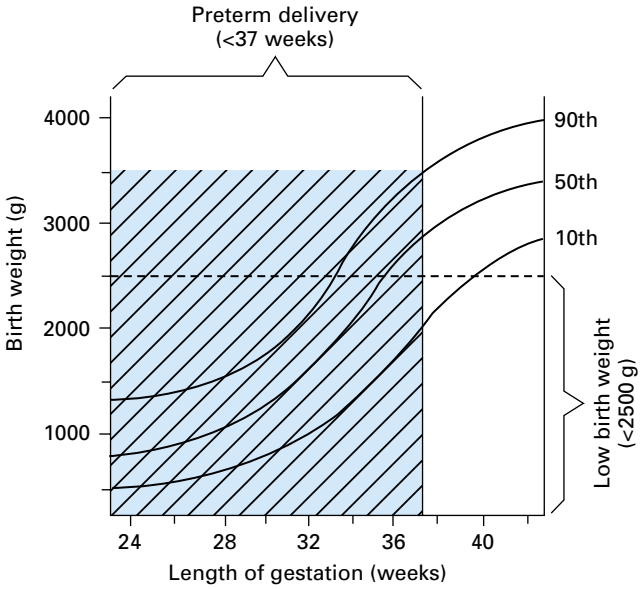


Figure 12.1 Relation between length of gestation and birth weight. Babies born in the crosshatched area are preterm irrespective of weight

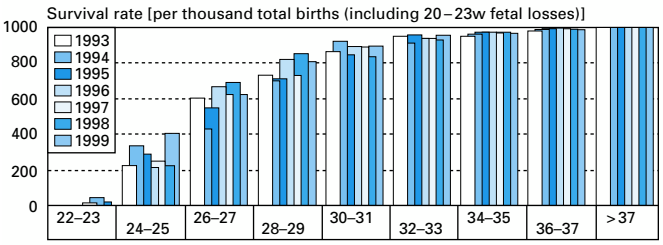


Figure 12.3 Birth weight specific survival to 28 days 1993–9 (Welsh Perinatal Survey)

Causes

Sociobiological background

The capacity for preterm labour is often predictable by a clustering of high risk factors. The mother's age, parity, and socioeconomic class bear strong associations with preterm labour. Socioeconomic class is an indicator of the woman's behaviour, nutrition, smoking, and previous preterm delivery. Also important is a woman's work in early pregnancy, particularly if it involves continuous standing. These may not be individual factors in their own right but are useful to identify women whose risk of preterm labour is increased.

Reproductive history

A multiparous woman's obstetric history may give prognostic clues; the chances of a preterm delivery are tripled after one previous preterm birth and increased sixfold after two. These are two simple sets of risks; other outcomes bring in differential variables. Past studies have been diminished by not including the woman's total obstetric history, which needs careful consideration in the case of each woman.

Medical history

Recurrent lower urinary tract infections are not usually associated with recurrent preterm labour, although pyelonephritis may be. The renal tract should be investigated

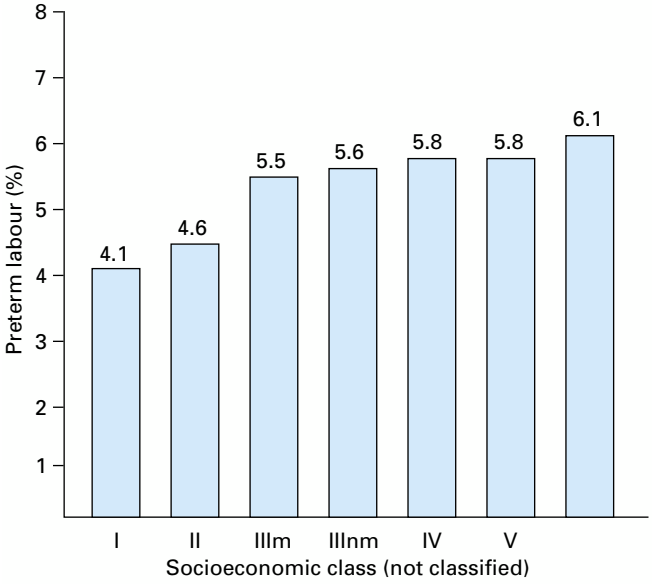


Figure 12.4 Singleton births before 37 weeks by socioeconomic class (Scotland 1990–95)

between pregnancies and reinfection prevented by prophylactic antibiotics in future pregnancies.

Uterine structural abnormalities can be a recurrent factor, the best documented being cervical incompetence. This follows damage or overstretch of the cervical internal os—an ill-formed muscular sphincter—and is a mechanical diagnosis first observed by obstetricians in the 1940s and made familiar by the work of Shirodkar and McDonald in the 1950s. The truer picture of the place of cervical incompetence and its management in preterm labour had to await a randomised controlled trial in the 1980s run jointly by the Medical Research Council and the Royal College of Obstetricians and Gynaecologists; the results put into proportion the importance of cervical incompetence as an individual factor in preterm delivery.[1] A much smaller report had more optimistic results (Table. 12.1).

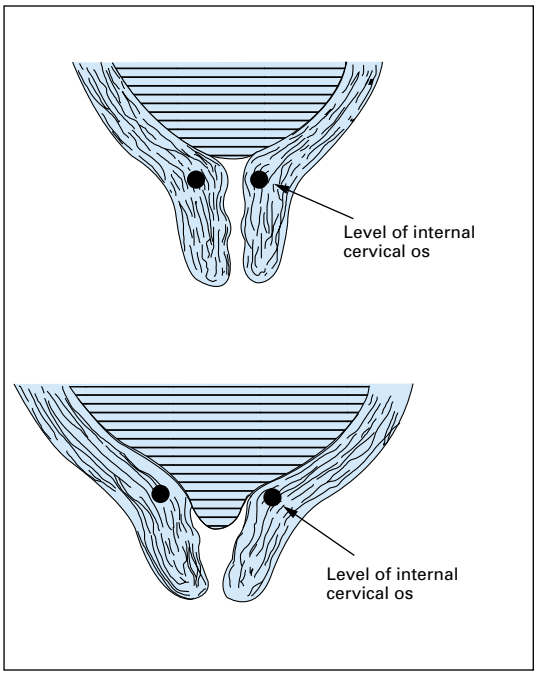


Figure 12.5 Cervical incompetence leads to a cone of unsupported membranes

Complications of pregnancy

Multiple pregnancy is a marker for preterm labour. The mean gestation of twins is 37 weeks and therefore many will be born before this time.

Several studies have shown the association of preterm labour with antepartum haemorrhage, irrespective of the cause of the bleeding.

Hard physical work in pregnancy is associated with preterm labour, particularly if it is repetitive and boring or in an unpleasant, noisy environment. This factor is discussed in Chapter 6.

Abnormalities of the fetus are often associated with preterm labour when there may also be polyhydramnios, which in itself can lead to premature membrane rupture and preterm labour.

Table 12.1 Effectiveness of cervical cerclage in reducing preterm delivery and rates of stillbirth, miscarriage, and neonatal death[2]

	No in group		Odds ratio and 95% confidence interval
	Experimental	Control	0.01 0.1 0.5 1 2 10 100
Delivered:			
Before 33 weeks	59/454	82/451	
Before 37 weeks	124/454	146/451	
Stillbirth, miscarriage, or neonatal death	37/454	54/451	

Data analysed 1993.

Infection and premature membrane rupture

Infection of the lower uterus and the membranes is an important feature that is poorly investigated epidemiologically. The presence of micro-organisms in the membranes is associated with an increased production of prostaglandins, one of the main factors associated with the onset of labour. Proteases, coagulases, and elastases are also produced by invading micro-organisms, whose endotoxins may stimulate labour directly as well as through prostaglandin metabolism. Low grade chorioamnionitis (infection of the membranes) is much commoner after premature rupture of the membranes, when an ascending infection from the vagina may produce such biochemical changes. One of the commonest organisms is the β haemolytic streptococcus, which is found as a commensal in the vagina in about 5% of women but may be a factor in preterm labour in up to 20%. Other anaerobic vaginoses are more common in women with premature rupture of

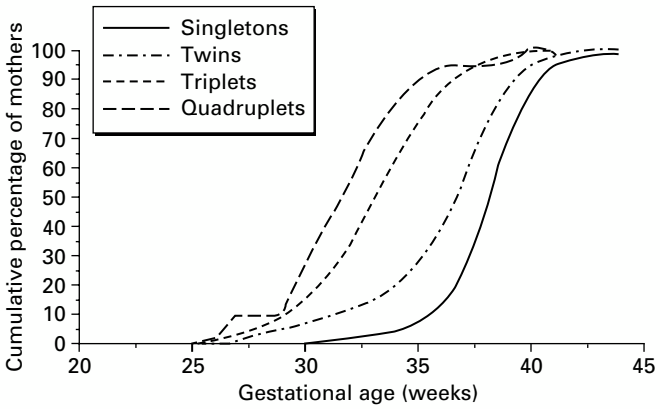


Figure 12.6 Cumulative gestational age at delivery for multiple pregnancies. By 37 weeks delivery has occurred in 97% of quadruplet, 96% of triplet, 55% of twin, and 25% of singleton pregnancies

membranes and preterm labour than in those who go to term. Changes in vaginal flora have been shown by culturing vaginal flora in mid-trimester and before labour. The persistence of *Bacteroides* in the last months of pregnancy was associated with preterm birth. Coliforms when associated with preterm labour were acquired later in pregnancy.[3] Premature membrane rupture itself is a commonly quoted cause of preterm labour; this may be due to the infection that weakened the forewaters or to the removal of the forewaters' mechanical support from the cervix.

Induction

Preterm labour is caused iatrogenically in about 15% of women in this country, though in some units the rate rises to 40%. Many of the inductions will obviously be in women after 37 weeks' gestation but some will be performed before this. The problem of rhesus incompatibility, previously a major indication, has reduced markedly; in its place is pregnancy-induced proteinuric hypertension and intrauterine growth restriction. Women with either of these problems should be delivered at hospitals that can cope with the neonatal sequelae of such induction; these groups produce a large proportion of babies who are born well before the 37th week of pregnancy.

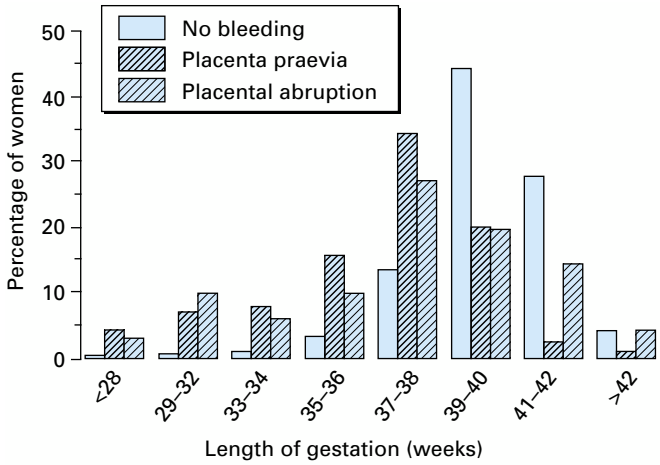


Figure 12.7 Time of delivery among women who had no vaginal bleeding in pregnancy compared with that among those with placenta praevia or placental abruption (n = 17 005)

Prevention

The recognition of some of the triggers of preterm labour has led some obstetricians to take action to prevent labour. There is little objective evidence that bedrest and the use of prophylactic tocolytic agents are helpful, although a doctor might use either of these managements to satisfy a mother who has previously undergone preterm labour and has faith in them. Repeated, carefully taken, high vaginal swabs to give the pattern of micro-organisms in the upper vagina may be useful. Active antibiotic treatment will eradicate colonisation and thus reduce the risks of preterm labour. Clindamycin is being used in this field and the whole approach is under evaluation.

Several centres have used programmes during early and mid-pregnancy to educate women with a history of preterm delivery to try to prevent a recurrence of the problem. There is no easy method of doing this in a group; the success of such programmes depends on the individual woman and her individual midwives and doctors. All the factors discussed so far must be considered, and the woman should obviously try to avoid those which seem to be the more relevant in her case. Even with the most intensive antenatal education programmes, preterm deliveries are not cut to less than about 3·5%, a background rate in many populations. Success in this subject may come eventually after a conscious effort to modify the lifestyle, socioeconomic conditions, and medical problems of each individual patient.

Box 12.1 Major indications for induction of preterm labour

- Pregnancy-induced proteinuric hypertension
- Intrauterine growth restriction

Box 12.2 Prevention of preterm labour

- Control vaginal infection
- Education about early signs of labour
- Cervical cerclage (if relevant)
- ? Bedrest
- ? Prophylactic tocolytics
- Better control of conditions that would require early iatrogenic induction of labour

Diagnosis

As with labour at term, diagnosing the onset of preterm labour is more easily performed retrospectively than at the time. You can look back and say a labour probably started at a certain time, but to do so prospectively is much harder. The general practitioner is left with the difficult task of deciding whether any group of uterine contractions will progress to cervical dilatation or whether they are just stronger Braxton Hicks contractions. The diagnosis may be assisted by external tocographic measurement of uterine contractions with a semiquantitative external monitor. Any woman thought to be in preterm labour should go to the local maternity unit as soon as possible for further assessment. There tocography may help, and assessment of the cervix may be valuable. About half of the women who present with regular, painful contractions will not

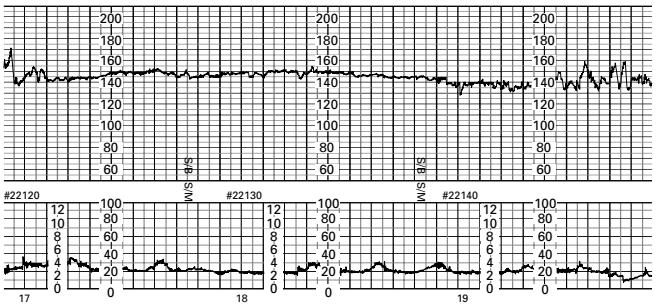


Figure 12.8 Cardiogram with fetal heart rate (above) and uterine pressure (below). The fetus is asleep for much of this trace and regular small contractions of the uterus are seen

progress to labour. Impending preterm labour may disrupt the cells of the cervix releasing fibronection. The cells can be detected fairly simply and are the basis of a test to check whether that woman is at risk of going into preterm labour soon. It is more useful in the early treatment of high-risk women but of less use in those at lower risk.

If preterm labour seems inevitable, treatment may be given to postpone it. Otherwise, the woman may be kept in for 24 hours to see if labour follows; if not, she can be discharged home to the care of her general practitioner. It is easy to be wise after the event, but only by sending every woman about whom there is reasonable doubt to the maternity unit will clinicians not miss the occasional woman who goes into very early preterm labour. Ultrasound evaluation of the cervix can sometimes warn of preterm labour. In some women dilatation, funnelling of the internal os and shortening of the canal all can be demonstrated but the specificity of the investigation is not high. A vaginal examination still gives the clearest evaluation of preterm labour.

Before 32 weeks' gestation short-term inhibition of labour allows:

- **Transfer to the delivery unit best equipped for special neonatal care**
- **Steroids to be given to help mature the fetal respiratory system**

A third of children born between 32 and 35 weeks' gestation may expect to have problems at school by the age of seven.

Inhibition of established preterm labour

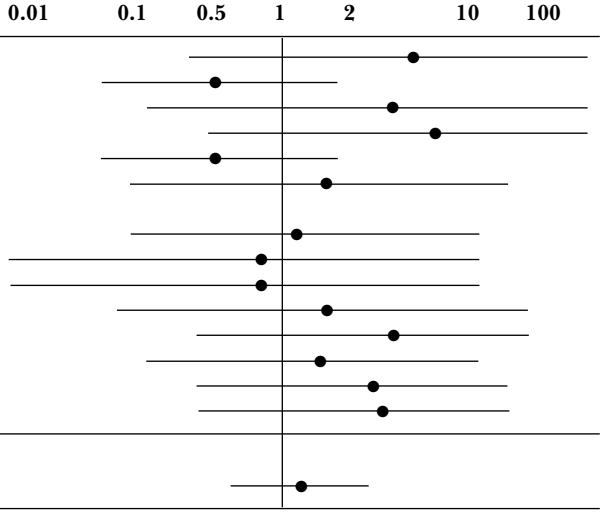
If a woman is in real preterm labour a decision has to be made whether labour should be stopped. It is probably wise not to do so if the mother's blood pressure is raised, there is proved infection in the endocervical or decidual regions, or the fetus has a lethal abnormality. Some obstetricians would consider further that it was unwise to inhibit labour in the presence of long-term rupture of the membranes, severe intrauterine growth restriction, or an antepartum haemorrhage. Each of these cases must be decided on their own merits.

Other than these exceptions, in most cases before 28 weeks it is worth trying to stop preterm labour to buy intrauterine time for the fetus. In the short term this can allow emergency treatment such as steroids to help maturation of the fetal respiratory system or allow transfer of the woman to a centre with good neonatal care facilities for a very small baby after delivery. These decisions must be made in consultation with the paediatricians as the practical management of any baby resulting from a preterm labour will depend on their skills and facilities. In a large well equipped obstetric-paediatric unit the borderline comes at about 27 weeks, provided that all other features of the pregnancy are normal.

If it is considered necessary to stop preterm labour a range of agents exist. Alcohol and the progestogens are obsolete.

Table 12.2 Effectiveness of β mimetic tocolytics used in preterm labour in reducing preterm delivery. The numbers are the proportions of women delivering before 37 weeks[2]

Study	No in group		Odds ratio and 95% confidence interval
	Experimental	Control	0.01 0.1 0.5 1 2 10 100
Christensen *et al.* (1980)	1/14	0/16	
Spellacy *et al.* (1979)	1/15	4/15	
Barden (unpublished)	1/12	0/13	
Hoebel (unpublished)	2/17	0/16	
Cotton *et al.* (1984)	1/19	4/19	
Howard *et al.* (1982)	1/16	1/21	
Ingemarsson (1976)	0/15	0/15	
Larsen *et al.* (1986)	1/49	2/50	
Calder and Patel (1985)	0/37	1/39	
Scommegna (unpublished)	0/16	1/17	
Mariona (unpublished)	1/4	1/5	
Wesselius-De Casparis *et al.* (1971)	2/33	1/30	
Leveno *et al.* (1986)	2/56	3/55	
Larsen *et al.* (1980)	11/131	2/45	
Adam (1966)	9/28	7/24	
Typical odds ratio and 95% confidence interval			

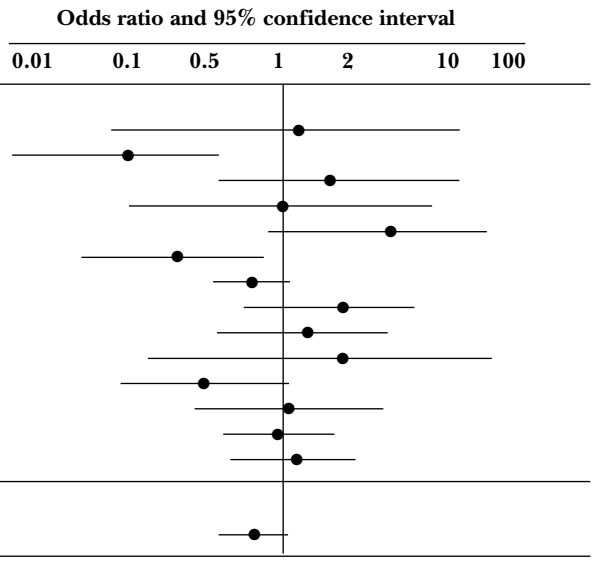


Prostaglandin synthesis inhibitors (such as indomethacin) are effective but may have side effects in mother and fetus. Both β antagonists (ritodrine and salbutamol) and calcium antagonists (nifedipine) are used in the United Kingdom to suppress labour. These drugs work equally well on the myometrial cells and may postpone labour for a time. There is little evidence from meta-analyses of many studies that they reduce either perinatal mortality rates or postponement of labour over a long period.[2] Their use will depend on the ripeness of the cervix—the less ripe the more likely that action will be effective. They are best used before 32 weeks of pregnancy and probably work better in the absence of infection. Though little evidence shows that prophylactic oral β mimetic agents prevent preterm labour, oral maintenance after intravenous inhibition has some effect. Because of this, the recent introduction of oral nifedipine has been popular. The side effect profile is better, with occasional flushing, palpitations, and transient hypotension.

Once treatment with tocolytic agents has been started, the next decision is where the woman is to deliver if labour

Table 12.3 Effectiveness of β mimetic tocolytics used in preterm labour in reducing perinatal death. The numbers are the proportions of perinatal deaths[2]

Study	No in group: Experimental	No in group: Control
Christensen *et al.* (1980)	14/14	16/16
Spellacy *et al.* (1979)	12/14	13/15
Barden (unpublished)	6/12	13/13
Hoebel (unpublished)	10/16	8/15
Cotton *et al.* (1984)	15/19	16/19
Howard *et al.* (1982)	9/15	5/18
Ingemarsson (1976)	3/15	12/15
Larsen *et al.* (1986)	14/49	23/50
Calder and Patel (1985)	23/37	19/39
Scommegna (unpublished)	10/15	10/16
Mariona (unpublished)	3/4	3/5
Wesselius-De Casparis *et al* (1971)	13/33	21/30
Leveno *et al.* (1986)	40/54	42/52
Larsen *et al.* (1980)	65/131	21/45
Sivasamboo (1972)	14/33	20/32
Typical odds ratio and 95% confidence interval		



Note the small numbers and confidence intervals in some of the studies in this meta-analysis and the one on p. 75.

proceeds. If the unit cannot cope with very small babies, *in utero* transfer must be considered. The woman should go to a tertiary referral centre in the region that can manage babies of this degree of immaturity. The alternative philosophy is to allow the baby to be delivered in the peripheral centre and, if necessary, transfer the child to the tertiary referral unit by *ex utero* transfer. *In utero* transfer may not be necessary every time; it is used as a precaution but it allows the woman to be in the tertiary referral centre that is able to provide more sophisticated obstetric as well as neonatal care, for example Doppler flow studies. *Ex utero* transfer allows the woman to stay closer to her home at the local hospital she has chosen. However, specialist antenatal tests may not be available, obstetricians may not be as experienced in the delivery of very small babies, and expert paediatric teams may not be available at the time of delivery because of the many other calls on obstetricians' and paediatricians' time. In addition, with road traffic conditions in the UK there is no guarantee that help can get to even the nearest district hospital quickly. At present the

Box 12.3 Expert care for babies expected to be very small

- *In utero* transfer to obstetric/neonatal referral centre
- Delivery in district general hospital and *ex utero* transfer to specialist centre

philosophy is in favour of *in utero* transfer, but it may not stay so for long in the reorganised NHS.

Steroids, given to the mother before delivery, pass across the placenta to the fetus. Between 26 and 34 weeks' gestation, they have been shown to decrease morbidity and mortality associated with the respiratory immaturity of preterm delivery.[4] They are of optimal use if more than 24 hours passes from the first dose to delivery but often there is not enough time from the admission of the woman to her inevitable delivery. The use of tocolytic agents is used to postpone delivery and so extend the time available for the steroids to work on the fetal lung helping to produce surfactant. The steroids are commonly given as betamethasone in two intramuscular doses at 12-hour intervals. If labour is delayed by more than a week and the woman is still less than 34 weeks' gestation, the course of steroids is often repeated although the evidence of the benefits of this is hard to show.[5]

Conclusion

This and the previous chapter are concerned with the most serious problems of current obstetrics. Getting the best results for very small babies is the most hopeful line of advance at present. It needs coordination from family doctors, obstetricians, midwives, and neonatal paediatricians with individual treatments tailored to individual mothers.

The data for perinatal and neonatal mortality rates for Scotland are taken from the Scottish stillbirths and neonatal deaths report produced by the Information Office of the Scottish Health Service. The figure showing the cumulative distribution of singleton and multiple births is reproduced by permission of the Office of Population Censuses and Surveys from *Three, four and more*, published by HMSO. Other data come from CESDI reports.

Preterm labour and small for gestational age fetuses constitute the most serious current problems in obstetrics.

Reference

1 MacNaughton C, Chalmers T, Dubowitz V, *et al.* Report of the MRC/RCOG Multicentre randomised trial of cervical circlage. *Brit J Obstet Gynaec* 1993;**100**:516–20.
2 Chambers l, Enkin M, Keirse MJNC, eds. *Effective care in pregnancy and childbirth.* Vol 2. *Childbirth.* Oxford: Oxford University Press, 1989:705, 707.
3 McDonald H, Loughlin J, Jolley P, *et al.* Changes in vaginal flora during pregnancy and association with preterm birth. *J Infect Dis* 1994;**170**:724–8.
4 Crowley P, Chalmers I, Kierse M. The effects of cortical steroid administration before preterm delivery. *Br J Obstet Gynaec* 1990;**97**:11–25.
5 RCOG. *Antenatal corticosteroids to prevent respiratory disease syndrome.* RCOG Guideline No. 7 London: RCOG, 1996.

Recommended reading

- Chamberlain G. Antenatal care of the very small baby. In Harvey P, Cooke R, eds. *The baby under 1000* g. Oxford: Butterworth Heinemann, 1999.
- Hyert J, Thilaganathan B. Obstetrics management at the limits of neonatal viability. In Studd J, ed. *Progress in obstetrics and gynaecology*, no. 14. London: Harcourt Brace, 2000.
- RCOG. *Antenatal corticosteroids to prevent respiratory distress syndrome. Guidelines no. 7.* London: RCOG, 1996.

13 Multiple pregnancy

Multiple pregnancy is a mixed blessing. On the one hand is the instant family, on the other are the increased perinatal mortality and morbidity as well as a much greater load for the mother after delivery.

Types

Multiple pregnancy follows either the division of an oocyte fertilised by one sperm into two separate bodies (identical or monozygotic twins) or the fertilisation of more than one egg by separate sperm (non-identical or dizygotic twins). In higher multiple pregnancies than twins a combination of these two mechanisms happens.

In monozygotic twins, division into two separate bodies was thought to occur only at a very early stage but it can in fact take place up to several days after fertilisation. The later this is, the more likely is the rare abnormality of conjoined twins.

Prevalence

The prevalence of twin births in the UK is 11.3/1000 deliveries, of triplets 0.3/1000, and of quadruplets about 0.01/1000 deliveries. There is a natural variation between races; Japanese women have one of the lowest rate of twins and those from some African countries have a much higher rate, up to one in 30 deliveries. Multiple pregnancies also increase with maternal age. These biological variations are due to an increase in the dizygotic twinning rates, based on the capacity of the woman to produce more than one oocyte at the time of ovulation.

The prevalence of multiple pregnancy has been increasing in the UK in the past decade. For higher multiples than twins the rate trebled from 12 per 100 000 to 40 per 100 000 between 1980 and 1998. Though a part of this is due to the increasing number of mothers over 35, the iatrogenic effect of ovarian stimulation and *in vitro* fertilisation programmes is also important. Concern about this led to the formation of a statutory body, the Human Fertilisation and Embryology Authority which made recommendations about the maximum number of oocytes or embryos transferred at assisted fertilisation, a limit of two.

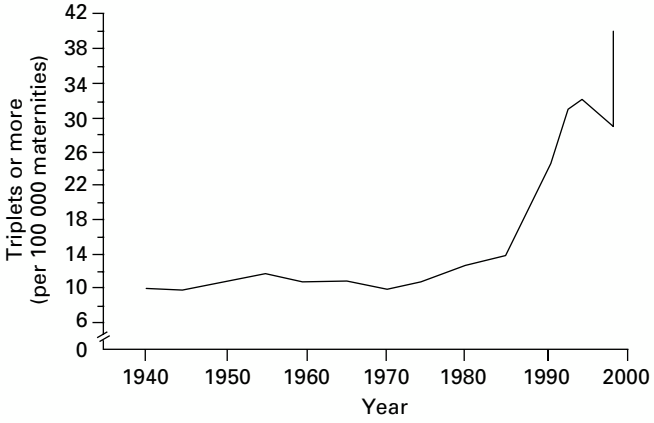


Figure 13.3 Proportion of maternities resulting in triplets (England and Wales, 1939–98)

Diagnosis

Twin pregnancies used to be diagnosed clinically when the woman reported her symptoms of pregnancy were worse than usual and the uterus was found to be bigger than would be

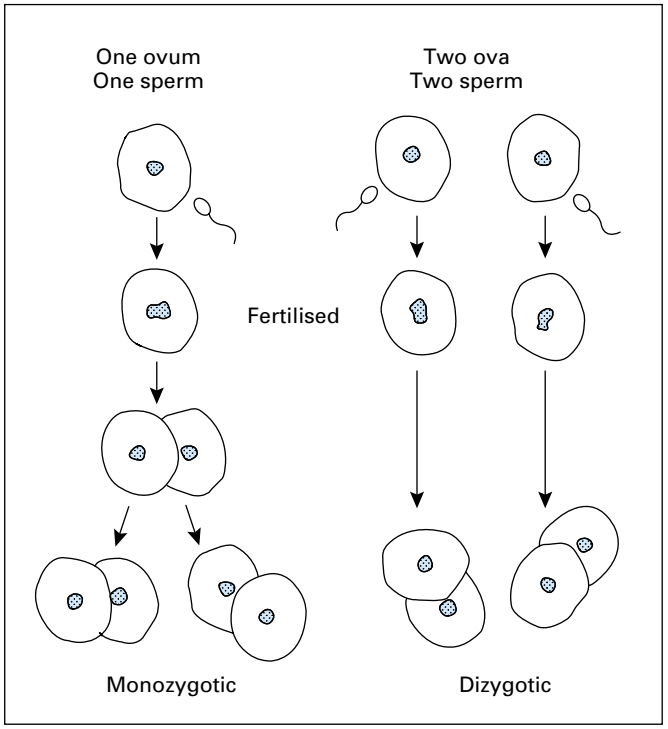


Figure 13.1 Monovular and binovular twins

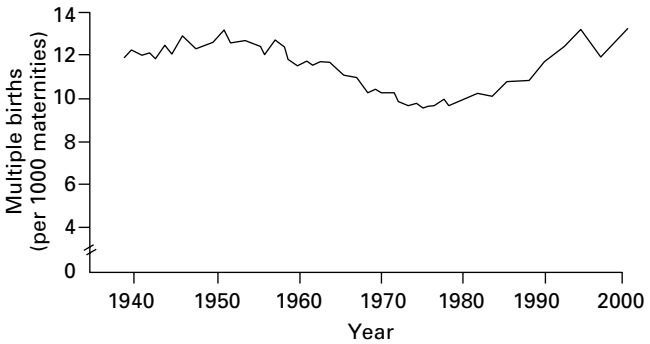


Figure 13.2 Proportion of maternities resulting in multiple births (England and Wales, 1939–98)

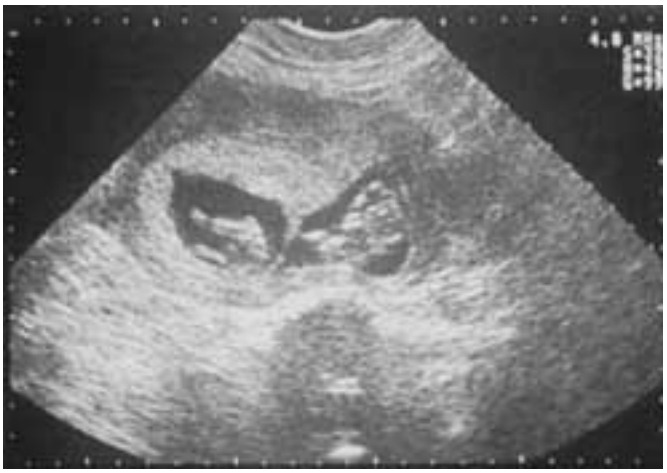


Figure 13.4 In early pregnancy twin sacs and embryos can be shown by ultrasonography

expected from gestational dates (after about 20 weeks); sometimes twins were diagnosed for the first time in labour. Often the fetal parts are hard to determine but palpation of more than two poles is suggestive of twins. Now in the UK most women have an ultrasound scan by 16 to 20 weeks and so multiple pregnancies are usually diagnosed much sooner. In the Scottish twin survey, 70% of multiple pregnancies were diagnosed by ultrasonography before 20 weeks and 95% in all were diagnosed in the antenatal period. The rest were diagnosed in labour. When a twin pregnancy is diagnosed by ultrasonography the increased incidence of congenital abnormalities should be remembered and a thorough ultrasound examination of each fetus performed between 20 and 24 weeks.

The increased uterine size leads to greater pressure on venous return. The frequency of the group of conditions that obstetricians (but not women) call minor problems (for example, varicose veins in the leg) is increased. Furthermore, the woman may have more symptoms of nausea in early pregnancy associated with the higher human chorionic gonadotrophin concentrations.

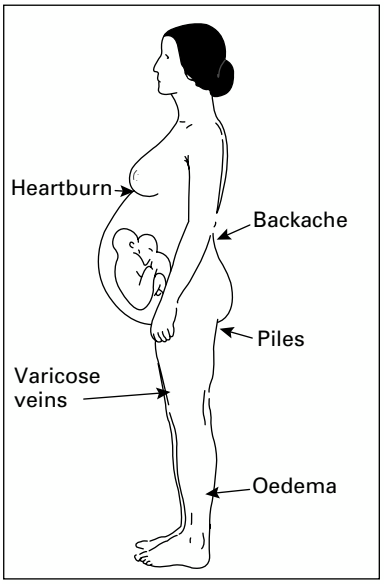


Figure 13.5 Changes that may follow an overdistended uterus with a multiple pregnancy

Death of one fetus

When one of a pair of twins dies *in utero* there is a risk to the mother of coagulopathy. For the surviving fetus there is also a risk of neurological lesions, preterm delivery with its problems of immaturity, and even intrauterine death. In very early pregnancy the complete absorption of the fetus that dies is usual (the vanishing twin phenomenon) probably happening in 5–8% of twin pregnancies. When fetal death comes later, it is best managed expectantly with close surveillance of the mother and the remaining fetus. In a dizygous twin pregnancy, the risks to the surviving fetus are relatively low.

One previously underconsidered feature of this problem has been the disaccord of the mother's reaction in grieving for one baby whilst looking forward hopefully to the birth of the other.

Congenital abnormalities

Many congenital abnormalities are more frequent in twins, especially those who are monozygous. Neural tube defects, heart abnormalities, and the incidences of Turner's and Klinefelter's syndromes are all increased. About twice as many live births from multiple pregnancies have a major congenital abnormality compared with singleton pregnancies.

Some of these abnormalities may be detected by ultrasound, others require amniocentesis. In multiple pregnancy this test is associated with a 3% rate of miscarriage compared with about 0.5% in singleton pregnancies. Care must be taken to identify the fluid from each sac, by proper labelling of the sample container, as the abnormality may be in one fetus only. Should severe abnormality be found in one fetus of a multiple pregnancy with two sacs the obstetrician may consider that the normal fetus is at increased risk and recommend selective fetocide. This can be by cardiac puncture, intravenous injection of potassium chloride or clipping the umbilical cord using a hysteroscope. Such management should be at a regional centre well used not just to performing these procedures but to the very important counselling that goes on before and after such an event. The risk of preterm labour in the unaffected pregnancy is increased.

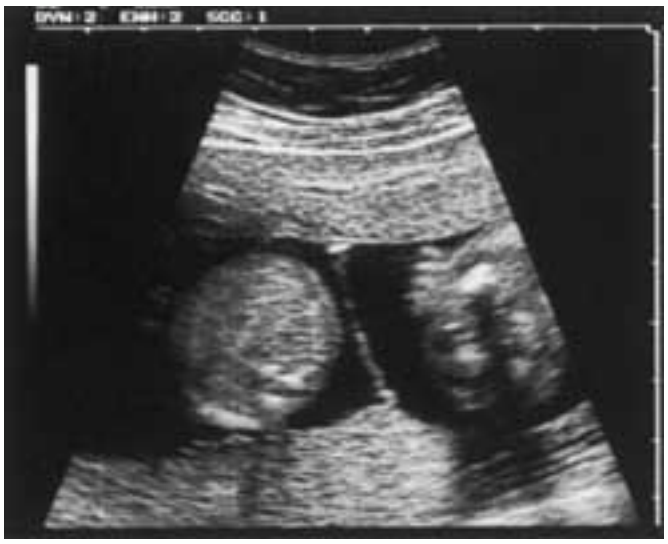


Figure 13.6 Twin fetuses at 16 weeks just before amniocentesis. Twin sacs are easily seen

Pregnancy-induced hypertension

The incidence of pregnancy-induced hypertension is increased in multiple pregnancies and eclampsia is also commoner. Antihypertensive treatment should be used as in any other pregnancy complicated by proteinuric hypertension

(see Chapter 9), and the ultimate treatment of delivery may be required earlier than for singletons, a more difficult decision as preterm twin babies fare less well than do preterm singletons. A caesarean section is more frequently needed.

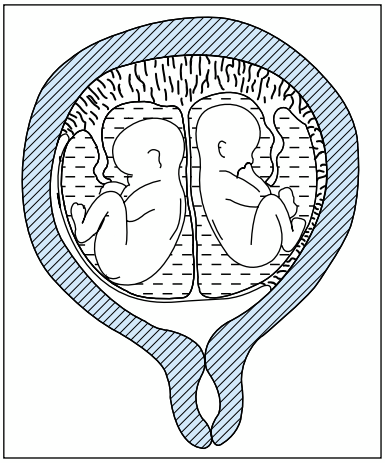


Figure 13.7 The greater area of placental implantation on the left of the uterus means that it may encroach on the lower segment

Anaemia

Commonly anaemia is reported as being more frequent in multiple than in singleton pregnancies. Some of this is due to the greater expansion in maternal blood volume with twins whereby the plasma increases more than the red cell bulk, so lowering the haemoglobin concentration. If the mean corpuscular haemoglobin concentration is used as a measure, anaemia is not more frequent in multiple pregnancies than in singletons provided that adequate nutrition and iron and folate intake are maintained. Greater demands of the growing fetuses for folate have led to some reports of megaloblastic anaemia, so folate supplements are commonly given.

Antepartum haemorrhage

Antepartum haemorrhage would be thought to be commoner in multiple pregnancy because of the greater surface area of the placental bed. The Aberdeen twin data set showed rates of antepartum haemorrhage in twin pregnancies to be 6% compared with 4.7% in singleton pregnancies ($p = <0.05$). Much of this difference, however, was made up of antepartum haemorrhage from unknown origin; only a few were caused by placental abruption or placenta praevia.

Intrauterine growth restriction

The growth of each fetus in multiple pregnancies mirrors that of the singleton until about 24 weeks of gestation; thence growth rates for most twins are still as for singletons but occasionally one or both may show a decrease. This is difficult to detect on clinical examination for polyhydramnios may cause imprecision in estimation of fetal size. Repeated serial ultrasound estimations of fetal size are the most useful way to check growth by plotting measurements of individual fetuses longitudinally through pregnancy. These data are not very different from the standard head or abdomen growth curves from singleton pregnancies until the last weeks. The estimation of fetal weight by various formulas based on the diameters of the fetus are not as useful in twin pregnancies as in the singleton.

Twin-to-twin transfusion

Twin-to-twin transfusion may be suspected when there is gross discordance in growth of a pair of twins or if there is

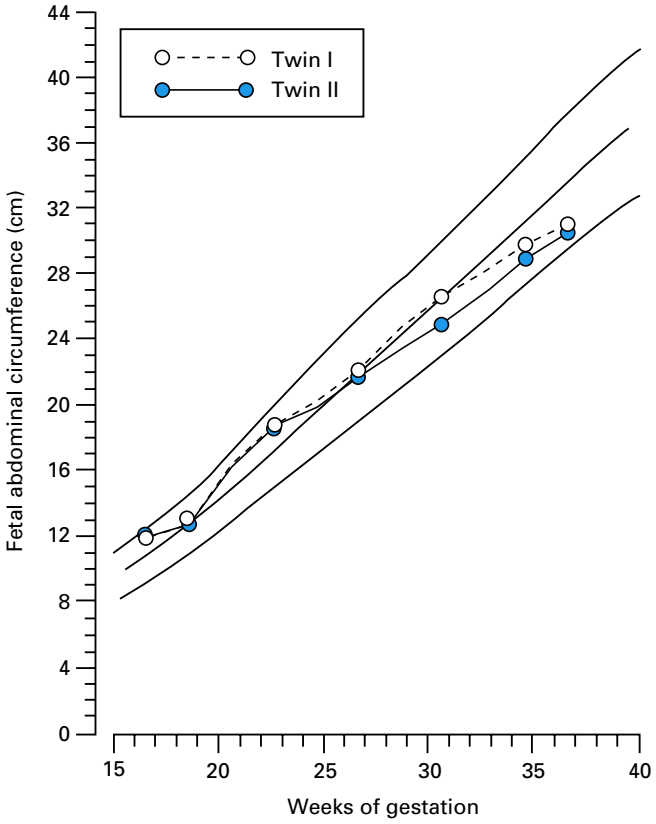


Figure 13.8 Growth of non-identical twins through pregnancy set against the mean (2 SD) singleton growth curve

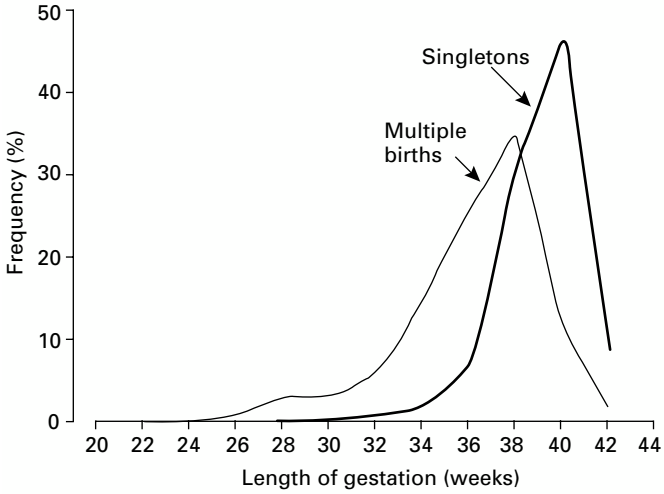


Figure 13.9 Percentage frequency distribution of length of gestation for singleton and twin pregnancies

polyhydramnios with one twin and oligohydramnios with the other. About 20% of monozygotic twin pregnancies have vascular connections between the two circulations inside the placenta but the imbalance of flow occurs in few of these, probably only 1% of such pregnancies. Clinical signs of stealing circulating blood from one twin by the other occurs less commonly and usually associated with a change in amniotic fluid volume which is quite obvious with ultrasound. If twins of the same sex are diagnosed by about 20 weeks gestation, careful ultrasound examination is made to determine chorionicity by assessing the number of discrete placental masses and the thickness of the membranes and their angle of approach to the decidual bed—the Lambda sign. While this warns of potential anastomoses, it is not of necessity grounds for fetocide, for sometimes the twin-to-twin transfusion can be compensated. It should lead to extra surveillance of both mother and fetus for the rest of the pregnancy.

Laser ablation of the communicating vessels of the placenta with intrauterine amnioscopes and narrow beam YAG laser is used; amnioreductions by amniocentesis may also be helpful.

Onset of preterm labour

The median gestation for human singleton pregnancy is just over 40 weeks whereas that for twins is 37 weeks and that for triplets about 33 weeks. The commonest single cause of perinatal mortality in multiple pregnancies is low birth weight. Though intrauterine growth restriction might also be present, birth weight is low mostly because of a preterm delivery. A measure of this problem is seen in the 30% of all liveborn triplets and 60% of liveborn quadruplets who have to stay in a neonatal intensive care unit for more than a month after delivery. The incidence of preterm labour (before 37 weeks) in twin pregnancies ranges from 20 to 50% compared with from 5 to 10% in singleton pregnancies.

An important part of antenatal care for multiple pregnancy is trying to detect those women who are likely to go into early preterm labour and prevent this if possible; if not, ensuring that they are delivered in the correct surroundings with neonatal unit facilities to look after immature babies. Some obstetricians find the examination of the cervix from 28 weeks gives a clue to its increasing ripeness (length, firmness, and dilatation). This seems to be of more use in primiparous than multiparous women. Others assess the cervix with ultrasound, endeavouring to predict early labour.

An essential element lies in informing the mother; antenatal education of women with twins about the signs of early preterm labour may be helpful.

The greater stretch of the myometrium imposed by multiple pregnancy increases the risk of preterm labour and several measures have been tried in the antenatal period to prevent this. Sympathomimetic drugs such as ritodrine have been given prophylactically, but most controlled trials have shown no benefit of this in twin pregnancy. Cervical cerclage inserted when a twin pregnancy is diagnosed does not seem to confer any increased benefits. Some consider that coitus may tip the balance in a woman who is on the edge of going into preterm labour, because of both the mechanical stimulation and the release into the vagina of prostaglandin-rich fluid. The avoidance of coitus in later pregnancy by women with twin pregnancies, however, does not seem to be associated with any significant prolongation of gestation.

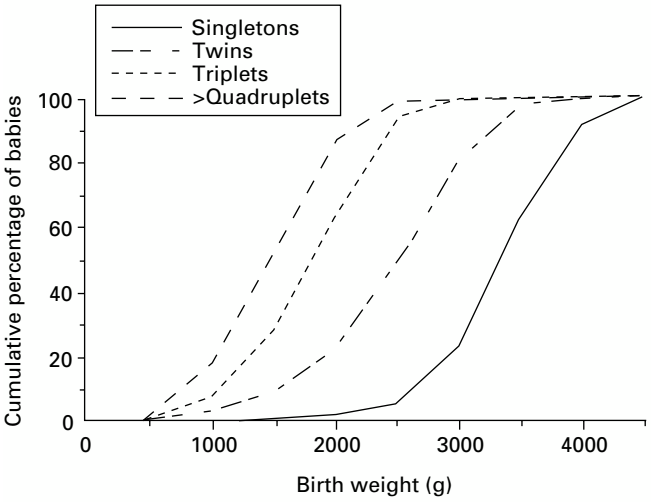


Figure 13.10 Birth weight distribution of singleton and multiple births

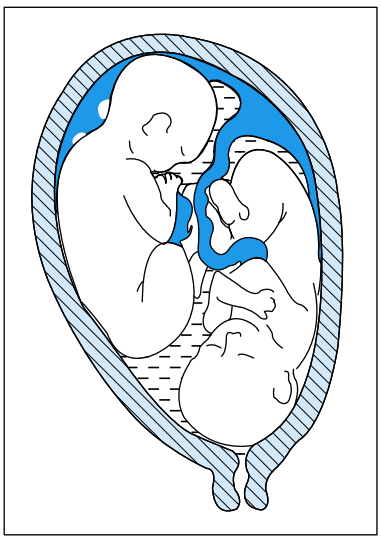


Figure 13.11 The extra stretch that twins place on the myometrium usually ensures that labour starts well before term

Box 13.1 Preparation of parents

- In pregnancy.
  - Why make frequent AN visits
  - When to give up outside work
  - Suitable diet for the mother
  - Potential delivery methods, e.g. CS
  - Visit to neonatal unit
- Discuss the future after twins are born.
  - Extra load for mother with two at once
  - Care of other children and husband
  - Help in home from relatives
  - Breast feeding
  - Local twins club
  - Is housing suitable?

Management

Antenatal care of a woman with a multiple pregnancy needs more vigilance than that of a woman with a singleton

pregnancy. The woman with a multiple pregnancy will need more social support and advice for she is embarking on an extra load before, during, and after delivery. Her socioeconomic state and its implications should be explored. She needs to be seen more often and will require more ultrasound investigations. Care can be shared with the general practitioner in early pregnancy but after about 24 weeks many obstetricians would prefer the care to be where tests can take place—the hospital antenatal clinic.

Antenatal diagnosis of fetal problems in multiple pregnancy must be preceded by careful counselling. All twins should have a detailed ultrasound scan for anomallies at 18–20 weeks and preferably a detailed fetal cardiac scan at 22–24 weeks. At least monthly ultrasound scans in the last trimester should be performed to monitor fetal growth.

Blood pressure and urinary protein concentration are checked at each clinic, as is the symphysio-fundal height. Palpation is performed by an experienced doctor or midwife.

Because of the increased risk of pregnancy-induced hypertension, women carrying twins were traditionally admitted to hospital from 32 weeks to ensure bedrest. The other justification for this was that it postponed preterm labour and so prolonged pregnancy. It is now realised that antenatal time in bed in hospital is not always the best rest: home is more relaxing. Furthermore, it would be more logical to bring the woman into hospital from 24 to 30 weeks, rather than at a later stage of pregnancy. Neither of the desirable aims has been fulfilled in randomised controlled trials of hospital admission after 32 weeks. Though reports from previous decades seemed to show a benefit in one or other of these aims (preventing raised blood pressure or postponing early labour), truly randomised studies in the 1990s have been unable to show benefit. When the disadvantages of separating the woman from her household, as well as the cost to society, are considered, the disadvantages of a routine policy of hospital admission outweigh the advantages. A woman should be advised, however, to come into hospital at a much lower critical level if, in her individual case, specific symptoms arise. These might include the development of hypertension or the threat of early preterm labour. The woman should be made well aware of the warning signs of preterm labour (see Chapter 12) and be encouraged to come in on a low level of suspicion.

Determination of the exact lie and presentation of each twin is often difficult in the last weeks of pregnancy. In many ways detail is not vital but the examiner should ensure that the leading twin is longitudinal. Nearly always the head or a breech is the lower presenting part. In cases of doubt a vaginal examination will usually give a clearer idea for if a presenting part is in or above the pelvis it can be identified more easily by the vaginal examination than through the abdominal wall. Ultrasonography will always confirm lie and presentation.

Delivery should be in a unit with experienced and sufficient staff to look after the resuscitation of both babies.

Many labours are complicated by the presence of one twin as a breech (up to 50%). Monitoring of each twin separately is necessary. An epidural anaesthetic provides good pain relief and less delay if operative delivery is needed quickly. A full account appears in *ABC of Labour*.

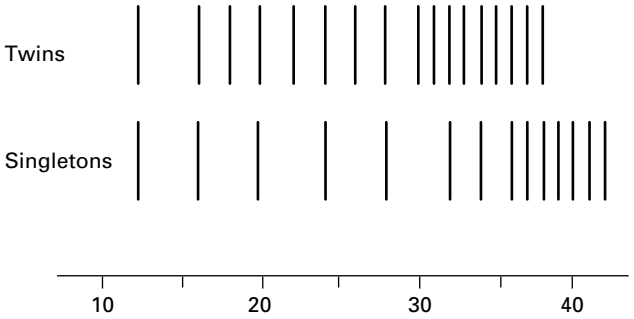


Figure 13.12 Frequency of antenatal visits

Box 13.2 Management of twin pregnancy

- Detailed ultrasound scan for abnormalities at 18–20 weeks
- Antenatal care at hospital clinic after 24 weeks
- More frequent antenatal visits
- Serial ultrasound scans to monitor fetal growth
- Watch for increased risk of maternal complications

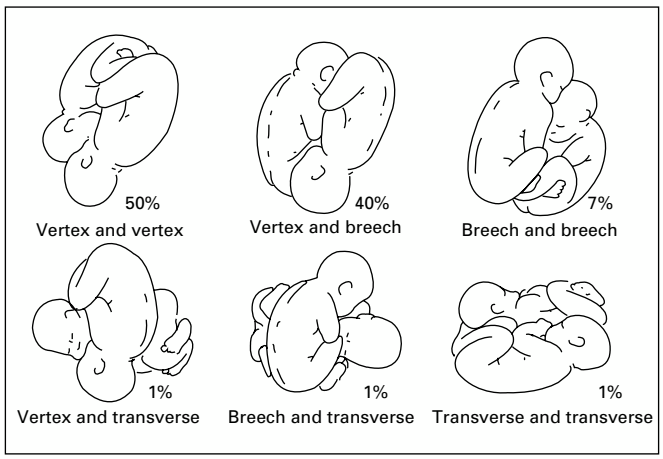


Figure 13.13 Lie and presentation of twins at the start of labour

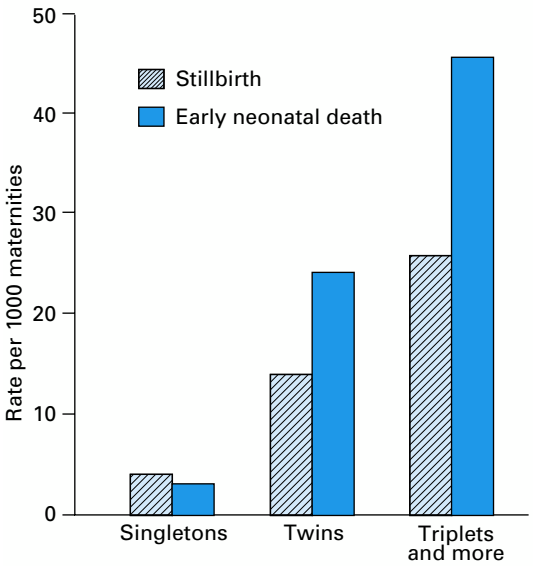


Figure 13.14 Still birth and early neonatal death rates for singleton and multiple pregnancies (England and Wales, 1998)

Outcome

Multiple pregnancies have increased risks for both mother and fetuses. Perinatal mortality rates are about four times higher among twins than singletons, being higher still among monozygotic twins. Rates are even greater in triplets and

quadruplets. About three quarters of the increased mortality is caused by immaturity following preterm delivery, by intrauterine growth restriction or by some combination of the two. The perinatal mortality rate for the second twin at vaginal deliveries is much higher than that of the first, depending on the skills of the professional in charge of the delivery. The perinatal mortality rates associated with antepartum haemorrhage, premature rupture of the membranes, and proteinuric hypertension are increased.

Though some of the increased perinatal mortality rate in twins can be reduced by careful delivery, a large component can be helped by good antenatal care. This includes diagnosing the multiple pregnancy early, carefully managing the woman throughout pregnancy, and either postponing early preterm labour or if it must, ensuring that it takes place in an appropriate hospital with a good neonatal unit.

Triplets and more

With the increasing use of ovarian stimulation and other assorted fertility techniques, more women are becoming pregnant with higher multiples of fetus. All the complications that apply to twins can happen and more so. The risks are multiplied by the increased number; most triplets deliver before 35 weeks and usually by caesarean section. The survival to live birth of more than five fetuses is most unlikely and many doctors would advice fetal reduction down to two or three babies if the woman wishes this.

Recommended reading

- Botting B, MacFarlane A, Price F. *Three, four and more—a study of triplets and higher order births*. London: HMSO, 1990.
- Bryan E, Denton J, Hallet F. *Guidelines for professionals: multiple pregnancy*. London: Multiple Births Foundation, 1997.
- Ward R, Whittle M, eds. *Multiple pregnancy*. London: RCOG, 1995.
- Warner B, Keily J, Donovan E. Multiple births. *Clin Perinatol* 2000;**27**:347–61.

14 The audit of birth

Doctors are mostly literate but are commonly innumerate. We are largely ignorant and frightened of the safe and helpful use of figures because we have never been taught to understand them properly. We often try to dismiss them, believing that they are used during medical debate in a biased fashion to support the arguments of the proponents but are put to one side as non-relevant or non-significant by the opponents. This is a head in the sand attitude as statistics are extremely helpful in providing evidence of changes. Obstetricians should be well used to monitoring their activities statistically, having collected and published data long before the current fashion for audit started.

To be useful medical statistics must be:

- collected properly from a prescribed population;
- analysed in a valid fashion so as not to produce bias;
- presented promptly in a digestible, unbiased form.

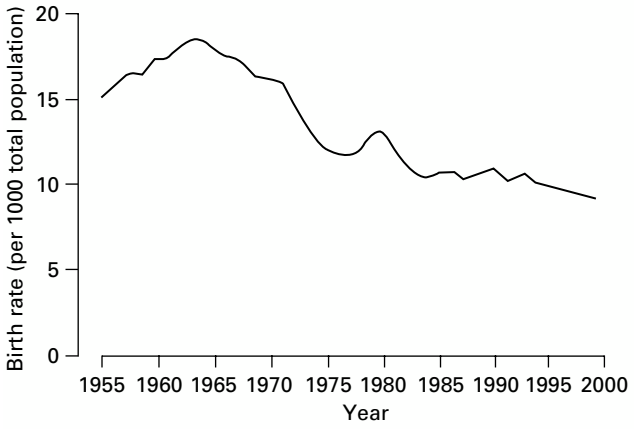


Figure 14.1 Birth rates in England and Wales, 1955–2000

Birth rates

The number of babies born is counted by two processes, birth registration and birth notification. These are two statutory obligations—registration by parents and notification by professional staff.

Birth rates are often expressed as a ratio of the number of births to the number of people in the existing population, gathered from the decennial census.

$$\text{Birth rate} = \frac{\text{No. of births} \times 1000}{\text{No. of people in the population}}$$

The birth rate in the UK in 1998 was 7.8 per 1000.

The denominator in this birth rate formula includes, however, men, who never give birth, and women under 15 and over 44, who are mostly outside the reproductive age group. Hence the denominator does not relate to the numerator very well; an alternative measure is more commonly used in the Western world:

$$\text{General fertility rate} = \frac{\text{No. of babies born} \times 1000}{\text{No. of women in the population aged 15–44}}$$

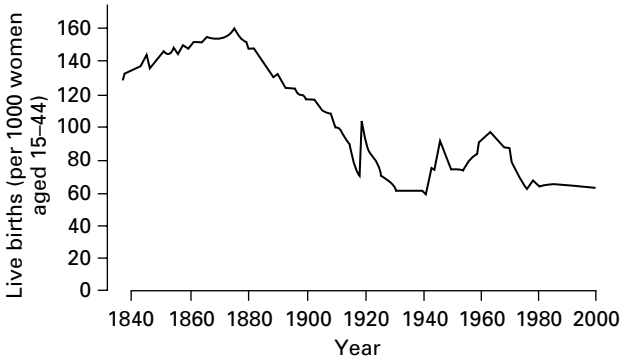


Figure 14.2 General fertility rates in England and Wales, 1840–2000

The general fertility rate in England and Wales in 1998 was 59 per 1000. International comparisons are harder because only countries with good census systems can break down population data to determine the number of women aged 15–44.

For the less numerically minded, completed family size is a user friendly statistic: we can all imagine the size of a family. Unfortunately, these data depend on uncertain estimates and are usually produced some years after the women concerned have passed their reproductive years and completed their family. Obviously, to increase any population the number in a family needs to be more than two. In much of western Europe it is 1.7 to 2.2, whereas in Kenya it is 6.9, showing a rapidly increasing population.

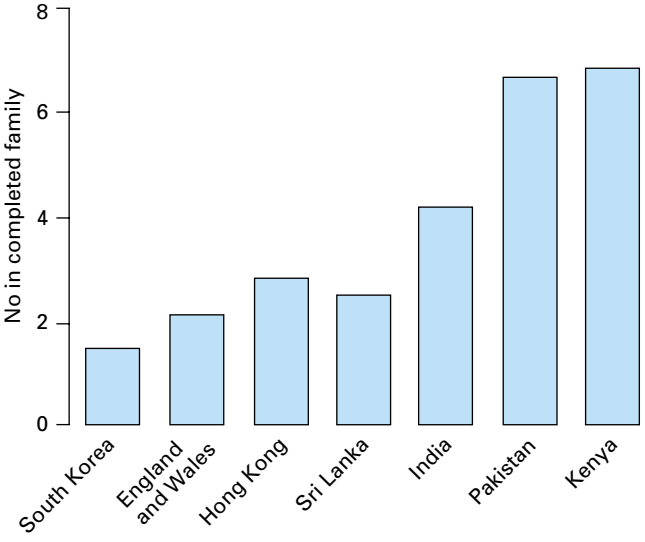


Figure 14.3 Completed family size

Perinatal mortality

Deaths of babies around the time of birth are assessed by three sets of statistics.

(1) Stillbirths or late intrauterine deaths occur when a child is delivered after the 24th completed week of pregnancy but shows no signs of life at birth:

$$\text{Stillbirth rate} = \frac{\text{No. of babies born dead after 24 weeks} \times 1000}{\text{Total births (live and stillborn)}}$$

The stillbirth rate for England and Wales in 1998 was 5.3 per 1000 total births.

(2) Neonatal death is recorded when babies who are born alive (regardless of gestation) die in the first 28 days of life; early neonatal deaths refer to babies who die in the first seven days after birth. All babies who die in the first year of life are recorded as infant deaths but those who die after the first four weeks are defined as postneonatal deaths.

$$\text{Neonatal death rate} = \frac{\text{No. of babies dying between 1–28 days} \times 1000}{\text{No. of live births}}$$

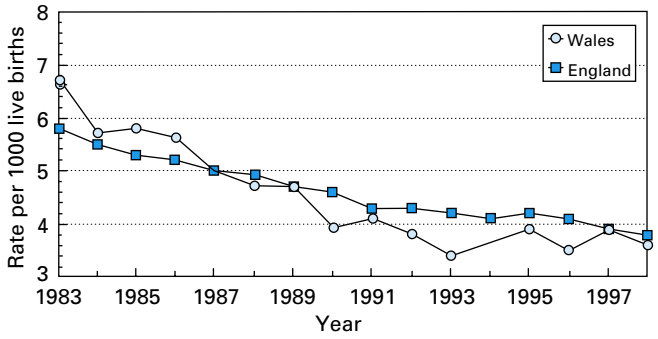


Figure 14.5 Neonatal mortality rates in England and Wales, 1983–97

The neonatal death rate for England and Wales in 1998 was 3.8 per 1000 live births.

(3) In the past 50 years perinatal mortality rates have been used to group together all babies whose deaths may have some relation to obstetric events; thus all stillbirths and neonatal deaths in the first week after birth are considered.

$$\text{Perinatal mortality rate} = \frac{\text{Stillbirths} + \text{neonatal deaths in the first 7 days} \times 1000}{\text{Total births (live and stillborn)}}$$

The perinatal mortality rate in England and Wales in 1998 was 7.9 per 1000 total births.

There is some degree of dissatisfaction with the use of perinatal mortality rates as an index of obstetric performance. Many babies born early now survive in neonatal units. Others with congenital lethal malformations may be kept alive in such units until the second or third week and so are not included in the perinatal mortality rate. We may return to looking at stillbirth rates and neonatal death rates as separate statistics. In 1992 in the UK, the gestation stage for viability was reduced from 28 to 24 weeks and so rates increased slightly around this time—a statistical but not a real blip.

The perinatal mortality rate has fallen steadily since the second world war. When data are compared from different countries, rates are falling in most of them at about the same rate, though some countries start worse off and stay there. This reflects the influence of socioeconomic factors and patterns of reproduction more than the quality of obstetric facilities. A similar pattern can be seen to a smaller extent in the regions of the UK.

The three main causes of perinatal mortality in the UK are low birth weight, hypoxia, and congenital abnormalities. Unfortunately, even after careful reexamination of notes and autopsy, some 70% of stillbirths are unexplained. Low birth weight is currently one of the biggest problems in the Western world (see Chapters 11 and 12). Hypoxia is mostly a problem of

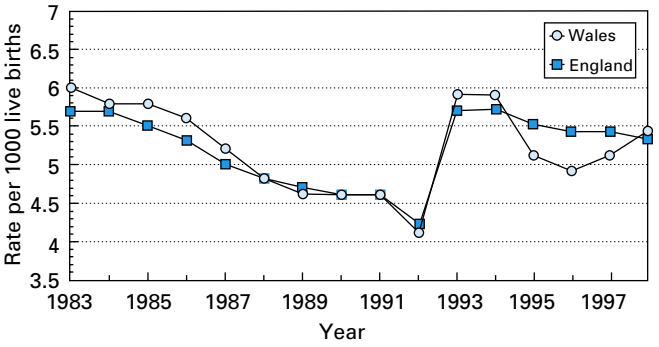


Figure 14.4 Stillbirth rates in England and Wales, 1983–97

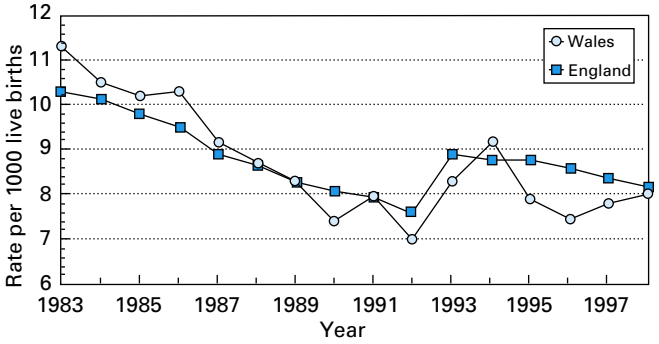


Figure 14.6 Perinatal mortality rates in England and Wales, 1983–97

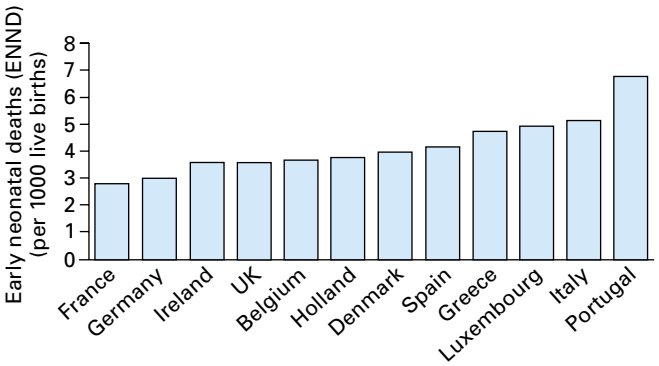


Figure 14.7 Early neonatal mortality rates (≤7 days) in the 12 countries of the then European community. (Source: Europe en Chiffers, Eurostat Office 1995)

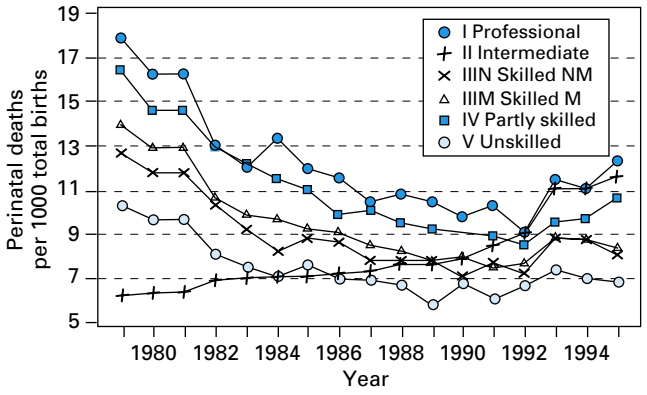


Figure 14.8 Perinatal mortality by father's social class 1979–95. (Source: ONS Mortality Statistics, DH3 series)

labour and to some extent is improved by monitoring women at high risk. Congenital abnormalities may be detected at antenatal examination (see Chapter 5) but the real cure of this problem would be to prevent malformations rather than to detect them and then abort the fetus.

Perinatal mortality rates are not a valid measure of obstetric or midwifery performance. In a developed society they are a mixed measure of a country's educational, social, nutritional, and public health systems as well as of obstetric acute medicine. The rate of deaths in the UK by socioeconomic class of the father has narrowed over the years but still in 1995 PNMR of Social Class I was 6.9 compared with 12.2 in Social Class V, almost double. Probably only a third of the improvement in perinatal mortality statistics is due to improvements in medicine and midwifery. The rest is due to social and economic factors.

A nationwide examination of stillbirths and neonatal deaths together with infant deaths has now been developed, the Confidential Enquiry into Stillbirths and Deaths in Infancy. Reports are made to the regional centres and concentrated in the Health Departments whence they are published each year.

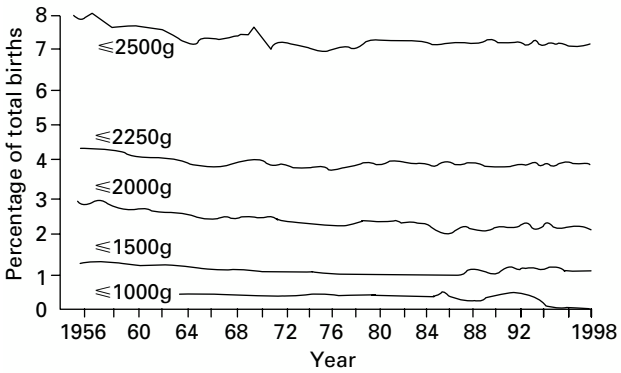


Figure 14.9 The proportions of babies in different birth weight bands have altered little in the past 30 years

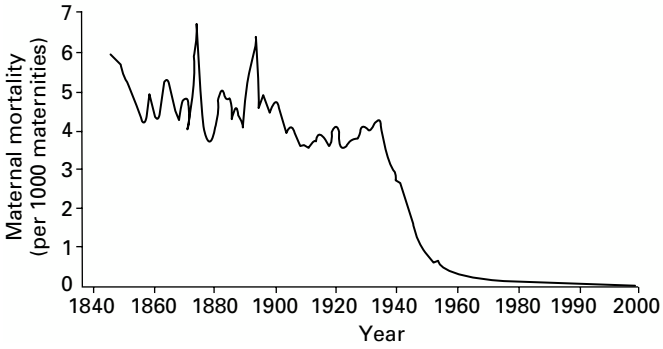


Figure 14.10 Maternal mortality in England and Wales, 1845–2000

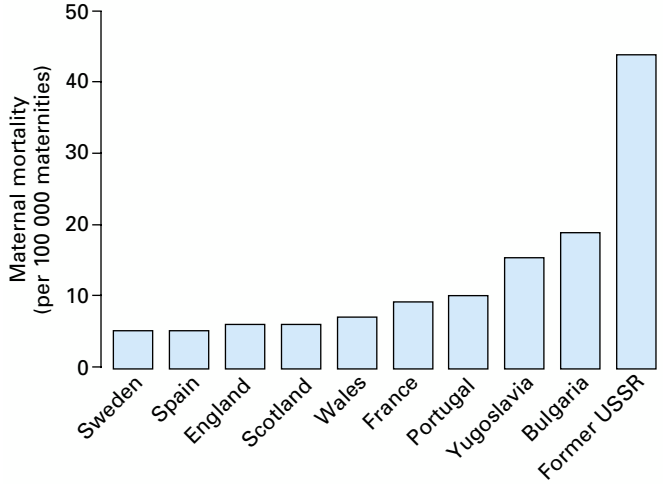


Figure 14.11 Maternal mortality in various European countries (excluding deaths from abortion)

Maternal mortality

Maternal deaths are rare in the Western world but this is not so everywhere: in Kenya a woman has a chance as high as one in 20 of dying during one of her several pregnancies.

Maternal death usually refers to a woman dying in pregnancy, childbirth, or within 42 days of the end of pregnancy. In many countries, including the UK, it includes deaths after an abortion or an ectopic pregnancy but in some countries it does not. The definition in Britain used to include deaths up to one year but has now come in line with World Health Organisation recommendations.

$$\text{Maternal death rate} = \frac{\text{Deaths in pregnancy, childbirth and 6 weeks afterwards} \times 1000}{\text{Total maternities}}$$

Maternal death rates in the UK did not reduce in this century as swiftly as did the rates of perinatal death. Until the mid-1930s maternal mortality was the same as it had been in Victorian times. With the development of antibiotic therapy the rates of puerperal sepsis reduced; to this was added the improvements brought by a proper blood transfusion service catalysed by the Second World War. The founding of the colleges of midwives and obstetricians organised professional training and standards, and the unification of the antenatal and delivery services in the new NHS helped further.

International statistics on maternal mortality are less easy to determine in a comparable way as different countries have different exclusions. In general, however, maternal mortality is an index of medical and midwifery care more than are the perinatal mortality rates. Maternal death rates by region and by country within the UK also vary but differently from perinatal mortality rates.

In Britain the Confidential Enquiry into Maternal Deaths has been set up to provide information about maternal deaths. A complete case history of each maternal death is obtained and published triennially by the Department of Health, keeping all information confidential. The reports are published from the whole UK rather than separately for the four kingdoms.

The maternal mortality in the UK was reported to be 7.4 per 100 000 in 1994–96; principal causes of maternal death in

Table 14.1 A total life risk assessment of maternal deaths, derived from both the maternal mortality ratio and the number of children a woman has (WHO 1996)

Country	Risk
Norway	1 in 7300
Italy	1 in 5300
UK	1 in 5100
Australia	1 in 4900
USA	1 in 3500
Poland	1 in 2200
Cuba	1 in 490
China	1 in 400
Mexico	1 in 220
India	1 in 37
Zimbabwe	1 in 28
Kenya	1 in 20
Mali	1 in 10

England and Wales are hypertension and pulmonary embolism. To reduce the toll of hypertension the inquiry committee recommends that in each region there should be one or two hospitals with staff skilled at looking after pregnant hypertensive mothers and their fetuses. Women with severe degrees of this condition should be electively transferred to these centres. Pulmonary embolism commonly follows popliteal or pelvic vein thrombosis, which should be watched for, particularly in the puerperium after an operative delivery. An active policy of thromboprophylaxis could reduce this cause of death.

Other major killers in the past were infection and haemorrhage; currently these are much reduced. It must give satisfaction to those who fought for the Abortion Act of 1967 to find that in the last five triennia reported by the confidential inquiry committee (1982–96) there was not a single death from illegal abortion in England and Wales.

Near misses

An audit of serious complications such as haemorrhage over 1000 ml or pulmonary embolism in women who do not die gives an index of morbidity. Such near misses are harder to identify and collect but may be used in local audit. Definitions should be agreed and data collection should be prompt.

Conclusion

Too many doctors think of vital statistics in terms of Disraeli's, "Lies, damn lies and statistics".

Perhaps they should look at statistics in the same way as did Richard Asher: "When something can be expressed in a numerical way, it is an aid to precise and accurate thinking".

Most of the data in England and Wales are derived from the old Office of Population Censuses and Surveys, now the Office for National Statistics. The data on maternal mortality come from the Confidential Enquiries into Maternal Deaths for the UK and those of perinatal data from the Confidential Enquiry into Stillbirths and Deaths in Infancy for England and its parallel body in Wales.

Table 14.2 Major causes of maternal death (UK 1994–96)

		Rate per million maternities
Direct	Thromboembolism	21.8
	Hypertension	9.1
	Amniotic fluid embolism	7.7
	Haemorrhage	5.5
	Sepsis	6.4
	Anaesthesia	0.5
Indirect	Cardiac	16.4
	Psychiatric	4.1

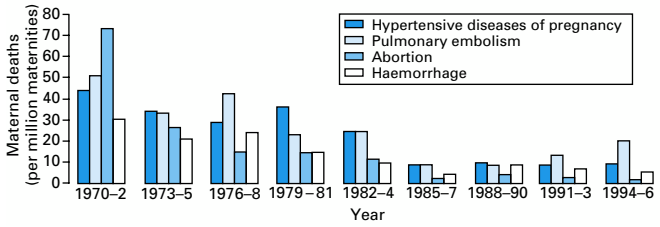


Figure 14.12 Major causes of maternal death in England and Wales, 1970–96

Recommended reading

- *Confidential enquiry into Stillbirths and Deaths in Infancy 1999.* London: ONS, 2000.
- Macfarlane A, Mugford M. *Birth counts.* London: Stationery Office, 2000.
- Office for National Statistics. *Why mothers die? Report on Confidential Enquiries into Maternal Deaths in the United Kingdom 1994–1996.* London: ONS, 1999.

L’envoi

Patterns of antenatal care shifted more in the last years of the 20th century than ever before. Less frequent visits for women with normal pregnancies and a wider sharing of professional responsibility are overtaking the hospital dominated and regimented patterns of the middle of the last century.

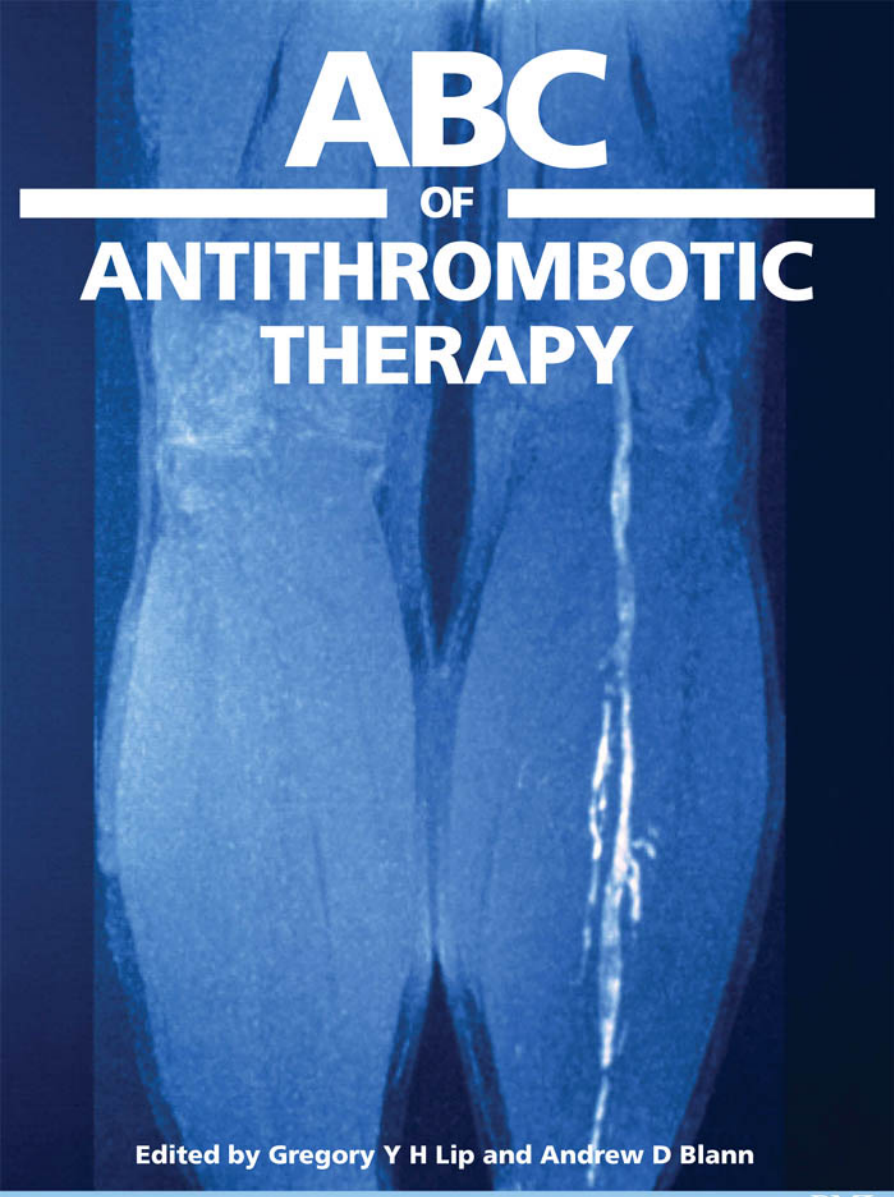
The development of antenatal care reflects what has happened in all of medicine—first came the clinical observations, then the mounting of investigations, each supported by some scientific pedigree, and only later a guilty sideways look at what value these all provided. In an ideal world all the investigations would have been subjected to rigorous scrutiny—randomised controlled trials and careful checks of sensitivity and specificity—but such intellectual disciplines were introduced after many of the antenatal tests had been started. We did not await the more scientifically assessed investigation because babies were still being born and women still needed to know. Meanwhile, we do the best with what we have. Clinical management should reflect the results of research studies and must depend in future more upon evidence based research promptly delivered.

If we were to see women in appropriate circumstances and make proper use of the proven valuable tests we already have, much effort and money would be saved. We could spend more time listening to and talking with the women we care for. The golden age of antenatal care would then have arrived.

Index

Page numbers in **bold** type refer to figures; those in italic refer to tables or boxed material

ABC OF ANTITHROMBOTIC THERAPY



Edited by Gregory Y H Lip and Andrew D Blann

BMJ Books

ABC OF
ANTITHROMBOTIC THERAPY

To Peck Lin, Philomena, and Aloysius
To Janet, Edward, Eleanor, and Rosalind

ABC OF ANTITHROMBOTIC THERAPY

Edited by

GREGORY Y H LIP

Professor of cardiovascular medicine and director, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

and

ANDREW D BLANN

Senior lecturer in medicine, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

BMJ Books

ЗНАНИЕ

First published in 2003

by BMJ Publishing Group Ltd, BMA House, Tavistock Square, London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0 7279 17714

Typeset by BMJ Electronic Production and Newgen Imaging Systems
Printed and bound in Spain by GraphyCems, Navarra

Cover image depicts a deep vein thrombosis scan of a leg vein blocked by a thrombus (blood clot, white) in a patient with deep vein thrombosis. With permission from James King-Holmes/Science Photo Library

Contents

Contributors

Andrew D Blann
Senior lecturer in medicine, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

Bernard S P Chin
Research fellow, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

Derek L Connolly
Consultant cardiologist, department of cardiology and vascular medicine, Sandwell and West Birmingham Hospitals NHS Trust, Sandwell Hospital, West Bromwich

Dwayne S G Conway
Research fellow, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

David A Fitzmaurice
Reader in primary care and general practice, Medical School, University of Birmingham, Edgbaston, Birmingham

Ira Goldsmith
Research fellow in cardiothoracic surgery, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

Robert G Hart
Professor of neurology, department of medicine (neurology), University of Texas Health Sciences Center, San Antonio, USA

Bernd Jilma
Associate professor in the department of clinical pharmacology, Vienna University Hospital, Vienna, Austria

Sridhar Kamath
Research fellow, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

Martin J Landray
Lecturer in medicine, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

Gregory Y H Lip
Professor of cardiovascular medicine and director, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

Andrew J Makin
Research fellow, haemostasis, thrombosis and vascular biology unit, university department of medicine, City Hospital, Birmingham

Neeraj Prasad
Consultant cardiologist, City Hospital, Birmingham

Stanley H Silverman
Consultant vascular surgeon, City Hospital, Birmingham

Alexander G G Turpie
Professor of medicine, McMaster University, Hamilton, Ontario, Canada

Robert D S Watson
Consultant cardiologist, City Hospital, Birmingham

Preface

The seeds for this book were sown with the establishment of the haemostasis, thrombosis and vascular biology unit at the university department of medicine, City Hospital, Birmingham—with the coming together of clinicians and scientists interested in thrombosis and vascular biology, bridging the previous divide in thrombosis between basic science research and the application to clinical practice. Indeed, thrombosis is the underlying pathophysiological process in a wide variety of conditions. A greater understanding of the mechanisms leading to thrombosis, and newer developments in the field of antithrombotic therapy make the field all the more dynamic and exciting.

The multidisciplinary team effort and the wide range of research areas studied in our unit forms the core content of the *ABC of Antithrombotic Therapy*. In major textbooks on thrombosis the scope is comprehensive, background details on physiology and pathophysiology are abundant, and treatment options are listed to exhaustion—the patient may sometimes almost disappear in the wealth of information. Our approach in this book—typical of the ABC series in the *British Medical Journal*—tries to synthesise and integrate the extensive research and clinical data that are needed to manage a particular situation as masterly as it is possible. We hope we have produced a patient-oriented guide with relevant information from clinical epidemiology, pathophysiology, common sense clinical judgement, and evidence based treatment options, with reference to recently published antithrombotic therapy guidelines from the American College of Chest Physicians, British Society for Haematology, European Society of Cardiology, American College of Cardiology, and American Heart Association.

Our expectant readers are physicians, general practitioners, medical or nursing students, nurses, and healthcare scientists who care for patients presenting with thrombosis-related problems, and thus, the scope is necessarily wide, ranging from venous thromboembolism to atrial fibrillation and stroke, and to thrombosis in cancer and thrombophilic states. Chapters on clinical pharmacology and bleeding risk, as well as anticoagulation monitoring are included. Furthermore, this book includes additional chapters which were not included in the 14 issues of this series when it first appeared in the *British Medical Journal*.

We thank our excellent colleagues for their help, encouragement and contributions, as well as Sally Carter at BMJ Books for encouraging us to complete the series and book, nearly to schedule.

Gregory Y H Lip
Andrew D Blann
Birmingham, April 2003

1 An overview of antithrombotic therapy

Andrew D Blann, Martin J Landray, Gregory Y H Lip

Many of the common problems in clinical practice today relate to thrombosis. The underlying final pathophysiological process in myocardial infarction and stroke is thrombus formation (thrombogenesis). Common cardiovascular disorders such as atrial fibrillation and heart failure are also associated with thrombogenesis. Thrombosis is also a clinical problem in various cancers and after surgery, especially orthopaedic.

Pathophysiology

Over 150 years ago Virchow recognised three prerequisites for thrombogenesis: abnormal blood flow, vessel wall abnormalities, and blood constituent abnormalities. This concept has been extended by modern knowledge of the endothelial function, flow characteristics, and blood constituents including haemorheological factors, clotting factors, and platelet physiology. As thrombus consists of platelets and fibrin (and often bystanding erythrocytes and white blood cells), optimum antithrombotic prophylactic therapy can and should be directed towards both.

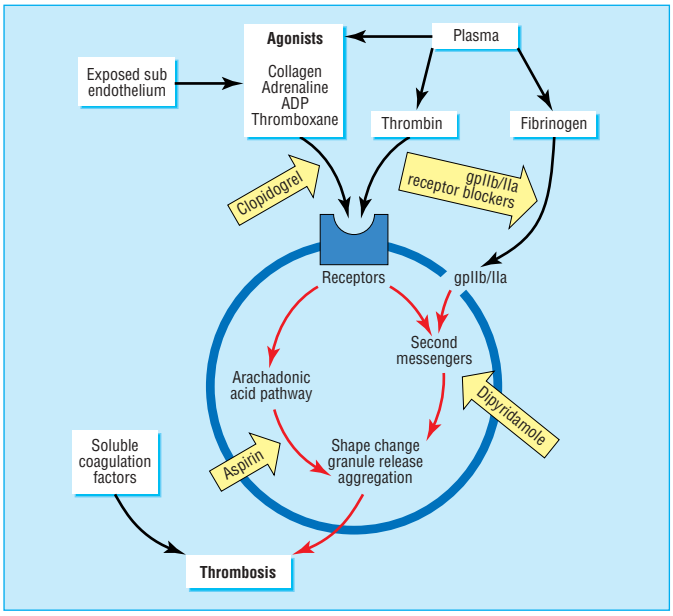
Antiplatelet drugs

Aspirin and agents acting on the cyclo-oxygenase pathway
Aspirin irreversibly inhibits cyclo-oxygenase by acetylation of amino acids that are next to the active site. In platelets, this is the rate limiting step in synthesis of thromboxane A_2, and inhibition occurs in the megakaryocyte so that all budding platelets are dysfunctional. Because platelets are unable to regenerate fresh cyclo-oxygenase in response, the effect of aspirin remains as long as the lifespan of the platelet (generally about 10 days). A severe weakness of aspirin is that its specificity for cyclo-oxygenase means it has little effect on other pathways of platelet activation. Thus aspirin fails to prevent aggregation induced by thrombin and only partially inhibits that induced by ADP and high dose collagen. Antithrombotic doses used in clinical trials have varied widely from less than 50 mg to over 1200 mg/day, with no evidence of any difference in clinical efficacy. Absorption is over 80% with extensive presystemic metabolism to salicylic acid. Only the parent acetylsalicylic acid has any significant effect on platelet function.

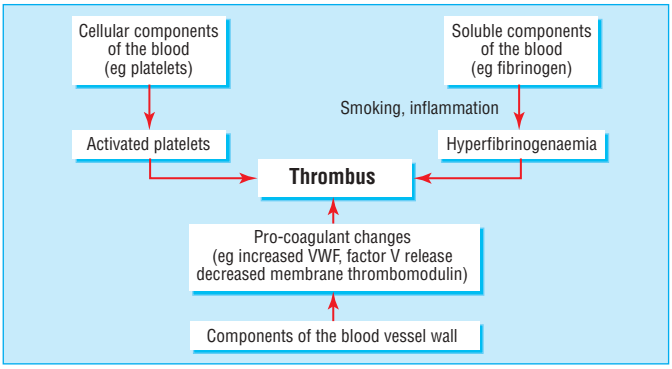
Adverse effects of aspirin include haemorrhage, hypersensitivity and skin rashes, alopecia, and purpura.

Sulfinpyrazone also inhibits cyclo-oxygenase (thus producing an aspirin-like state), but is reversible, and also inhibits serotonin uptake by platelets. Iloprost is a prostacyclin analogue that exerts its effects by promoting vasodilatation and inhibiting platelet aggregation induced by ADP, thereby opposing the effects of thromboxane A_2.

Dipyridamole
Dipyridamole inhibits phosphodiesterase, thus preventing the inactivation of cyclic AMP, intraplatelet levels of which are increased, resulting in reduced activation of cytoplasmic second messengers. However, it may also exert its effect in other ways, such as stimulating prostacyclin release and inhibiting thromboxane A_2 formation. The influence of this drug on these pathways causes reduced platelet aggregability and adhesion in



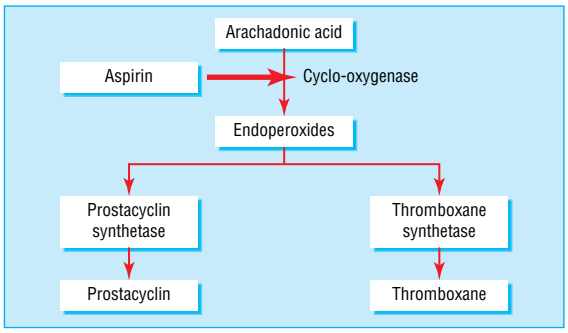
Routes to inhibiting platelet function



Key components of Virchow's triad (VWF=von Willebrand factor)

Contraindications to aspirin

Absolute	Relative
• Active gastrointestinal ulceration	• History of ulceration or dyspepsia
• Hypersensitivity	• Children under 12 years old
• Thrombocytopenia	• Bleeding disorders
	• Warfarin treatment



Platelet metabolism influenced by aspirin

vitro with increased platelet survival in vivo. Its effect is relatively short lasting, and repeated dosing or slow release preparations are needed to achieve 24 hour inhibition of platelet function.

Clopidogrel and ticlopidine

These thienopyridine derivatives inhibit platelet aggregation induced by agonists such as platelet activating factor and collagen, and also dramatically reduce the binding of ADP to a platelet surface purinoreceptor. The mechanism of this inhibitory action seems to be independent of cyclo-oxygenase. There is also impairment of the platelet response to thrombin, collagen, fibrinogen, and von Willebrand factor. The peak action on platelet function occurs after several days of oral dosing. Adverse effects include evidence of bone marrow suppression, in particular leucopenia, especially with ticlopidine.

Other receptor blockers

Signal transduction generally occurs when specific receptors on the surface are occupied by ligands such as ADP, leading to structural modification of the glycoprotein IIb/IIIa receptor on the surface of the platelet. This is the commonest receptor on the platelet surface and represents the final common pathway for platelet aggregation, resulting in crosslinking of platelets.

After intravenous administration of glycoprotein IIb/IIIa receptor inhibitors such as abciximab, platelet aggregation is 90% inhibited within two hours, but function recovers over the course of two days. The major adverse effect is haemorrhage, and concurrent use of oral anticoagulants is contraindicated. Eptifibatide is a cyclic heptapeptide that mimics the part of the structure of fibrinogen that interacts with glycoprotein IIb/IIIa. Thus it is a fraction of the size of abciximab and is targeted at the same structure on the platelet surface.

Clinical trials with oral glycoprotein IIb/IIIa receptor inhibitors have been disappointing, with no beneficial effects seen and even some evidence of harm.

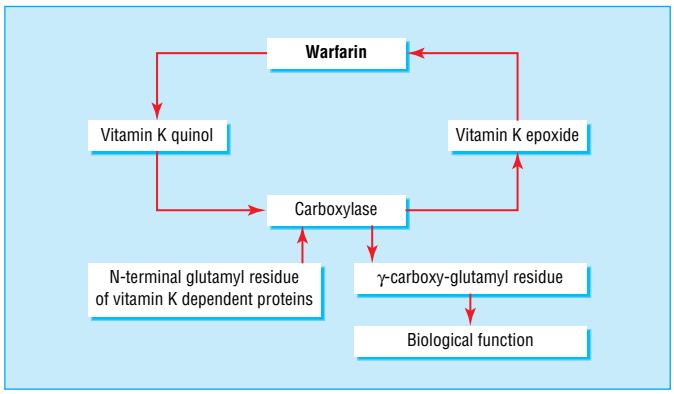
Anticoagulant drugs

Warfarin

This 4-hydroxycoumarin compound, the most widely used anticoagulant in Britain and the Western world, inhibits the synthesis of factors dependent on vitamin K (prothrombin; factors VII, IX, and X; protein C; protein S). Factor VII levels fall rapidly (in <24 hours) but factor II has a longer half life and only falls to 50% of normal after three days. Warfarin is approximately 97% bound to albumin, and free warfarin enters liver parenchymal cells and is degraded in microsomes to an inactive water soluble metabolite that is conjugated and excreted in the bile. Partial reabsorption is followed by renal excretion of conjugated metabolites.

There is a considerable variability in warfarin's effect on patients, its effectiveness being influenced by age, racial background, diet, and co-medications such as antibiotics. Thus it demands frequent laboratory monitoring, the prothrombin time being compared with a standard to produce the international normalised ratio. The degree of anticoagulation required varies with clinical circumstance, but the target international normalised ratio usually ranges from 2 to 4. Phenindione is an alternative oral vitamin K antagonist, but concerns regarding the potential for hepatotoxicity, nephrotoxicity, and blood dyscrasias have reduced its role largely to individuals with documented hypersensitivity to warfarin.

Adverse effects of warfarin include haemorrhage, hypersensitivity and skin rashes, alopecia, and purpura.



Vitamin K metabolism and the effect of warfarin

Factors that influence the efficacy of warfarin*

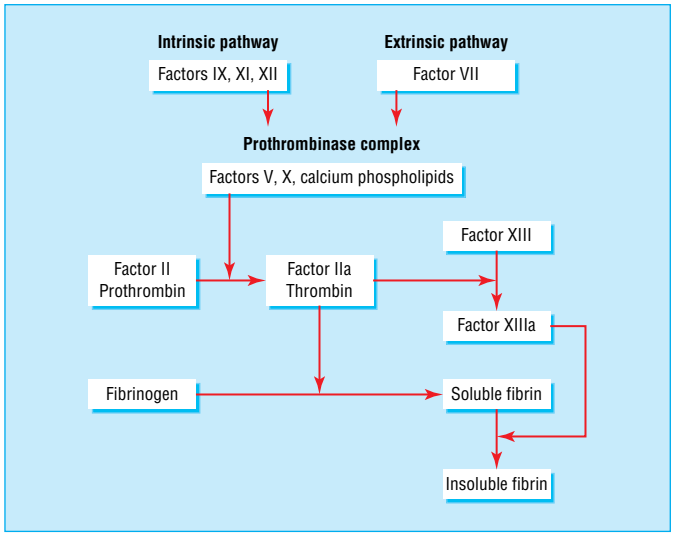
Patient factors

- *Enhanced anticoagulant effect*—Weight loss, increased age (>80 years), acute illness, impaired liver function, heart failure, renal failure, excess alcohol ingestion
- *Reduced anticoagulant effect*—Weight gain, diarrhoea and vomiting, relative youth (<40 years), Asian or African-Caribbean background

Examples of drug interactions with warfarin

- *Reduced protein binding*—Aspirin, phenylbutazone, sulfinpyrazone, chlorpromazine
- *Inhibition of metabolism of warfarin*—Cimetidine, erythromycin, sodium valproate
- *Enhanced metabolism of warfarin*—Barbiturates, phenytoin, carbamazepine
- *Reduced synthesis of factors II, VII, IX, X*—Phenytoin, salicylates
- *Reduced absorption of vitamin K*—Broad spectrum antibiotics, laxatives
- *Enhanced risk of peptic ulceration*—Aspirin, NSAIDs, corticosteroids
- *Thrombolytics*—Streptokinase, tissue plasminogen activator
- *Antiplatelet drugs*—Aspirin, NSAIDs

*This list is intended to be illustrative not exhaustive



Simplified coagulation cascade

Melagatran
This oral thrombin inhibitor undergoing phase III trials seems to be well tolerated, with few clinically significant bleeding problems, in patients with venous thromboembolism. Although considerable pharmacokinetic and animal data exist, solid evidence of its effectiveness compared with warfarin and heparin in patients at high or low risk is still awaited.

Heparin
Heparin is a glycosaminoglycan whose major anticoagulant effect is accounted for by a pentasaccharide with a high affinity for antithrombin III. This binding results in a conformational change in antithrombin III so that inactivation of coagulation enzymes thrombin (IIa), factor IXa, and factor Xa is markedly enhanced. Its short half life means it must be given continuously, and its extensive first pass metabolism means it must be given parenterally, preferably by continuous intravenous infusion, and it is therefore inappropriate for home use. The effect on the intrinsic clotting cascade must be monitored carefully by measuring the activated partial thromboplastin time (APTT), generally aiming for a value 1.5 to 2.5 times that of control.

Unfractionated heparin consists of a heterogeneous mixture of polysaccharides with an average molecular weight of 15 000 Da. Low molecular weight heparins (4000-6000 Da) are weaker inhibitors of thrombin but inhibit factor Xa to a similar extent. Different commercial preparations of low molecular weight heparin vary in the ratio of anti-Xa to antithrombin activity, although the clinical relevance of this is uncertain. Better absorption after subcutaneous administration and reduced protein binding result in greatly improved bioavailability. The effective half life after subcutaneous injection is four hours, allowing an injection once daily in most circumstances. These more predictable pharmacokinetics allow the dose to be calculated on the basis of the patient's weight and reduce the requirement for frequent monitoring. In those rare cases where monitoring is deemed necessary, measurement of plasma levels of anti-Xa activity is needed. Tests of APTT are unhelpful.

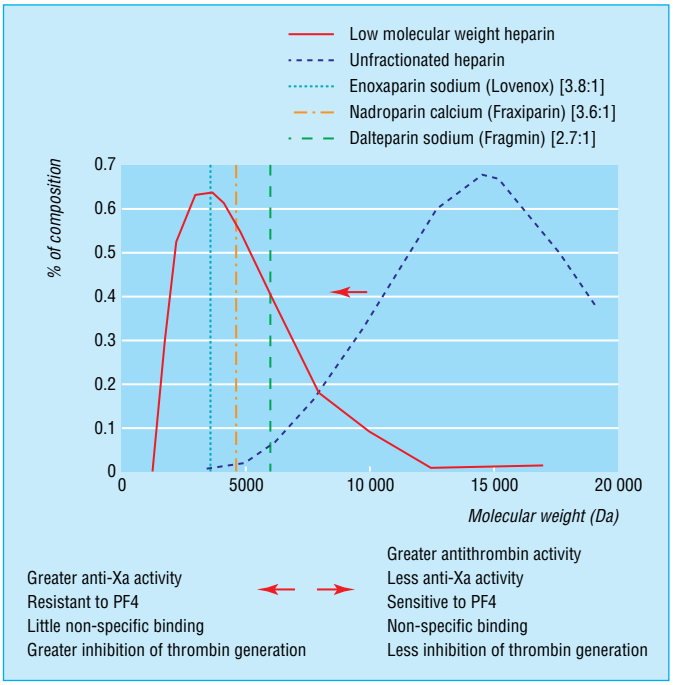
Major adverse effects of heparin include haemorrhage, osteoporosis, alopecia, thrombocytopenia, and hypersensitivity. At present, the risk of haemorrhage seems to be similar with low molecular weight and unfractionated heparin. However, the risk of heparin induced thrombocytopenia seems to be less with the low molecular weight form.

Hirudin and direct thrombin inhibitors
Hirudin, a 65 amino acid residue anticoagulant peptide with a relative molecular mass of 7000 Da purified from the leech *Hirudo medicinalis*, binds thrombin with high specificity and sensitivity. With a true half life of about an hour and a half life effect on the APTT of two to three hours, it may be seen as an alternative to heparin in indications such as unstable angina and in coronary angioplasty.

Many derivatives are available, with hirulog and argatroban among the best developed. However, trials of the former have been discouraging: no clear benefit over heparin was shown. Conversely, argatroban may have a role in the anticoagulation of patients unable to tolerate heparin as a result of heparin induced thrombocytopenia. Furthermore, in a clinical trial of patients with heparin induced thrombocytopenia, use of argatroban was associated with a reduction in levels of plasma platelet activation markers.

Thrombolytic agents

These agents lyse pre-existing thrombus, either by potentiating the body's own fibrinolytic pathways (such as streptokinase) or



The three low molecular weight heparins that have been evaluated in clinical trials of acute coronary syndromes are shown with their respective anti-Xa and antithrombin activity (PF4=platelet factor 4)

Comparison of low molecular weight and unfractionated heparins

	Unfractionated heparin	Low molecular weight heparin
Action	Anti-XIIa, XIa, IXa, VIIa, antithrombin	Mostly anti-Xa
Route of administration	Subcutaneous Intravenous	Subcutaneous
Absorption from subcutaneous route	Slow	Improved
Protein binding	Proteins in plasma and on endothelium	Reduced
Bioavailability	Subcutaneous—10-30% at low doses, 90% at higher doses Intravenous—100% by definition	>90%
Effective half life	Subcutaneous—1.5 hours Intravenous—30 min	4 hours
Between and within individual variation	Extensive	Minimal
Monitoring	APTT	Not required (anti-Xa activity)
Elimination	Liver and kidney	Kidney

by mimicking natural thrombolytic molecules (such as tissue plasminogen activator). The common agents in clinical use are derived from bacterial products (streptokinase) or manufactured using recombinant DNA technology (recombinant tissue plasminogen activator). Newer drugs aim to be less antigenic and more thrombus specific in an attempt to increase efficacy and specificity of various agents; on present evidence, however, the differences between thrombolytic agents are only marginal. Because of the lack of site specificity for these drugs, the major adverse effect is that of haemorrhage (gastrointestinal, intracranial, etc). The other important adverse effect is that of hypersensitivity reaction, especially with streptokinase. This usually manifests as flushing, breathlessness, rash, urticaria, and hypotension. Severe anaphylaxis is rare. Hypersensitivity reactions are avoided by using tissue plasminogen activator or recombinant tissue plasminogen activator, which are not antigenic.

Streptokinase

Derived from streptococci, this product is an effective thrombolytic agent for the treatment of acute myocardial infarction and pulmonary thromboembolism. Acting by converting plasminogen to plasmin, the main fibrinolytic enzyme, it potentiates fibrinolysis. However, it is not site specific, lysing thrombus anywhere in the body. Being bacteria derived, it is antigenic, and repeated administration results in neutralising antibodies and allergic reactions. For example, a single administration of 1.5 MU for acute myocardial infarction results in neutralising antibodies that have been shown to persist for up to four years and are sufficient to neutralise a repeat administration of a similar dose of the agent in half of cases.

Tissue plasminogen activator

In clinical use this is produced by recombinant DNA technology and mimics an endogenous molecule that activates the fibrinolytic system. Thus, recombinant tissue plasminogen activator does not elicit an allergic response and is considered more clot specific. Nevertheless, it has a short half life and needs continuous infusion to achieve its greatest efficacy. Accelerated administration of tissue plasminogen activator gives a slight mortality advantage over streptokinase at the cost of a marginal increase in stroke rate.

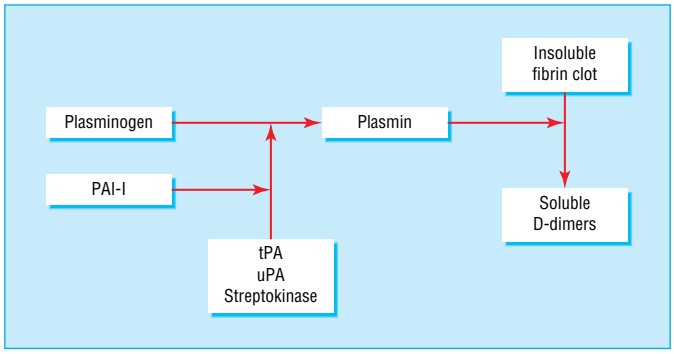
Fibrinolytic drugs

Examples	Source	Mechanism of action
Streptokinase	Group C β haemolytic streptococci	Complexes with and activates plasminogen
Urokinase	Trypsin-like chemical produced by kidney	Direct acting plasminogen activator
Reteplase (recombinant tissue plasminogen activator)	Recombinant DNA technology	Acivates plasminogen, non-immunogenic

Contraindications to thrombolysis

Absolute
- Recent or current haemorrhage, trauma, or surgery
- Active peptic ulceration
- Coagulation defects
- Oesophageal varices
- Coma
- Recent or disabling cerebrovascular accident
- Hypertension
- Aortic dissection

Relative
- Previous peptic ulceration
- Warfarin
- Liver disease
- Previous use of anistreplase or streptokinase within four years (use alternative agent)
- Hypersensitivity (anistreplase, streptokinase)
- Heavy vaginal bleeding



Simplified fibrinolysis (PAI-1=plasminogen activator inhibitor, tPA=tissue plasminogen activator, uPA=urokinase plasminogen activator)

Further reading

- Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *BMJ* 1994;308:81-106
- Blann AD, Lip GYH. Virchow's triad revisited: the importance of soluble coagulation factors, the endothelium, and platelets. *Thromb Res* 2001;101:321-7
- CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *Lancet* 1996;348:1329-39
- Catella-Lawson F. Direct thrombin inhibitors in cardiovascular disease. *Coron Artery Dis* 1997;8:105-11
- Eriksson H, Eriksson UG, Frison L. Pharmacokinetic and pharmacodynamics of melagatran, a novel synthetic LMW thrombin inhibitor, in patients with a DVT. *Thromb Haemost* 1999;81: 358-63
- International Stroke Trial Collaborative Group. The international stroke trial (IST): a randomised trial of aspirin, subcutaneous heparin, or both, or neither among 19 435 patients with acute ischaemic stroke. *Lancet* 1997;349:1569-81
- Lewis BE, Wallis DE, Berkowitz SD, Matthai WH, Fareed J, Walenga JM, et al. Argatroban anticoagulant therapy in patients with heparin-induced thrombocytopenia. *Circulation* 2001;103:1838-43
- Nurden AT. New thoughts on strategies for modulating platelet function through the inhibition of surface receptors. *Haemostasis* 1996;20:78-88
- Stirling Y. Warfarin-induced changes in procoagulant and anticoagulant proteins. *Blood Coagul Fibrinolysis* 1995;6:361-73

The figure showing percentage of composition of unfractionated and low molecular weight heparin in terms of molecular weight is adapted from Levine GN, Ali MN, Schafer AI. *Arch Intern Med* 2001;161: 937-48.

2 Bleeding risks of antithrombotic therapy

David A Fitzmaurice, Andrew D Blann, Gregory Y H Lip

Many of the common cardiovascular disorders (especially in elderly people) are linked to thrombosis—such as ischaemic heart disease, atrial fibrillation, valve disease, hypertension, and atherosclerotic vascular disease—requiring the use of antithrombotic therapy. This raises questions regarding the appropriate use of antithrombotic therapy in older people, especially because strategies such as anticoagulation with warfarin need regular monitoring of the international normalised ratio (INR), a measure of the induced haemorrhagic tendency, and carry a risk of bleeding. The presence of concomitant physical and medical problems increases the interactions and risks associated with warfarin, and anticoagulation in elderly patients often needs an assessment of the overall risk:benefit ratio.

Physical frailty in elderly people may reduce access to anticoagulant clinics for INR monitoring. The decline in cognitive function in some elderly patients also may reduce compliance with anticoagulation and the appreciation of bleeding risks and drug interactions. However, in recent studies of anticoagulation in elderly people, no significant associations of anticoagulant control were found with age, sex, social circumstances, mobility, domicillary supervision of medication, or indications for anticoagulation.

Questions to ask when considering oral anticoagulation

- Is there a definite indication (such as atrial fibrillation)?
- Is there a high risk of bleeding or strong contraindication against anticoagulation?
- Will concurrent medication or disease states increase bleeding risk or interfere with anticoagulation control?
- Is drug compliance and attendance at anticoagulant clinic for monitoring likely to be a problem?
- Will there be regular review of the patient, especially with regard to risks and benefits of anticoagulation?

$$\text{INR}=\left(\frac{\text{patient's prothrombin time}}{\text{mean normal time}}\right)^{\text{ISI}}$$

ISI=international sensitivity ratio. The mean normal prothrombin time is often generated from samples from local healthy subjects or a commercially available standard. The exact value of the ISI depends on the thromboplastin used in the prothrombin time method

Warfarin

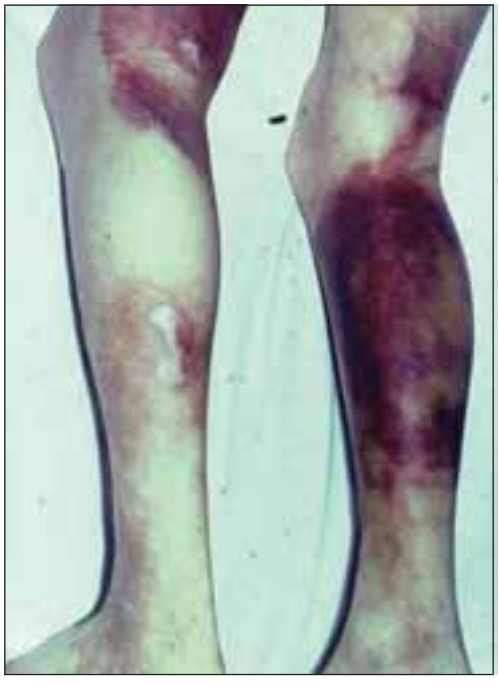
Bleeding is the most serious and common complication of warfarin treatment. For any given patient, the potential benefit from prevention of thromboembolic disease needs to be balanced against the potential harm from induced haemorrhagic side effects.

Minor bleeds

Most bleeding problems are clinically minor, although patients are unlikely to view such bleeds in these terms. The problems include nose bleeds, bruising, and excessive bleeding after minor injury such as shaving. Patients should be made aware of these common problems and be reassured that these events are expected in patients receiving warfarin treatment. Menorrhagia is surprisingly rare as a major clinical problem, even though it can be severe.

More serious problems

Patients need access to medical care if they have serious problems. Such problems are generally due to a high INR. Usually, spontaneous bruising, any bleeding that is difficult to arrest, frank haematuria, any evidence of gastrointestinal bleeding, and haemoptysis, need urgent assessment. The definition of minor or major bleeding lacks clarity: in many cases the patient presents with a concern that may need follow up, and a minor bleed can only be defined as such in retrospect. In most cases, evidence of bleeding suggests some underlying pathology but may also be due to drug interactions. For example, a patient with recurrent haemoptysis may be found to have hereditary telangectasia. Further investigation of the cause of bleeding should always be considered, particularly if the bleeding is recurrent. It is also important in these instances to check for concomitant drug use, particularly drugs received over the counter. Patients should be aware that aspirin and



Purpura, petechiae, and haematoma secondary to over-anticoagulation

Sudden, unexplained changes to the efficacy of warfarin may be caused by the consumption of over the counter multivitamin tablets or foodstuffs that contain high levels of vitamin K

non-steroidal anti-inflammatory drugs are particularly dangerous in combination with warfarin; however, even supposedly safe drugs such as paracetamol can affect a patient's bleeding tendency.

Incidence of bleeding problems

The incidence of severe bleeding problems that may bring patients to an accident and emergency unit has probably been overestimated. The annual incidence of fatality caused by warfarin administration has been estimated to be 1%. However, this is based on old data, and, although difficult to prove, the overall improvement in anticoagulation control in the past 10-15 years means that a more realistic figure is about 0.2%. Methodological problems have hampered the interpretation of previously reported data, particularly with regard to definitions of major and minor bleeding episodes, with some investigators accepting hospital admission for transfusion of up to 4 units of blood as being "minor." Certainly, the most serious "major" bleed is an intracranial haemorrhage. Reviews of observational and experimental studies showed annual bleeding rates of 0-4.8% for fatal bleeding and 2.4-8.1% for major bleeds. Minor bleeds are reported more often, with about 15% of patients having at least one minor event a year.

Risk factors for bleeding

Age is the main factor that increases risk of bleeding. One study showed a 32% increase in all bleeding and a 46% increase in major bleeding for every 10 year increase above the age of 40.

Early studies suggested an increased risk with increasing target INR, but the data were difficult to interpret because results were reported in both INR and prothrombin time. The actual risk of bleeding should be taken into account as well as the degree of anticoagulation (as measured by the INR). One study which achieved point prevalence of therapeutic INRs of 77% reported no association between bleeding episodes and target INR.

Data from an Italian study in 2745 patients with 2011 patient years of follow up reported much lower bleeding rates, with an overall rate of 7.6 per 100 patient years. The reported rates for fatal, major, and minor bleeds were 0.25, 1.1, and 6.2 per 100 patient years respectively. This study confirmed an increased risk with age and found a significantly increased risk during the first 90 days of treatment. Peripheral vascular and cerebrovascular disease carried a higher relative risk of bleeding, and target INR was strongly associated with bleeding with a relative risk of 7.9 (95% confidence interval 5.4 to11.5, $P < 0.0001$) when the most recent INR recorded was > 4.5. Data from a trial in a UK community showed 39.8 minor, 0.4 major, and no fatal haemorrhagic events per 100 patient years for the total study population, with 3.9 serious thromboembolic events per 100 patient years, of which 0.79 were fatal.

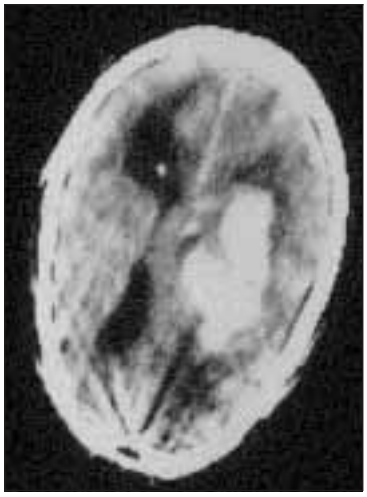
Warfarin is therefore a relatively safe drug, particularly if therapeutic monitoring is performed well. Analogies are often made between therapeutic monitoring of warfarin and monitoring of blood glucose for diabetic patients. Given the increase in numbers of patients receiving warfarin, particularly for atrial fibrillation, the scale of the problem is likely to be the same. There is no reason why warfarin monitoring cannot become as routine as glucose monitoring in diabetes: relevant small machines are available for generating an INR (with associated standards and quality control).

Overanticoagulation

Excessive anticoagulation without bleeding or with only minor bleeding can be remedied by dose reduction or discontinuation. The risk of bleeding is decreased dramatically by lowering the intended INR from 3-4.5 down to 2-3, although this increases

Patients at high risk of bleeding with warfarin

- Age > 75 years
- History of uncontrolled hypertension (defined as systolic blood pressure > 180mm Hg or diastolic blood pressure > 100 mm Hg)
- Alcohol excess (acute or chronic), liver disease
- Poor drug compliance or clinic attendance
- Bleeding lesions (especially gastrointestinal blood loss, such as peptic ulcer disease, or recent cerebral haemorrhage)
- Bleeding tendency (including coagulation defects, thrombocytopenia) or concomitant use of non-steroidal anti-inflammatory drugs and antibiotics
- Instability of INR control and INR > 3



Computed tomography scan showing intracerebral haemorrhage

Risk of bleeding associated with warfarin treatment

- Rate of bleeding episodes associated in the general patient population is decreasing (possibly due to better management)
- Risk increases with age
- Risk of bleeding is directly related to the achieved intensity of INR rather than the target INR (a clear dose-response effect)
- Temporal association between measured INR and risk of bleeding
- Relative risk of bleeding is increased in patients with cerebrovascular disease and venous thrombosis

the risk of thrombosis. If bleeding becomes substantial, 2-5 mg of oral or subcutaneous vitamin K may be needed. In patients with prosthetic valves, vitamin K should perhaps be avoided because of the risk of valve thrombosis unless there is life threatening intracranial bleeding. Alternatives to vitamin K include a concentrate of the prothrombin group of coagulation factors including II, IX, and X, fresh frozen plasma 15 ml/kg, and recombinant factor VIIa.

Aspirin

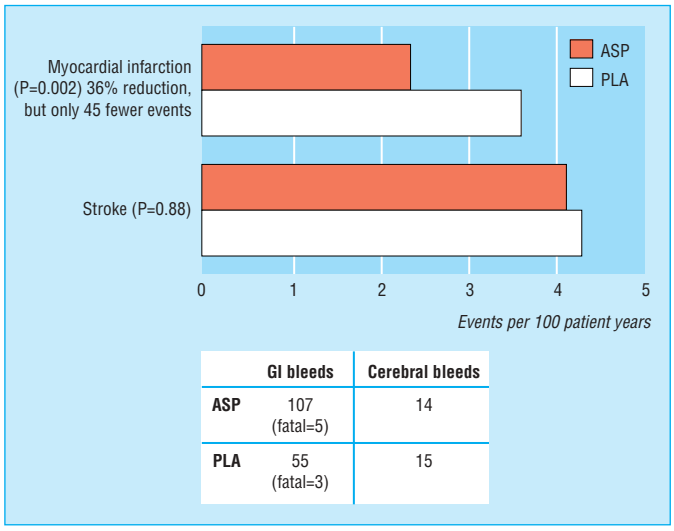
Aspirin has little effect in terms of bruising but can cause serious gastrointestinal bleeding. The risk of gastrointestinal bleeding is related to dose and should not be problematic at doses of 75 mg/day given as thromboprophylaxis. There is currently no consensus as to optimal dose of aspirin for stroke prevention in atrial fibrillation. A meta-analysis of randomised controlled trials using aspirin showed that a mean dose of 273 mg/day, increased absolute risk of haemorrhagic stroke to 12 events per 10 000 people. This relatively small increase must be weighed against the reduced risk of myocardial infarction (to 137 events per 10 000) and ischaemic stroke (to 39 events per 10 000). However, in one trial of patients with well controlled hypertension, use of aspirin 75 mg prevented 1.5 myocardial infarctions per 1000 patients a year, which was in addition to the benefit achieved by lowering the blood pressure, with no effect on stroke. Although there was no increase in the number of fatal bleeding events (seven in patients taking aspirin, compared with eight in the placebo group), there was a 1.8% increase in non-fatal, major bleeding events (129 events in patients taking aspirin, compared with 70 in the placebo group) and minor bleeds (156 and 87, respectively).

Risk of bleeding

There have been conflicting results concerning the role of age as an independent risk factor for haemorrhage induced by anticoagulants. Advanced age (>75 years), intensity of anticoagulation (especially INR >4), history of cerebral vascular disease (recent or remote), and concomitant use of drugs that interfere with haemostasis (aspirin or non-steroidal anti-inflammatory drugs) are probably the most important variables determining patients' risk of major life threatening bleeding complications while they are receiving anticoagulation treatment.

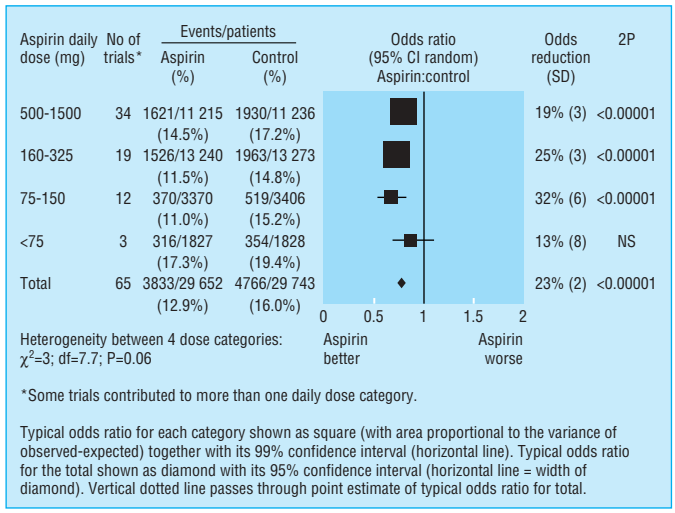
Generally elderly people have increased sensitivity to the anticoagulant effect of warfarin, and require a lower mean daily dose to achieve a given anticoagulant intensity. For example, patients aged >75 years need less than half the daily warfarin dose of patients aged <35 for an equivalent level of anticoagulation. Whatever the mechanism it is clear that warfarin therapy needs careful justification for being given to elderly patients, and the dose needs modification and careful monitoring.

As there is an exponential increase in bleeding risk with a linear increase in anticoagulant effect, there will be a substantial increase in bleeding risk with overanticoagulation. For example, the annual risk of bleeding rises from 1.6% in elderly people not treated with anticoagulant drugs (based on the "Sixty-Plus" study), to 5% (relative risk 3) at an INR of 2.5, and to 50% (relative risk 30) at an INR of 4. In another study, total bleeding events were 39% in a group of 31 patients with an INR of 7 compared with 13% in a group of 100 with a stable INR (odds ratio 5.4, 95% CI 2.1-13.9). The greatest risk factor for being in this group was (apart from having a high target INR) antibiotic therapy in the preceding four weeks.



	GI bleeds	Cerebral bleeds
ASP	107 (fatal=5)	14
PLA	55 (fatal=3)	15

Myocardial infarction, stroke, and bleeding in the hypertension optimal treatment trial (HOT) study (ASP=aspirin, PLA=placebo)



Effect of different doses of aspirin in secondary prevention of vascular events (There is no significant difference in benefit with different aspirin doses, but at higher doses adverse effects are more likely)

Variables that may influence the risk of bleeding in elderly people

- Increased sensitivity to the effect of anticoagulation, perhaps due to increased receptor affinity or lower dietary vitamin K intake
- Concurrent use of drugs that increase bleeding risk
- Associated comorbidity and other diseases that decrease compliance and increase the risk of bleeding

Possible reasons for increased sensitivity to anticoagulation in elderly people

- Lower body weight
- Differences in pharmacokinetics, with a tendency towards reduced drug clearance in the elderly either due to decreases in renal or hepatic blood flow and function with age per se or disease processes
- Change in receptor sensitivity
- Lower dietary vitamin K intake in the elderly may perhaps be the more important cause

Multiple drug therapy or polypharmacy is quite common, with the consequence of adverse drug interactions, the risk of which rises exponentially with the number of drugs given simultaneously and with concurrent diseases. Typical drug interactions include changes in absorption across intestinal mucosae and hepatic metabolism. Patients should be cautioned about the risk of warfarin-drug interactions when their medication list is altered. The decline in cognitive function in some elderly patients may mean they do not realise that some drugs can interact with anticoagulants and so they do not mention their use of oral anticoagulants to doctors or pharmacists. However, elderly patients are likely to attend clinic less often than younger patients, suggesting a greater degree of INR stability.

Many diseases associated with stroke and thromboembolism are more common with increasing age. Older patients are often at highest risk, and appropriate anticoagulation therapy reduces morbidity and mortality. Careful and continuing evaluation of patients is necessary to ensure that the risks of bleeding do not outweigh the benefits from anticoagulation.

The diagram showing the results of the hypertension optimal treatment trial is adapted from Hansson L, et al. *Lancet* 1998;351:1755-62. The figure showing the effect of different doses of aspirin in secondary prevention of vascular events is reproduced from *Clinical Evidence* (June issue 7), BMJ Publishing Group, 2002.

Further reading

- Blann AD, Hewitt J, Siddique F, Bareford D. Racial background is a determinant of average warfarin dose required to maintain the INR between 2.0 and 3.0. *Br J Haematol* 1999;10:207-9
- Erhardtsten E, Nony P, Dechavanne M, Ffrench P, Boissel JP, Hedner U. The effect of recombinant factor VIIa (NovoSeven™) in healthy volunteers receiving acenocoumarol to an International Normalized Ratio above 2.0. *Blood Coag Fibrin* 1998;9:741-8
- Fitzmaurice DA, Hobbs FDR, Murray ET, Hodder, RL, Allan TF, Rose, PE. Oral anticoagulation management in primary care with the use of computerised decision support and near-patient testing. A randomised controlled trial. *Arch Intern Med* 2000;160:2323-48
- Gurwitz JH, Goldberg RJ, Holden A, Knapic N, Ansell J. Age-related risks of long term oral anticoagulant therapy. *Arch Intern Med* 1988;148:1733-6
- He J, Whelton PK, Vu B, Klag MJ. Aspirin and risk of haemorrhagic stroke. *JAMA* 1998;280:1930-5
- Haemostasis and Thrombosis Task Force of the British Society for Haematology. Guidelines on oral anticoagulation: third edition. *Br J Haematol* 1998;101:374-87
- Landefeld CS, Beyth RJ. Anticoagulant related bleeding: clinical epidemiology, prediction, and prevention. *Am J Med* 1993;95:315-28
- Levine MN, Hirsh J, Landefeld CS, Raskob G. Haemorrhagic complications of anticoagulant treatment. *Chest* 1992;102:352-63S
- Panneerselvan S, Baglin C, Lefort W, Baglin T. Analysis of risk factors for over-anticoagulation in patients receiving long-term warfarin. *Br J Haematol* 1998;103:422-4
- Palareti G, Leali N, Coccheri S, Poggi M, Manotti C, D'Angelo A, et al. Bleeding complications of oral anticoagulant treatment: an inception-cohort, prospective collaborative study (ISCOAT). *Lancet* 1996;348:423-8
- van der Meer FJM, Rosendaal FR, Vandenbroucke, Briet E. Bleeding complications in oral anticoagulant therapy. *Arch Int Med* 1993;153:1557-62
- Hutton BA, Lensing AWA, Kraaijenhagen RA, Prins MH. Safety of treatment with oral anticoagulants in the elderly. *Drugs and Aging* 1999;14:303-12

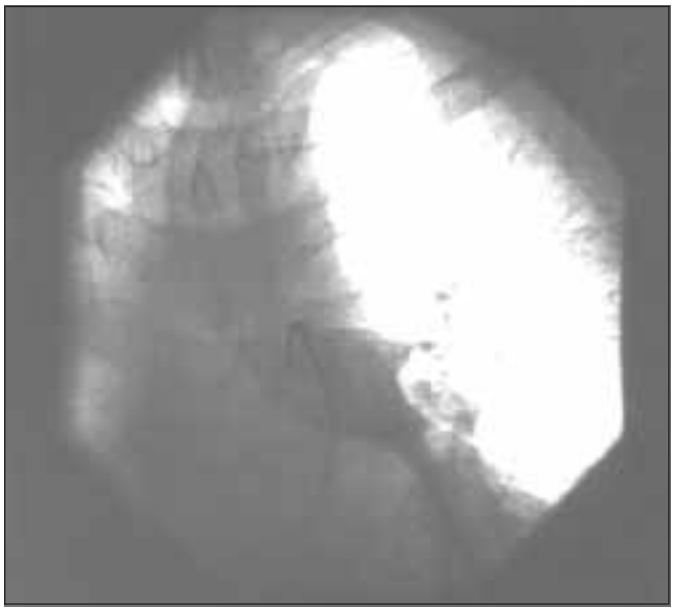
3 Venous thromboembolism: pathophysiology, clinical features, and prevention

Alexander G G Turpie, Bernard S P Chin, Gregory Y H Lip

Venous thromboembolism is a common complication among hospital inpatients and contributes to longer hospital stays, morbidity, and mortality. Some venous thromboembolisms may be subclinical, whereas others present as sudden pulmonary embolus or symptomatic deep vein thrombosis. Ultrasonic Doppler and venographic techniques have shown deep vein thrombosis of the lower limb to occur in half of all major lower limb orthopaedic operations performed without antithrombotic prophylaxis. Deep vein thrombosis of the lower limb is also seen in a quarter of patients with acute myocardial infarction, and more than half of patients with acute ischaemic stroke.

Deep vein thrombosis of the lower limb normally starts in the calf veins. About 10-20% of thromboses extend proximally, and a further 1-5% go on to develop fatal pulmonary embolism. Appropriate antithrombotic measures can reduce this complication. Until recently, some clinicians were reluctant to provide such prophylaxis routinely. As unfounded fears of major bleeding complications from anticoagulant regimens wane, preventive treatments are used more often with medical and surgical patients. However, the risk of bleeding can be serious and this has particular bearing in postoperative patients.

Venous thromboembolism can also arise spontaneously in ambulant individuals, particularly if they have associated risk factors such as thrombophilia, previous thrombosis, or cancer. However, in over half of these patients, no specific predisposing factors can be identified at presentation.



Pulmonary angiography showing large pulmonary embolus in left pulmonary artery

Venous thromboembolism often manifests clinically as deep vein thrombosis or pulmonary embolism, and is possibly one of the preventable complications that occur in hospitalised patients

Pathophysiology

Thrombus formation and propagation depend on the presence of abnormalities of blood flow, blood vessel wall, and blood clotting components, known collectively as Virchow's triad. Abnormalities of blood flow or venous stasis normally occur after prolonged immobility or confinement to bed. Venous obstruction can arise from external compression by enlarged lymph nodes, bulky tumours, or intravascular compression by previous thromboses. Increased oestrogens at pharmacological levels, as seen with oral contraceptive use and with hormone replacement therapy in postmenopausal women, have been associated with a threefold increased risk in the small initial risk of venous thromboembolism. Cancers, particularly adenocarcinomas and metastatic cancers, are also associated with increased venous thromboembolism. Indeed, on presentation, some idiopathic venous thromboembolisms have revealed occult cancers at follow up. Both oestrogens at pharmacological levels and cancer can also activate the clotting system.

Risk factors and conditions predisposing to venous thromboembolism

- History of venous thromboembolism
- Prolonged immobility
- Prolonged confinement to bed or lower limb paralysis
- Surgery, particularly lower limb orthopaedic operations, and major pelvic or abdominal operations
- Trauma—For example, hip fractures and acute spinal injury
- Obesity
- Major medical illnesses such as acute myocardial infarction, ischaemic stroke, congestive cardiac failure, acute respiratory failure
- Oestrogen use in pharmacological doses—For example, oral contraception pills, hormone replacement therapy
- Cancer, especially metastatic adenocarcinomas
- Age >40 years
- Aquired hypercoagulable states—Lupus anticoagulant and antiphospholipid antibodies, hyperhomocysteinaemia, dysfibrinogenaemia, myeloproliferative disorders such as polycythaemia rubra vera
- Inherited hypercoaguable states—Activated protein C resistance (factor V Leiden mutation), protein C deficiency, protein S deficiency, antithrombin deficiency, prothrombin gene mutation

Clinical presentation and diagnosis

Deep vein thrombosis

Deep vein thrombosis commonly presents with pain, erythema, tenderness, and swelling of the affected limb. Thus, in lower limb deep vein thrombosis, the affected leg is usually swollen with the circumference of the calf larger than the unaffected side. Other causes of leg swelling, erythema, and tenderness include a ruptured Baker's cyst and infective cellulitis. The

diagnosis of deep vein thrombosis is therefore more likely when risk factors are present and less so if there are features suggesting alternative diagnoses. For example, ruptured Baker's cysts commonly appear in the context of osteoarthritis and rheumatoid arthritis. Infective cellulitis is unlikely to be bilateral, with clearly demarcated areas of erythema extending proximally. Breaks in the skin, particularly between the toes, and coexistent fungal infection are additional clues to cellulitis.

Objective diagnosis of venous thromboembolism is important for optimal management. Although the clinical diagnosis of venous thromboembolism is imprecise, various probability models based on clinical features have proved to be practical and reliable (interobserver reliability, $\kappa = 0.85$) in predicting the likelihood of venous thromboembolism. These models should be used in conjunction with objective diagnostic tests.

Compression ultrasonography remains the non-invasive investigation of choice for the diagnosis of clinically suspected deep vein thrombosis. It is highly sensitive in detecting proximal deep vein thrombosis although less accurate for isolated calf deep vein thrombosis. In patients with suspected thrombosis and a negative compression ultrasound result, the test should be repeated in seven days because studies have shown that patients with two or more negative tests over a week who are untreated have a less than 2% risk of proximal extension or subsequent deep vein thrombosis.

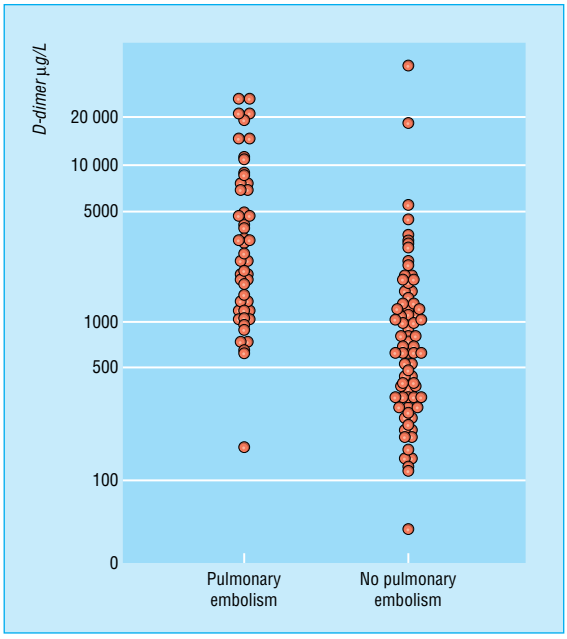
Impedance plethysmography is slightly less specific and sensitive than ultrasonography but may still have a role in pregnant women and suspected recurrent deep vein thrombosis. The gold standard is invasive contrast venography, which is still used when a definitive answer is needed. Newer imaging techniques are being developed, and tools such as magnetic resonance venography or computed tomography could possibly detect pelvic vein thromboses, but further testing is needed to establish their role in the diagnosis of deep vein thrombosis.

Blood tests such as fibrin D-dimer add to the diagnostic accuracy of the non-invasive tests. In one study, the sensitivity and specificity of a D-dimer concentration of >500 µg/l for the presence of pulmonary embolism were 98% and 39%, respectively, which give positive and negative predictive values of 44% and 98%. The sensitivity of the test even remained high at three and seven days after presentation (96% and 93%).

Modified pretest probability for deep vein thrombosis

Clinical feature	Score
Tenderness along entire deep vein system	1.0
Swelling of the entire leg	1.0
Greater than 3 cm difference in calf circumference	1.0
Pitting oedema	1.0
Collateral superficial veins	1.0
Risk factors present:	
Active cancer	1.0
Prolonged immobility or paralysis	1.0
Recent surgery or major medical illness	1.0
Alternative diagnosis likely (ruptured Baker's cyst in rheumatoid arthritis, superficial thrombophlebitis, or infective cellulitis)	−2.0

Score >3=high probablility; 1-2=moderate probability <0=low probability



Plasma D-dimer concentrations on day of presentation according to final diagnosis

Investigations for suspected venous thromboembolism by pretest clinical probability

Pretest probability	Fibrin D-dimer	Other investigations	Comment
Deep vein thrombosis:			
Low	Negative		No further investigations needed
Low or moderate	Positive	Negative ultrasound compression	Consider venography or repeat ultrasound after a week
Moderate or high	Positive	Positive ultrasound compression	Treat with anticoagulants
High	Positive	Negative ultrasound compression	Consider venography to rule out deep vein thrombosis, especially in high risk patients and those with recurrent pulmonary emboli
Pulmonary embolism:			
Low	Negative	No	No further investigation needed
Low or moderate	Positive	Proceed to ventilation-perfusion scan	If non-diagnostic ventilation-perfusion scan consider serial compression ultrasound over two weeks to rule out venous thromboembolism
High	Positive	Proceed to ventilation-perfusion scan (or ultrasound)	If non-diagnostic ventilation-perfusion scans, proceed to venography or pulmonary angiography as needed

Pulmonary embolism

Patients presenting with acute pulmonary embolism often complain of sudden onset of breathlessness with haemoptysis or pleuritic chest pain, or collapse with shock in the absence of other causes. Deep vein thrombosis may not be suspected clinically, but its presence, along with thrombotic risk factors, will make the diagnosis of pulmonary embolism more likely. A similar clinical probability model to that for deep vein thrombosis has been developed for pulmonary embolism.

Pulmonary angiography is the gold standard investigation for pulmonary embolism, but it is invasive and associated with 0.5% mortality. A ventilation-perfusion scan using technetium DTPA (ditriaminopentaric acid) is more widely used. However this investigation is non-specific, and is diagnostic in only 30% of cases. Spiral computed tomography scans are more reliable but diagnosis is limited to emboli in larger vessels only. Measurement of fibrin D-dimer levels, used for deep vein thrombosis, is helpful, as is compression ultrasound in the detection of occult deep vein thrombosis.

Clinical probability for pulmonary embolism

Clinical feature	Score
Deep vein thrombosis suspected:	
Clinical features of deep vein thrombosis	3.0
Recent prolonged immobility or surgery	1.5
Active cancer	1.0
History of deep vein thrombosis or pulmonary embolism	1.5
Haemoptysis	1.0
Resting heart rate >100 beats/min	1.5
No alternative explanation for acute breathlessness or pleuritic chest pain	3.0

>6=high probability (60%); 2-6=moderate probability (20%); <1.5=low probability (3-4%)

Prevention strategies

An appropriate strategy for the prevention of venous thromboembolism include pharmacological or physical methods. To optimise treatment, patients should be stratified into risk categories to allow the most appropriate prophylactic measure to be used.

Prophylactic drugs include unfractionated heparin, low molecular weight heparin, oral anticoagulants (such as coumarins), thrombin inhibitors (such as hirudin), and specific factor Xa inhibitors (such as fondaparinux). The recently approved fondaparinux reduces the risk of venous thromboembolism after orthopaedic surgery by more than half compared with low molecular weight heparin and seems likely to become the treatment of choice after universal availability.

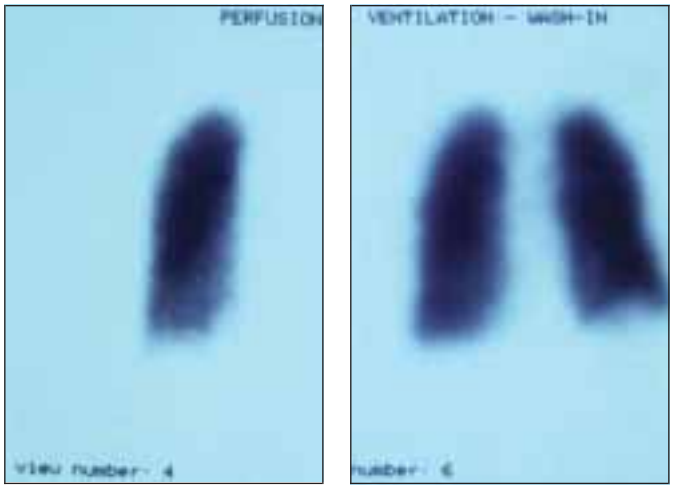
Prophylactic physical methods include the use of compression elastic stockings, intermittent pneumatic compression (which provides rythmic external compression at 35-40 mm Hg for about 10 seconds every minute), and early mobilisation to improve venous blood flow in conditions predisposing to venous stasis.

Thromboembolic risk stratification for surgery patients

- *Low risk*—Uncomplicated surgery in patients aged <40 years with minimal immobility postoperatively and no risk factors
- *Moderate risk*—Any surgery in patients aged 40-60 years, major surgery in patients <40 years and no other risk factors, minor surgery in patients with one or more risk factors
- *High risk*—Major surgery in patients aged >60 years, major surgery in patients aged 40-60 years with one or more risk factors
- *Very high risk*—Major surgery in patients aged >40 years with previous venous thromboembolism, cancer or known hypercoagulable state, major orthopaedic surgery, elective neurosurgery, multiple trauma, or acute spinal cord injury

General surgery

Patients at low risk undergoing general surgery do not need specific prophylaxis other than early mobilisation. In moderate risk patients, fixed low doses of unfractionated heparin (5000 IU every 12 hours) or low molecular weight heparin (3400 anti-Xa units or equivalent) once daily is sufficient. Higher doses of low molecular weight heparin (more than 3400 IU anti-Xa daily) should be reserved for high risk general surgery and orthopaedic operations. Compression elastic stockings and intermittent pneumatic compression may protect high risk patients when used with anticoagulants. They are also effective when used alone in moderate risk patients where anticoagulants are contraindicated.



Ventilation-perfusion scan showing massive pulmonary thromboembolism, showing a mismatch between (left) perfusion and (right) ventilation scans

Orthopaedic surgery

In very high risk patients, such as those undergoing major orthopaedic operations, high dose low molecular weight heparin or warfarin is appropriate. The current recommended length of anticoagulant prophylaxis is 7-10 days with low molecular weight heparin or warfarin. Extended use may provide additional benefit. Routine screening with duplex ultrasonography is not helpful. Hirudin seems to be superior to low molecular weight heparin and low dose unfractionated

Key points

- Understanding Virchow's triad aids the treatment of venous thromboembolism
- Numerous situations and risk factors can contribute to venous thromboembolism
- Diagnosis of venous thromboembolism depends upon a combination of history, risk factors, and investigations
- Antithrombotic prophylaxis is safe and effective

heparin as prophylaxis in patients undergoing elective hip replacements but is still not universally available.

Neurosurgery, multiple traumas, and spinal cord injuries
Intermittent pneumatic compression is the prophylaxis of choice for elective neurosurgery. Among the low molecular weight heparins, only enoxaparin 30 mg twice daily has been shown to reduce venous thromboembolism without excess bleeding after elective neurosurgery, multiple traumas, or spinal cord injuries and so may be used in these situations. Other low molecular weight heparins have either not been tested or have not conclusively been shown to be of benefit in this setting.

Medical conditions
In general medical patients including heart failure and respiratory failure, both unfractionated heparin and low molecular weight heparin have been shown to be effective in reducing the risk of venous thromboembolism. Low molecular weight heparin has been shown to be more effective than heparin in stroke. Low dose heparin has been shown to be effective in acute myocardial infarction but this is now largely historic because myocardial infarction patients receive therapeutic dose anticoagulants.

Other considerations
Combined approaches using drugs and physical methods may be better at preventing thromboembolism than physical methods alone. However, compression elastic stockings and intermittent pneumatic compression may be used for moderate or high risk patients when anticoagulation is contraindicated or best avoided. Inferior vena cava filter placement should be reserved for patients at very high risk of venous thromboembolism where anticoagulation as well as physical methods are contraindicated. Inferior vena cava filter placement tends to cause a long term increase of recurrent deep vein thrombosis, even though the immediate risk of postoperative pulmonary embolism is reduced.

Further reading

- Haemostasis and Thrombosis Task Force of the British Society for Haematology. Guidelines on anticoagulation: third edition. *Br J Haematol* 1998;101:374-87
- Hyers TM, Agnelli G, Hull RD, Morris TA, Samama M, Tapson V, et al. Antithrombotic therapy for venous thromboembolic disease. *Chest* 2001;119:176-93S
- Kearon C, Hirsh J. Management of anticoagulation before and after elective surgery. *N Engl J Med* 1997;336:1506-11
- Simonneau G, Sors H, Charbonnier B, Page Y, Labaan JP, Azarian R et al. A comparison of low molecular weight heparin with unfractionated heparin for acute pulmonary embolism. *N Engl J Med* 1997:337;663-9
- Turpie AG, Bauer KA, Eriksson BI, Lassen MR. Fondaparinus vs enoxaparin for the prevention of venous thromboembolism in major orthopedic surgery: a meta-analysis of 4 randomised double blind studies. *Arch Intern Med* 2002;162:1833-40
- Walker ID, Greaves M, Preston FE. Guideline: Investigation and management of heritable thrombophilia. *Br J Haematol* 2001:114;512-28

Evidence based use of antithrombotic prophylaxis

General surgery
- *Low risk*— Early mobilisation
- *Moderate risk*—UH 5000 IU 12 hourly starting two hours before surgery, or low molecular weight heparin <3400 anti-Xa IU daily*, or compression elastic stockings, or intermittent pneumatic compression
- *High risk*—Low molecular weight heparin >3400 anti-Xa IU daily† plus compression elastic stockings, or unfractionated heparin 5000 IU eight hourly starting two hours before surgery plus compression elastic stockings, or intermittent pneumatic compression if anticoagulation contraindicated
- *Very high risk*—Perioperative warfarin (INR 2-3), low molecular weight heparin >3400 anti-Xa IU daily† plus compression elastic stockings, or prolonged low molecular weight heparin therapy plus compression elastic stockings

Major orthopaedic surgery
- *Elective hip replacement*—Recombinant hirudin 15 mg twice daily, unfractionated heparin 3500 IU eight hourly with postoperative adjustments (APTT 1.2-1.5), or low molecular weight heparin >3400 anti-Xa IU daily†, or perioperative warfarin (INR 2-3), or fondaparinux 2.5 mg daily
- *Elective knee replacement*—Low molecular weight heparin >3400 anti-Xa IU daily†, or perioperative warfarin (INR 2-3), or fondaparinux 2.5 mg daily, intermittent pneumatic compression
- S*urgery for hip fracture*—Low molecular weight heparin >3400 anti-Xa IU daily†, or perioperative warfarin (INR 2-3), or fondaparinux 2.5 mg daily

Elective neurosurgery
- Intermittent pneumatic compression, enoxaparin 30 mg twice daily

Acute spinal cord injury
- Enoxaparin 30 mg twice daily

Trauma
- Enoxaparin 30 mg twice daily

Acute myocardial infarction
- Low dose unfractionated heparin 5000 IU twice daily, full dose unfractionated heparin 40 000 IU infusion over 24 hours, elastic stockings, and early mobilisation

Ischaemic stroke
- Low dose unfractionated heparin 5000 IU twice daily

Other medical conditions including congestive heart failure
- Enoxaparin 40 mg once daily or 30 U twice daily, dalteparin 2500 IU daily, low dose unfractionated heparin 5000 IU twice daily

Cancer patients receiving chemotherapy
- Low dose warfarin (INR <2), dalteparin 2500 IU daily

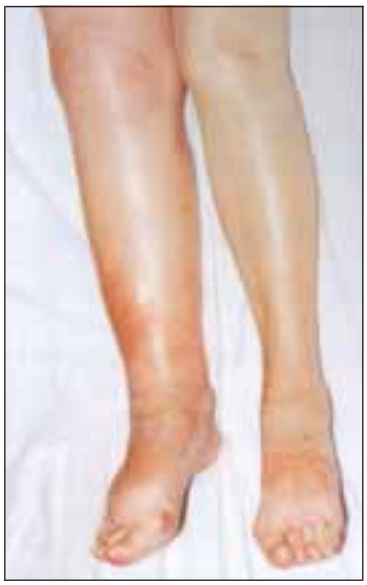
*Dalteparin 2500 IU once daily starting two hours before surgery
Enoxaparin 20 mg once daily starting two hours before surgery
Nadroparin 3100 IU once daily starting two hours before surgery
Tinzaparin 3500 IU once daily starting two hours before surgery
†Dalteparin 5000 IU once daily starting 10-12 hours before surgery
Danaparoid 750 IU twice daily starting one to two hours before surgery
Enoxaparin 40 mg once daily starting 10-12 hours before surgery or 30 mg twice daily starting after surgery
Tinzaparin 50 IU/kg once daily starting two hours before surgery

The box showing evidence based use of antithrombotic prophylaxis is adapted from the 6th ACCP guidelines Geerts WH, et al. *Chest* 2001;119:132-75S. The figure showing Plasma D-dimer concentrations on day of presentation according to final diagnosis is adapted from Bounameaux H, et al. *Lance*t 1991;337:196-200.

4 Venous thromboembolism: treatment strategies

Alexander G G Turpie, Bernard S P Chin, Gregory Y H Lip

Pulmonary embolism and deep vein thrombosis are treated using similar drugs and physical methods. The efficacy of intravenous infusion of unfractionated heparin was first proved in a randomised trial in 1960. Subsequently, trials concentrated on the dose, duration of infusion, mode of administration, and combination with warfarin treatment. Later trials have reported the efficacy and cost effectiveness of low molecular weight heparin compared with unfractionated heparin.



Right ileofemoral deep vein thrombosis

Unfractionated heparin

Unfractionated heparin, administered by continuous infusion or subcutaneous injections adjusted to achieve activated partial thromboplastin time (APTT) greater than 1.5, is effective as initial treatment of venous thromboembolism. Initial heparinisation should be followed by long term anticoagulation with oral anticoagulants. APTT is a global coagulation test and not specific for heparin, and it is also influenced by various plasma proteins and clotting factors. Measuring plasma heparin levels is more accurate but it is impractical and expensive. A sensible approach is to standardise the APTT with plasma heparin within each laboratory.

The most common mistake when starting heparin treatment is failure to achieve adequate anticoagulation. APTT ratios of less than 1.5 during the first few days of heparin therapy increase the long term risk of venous thromboembolism recurrence. Hence, the initial bolus dose should be adequate and APTT monitored every six hours during the first 24 hours of heparin infusion.

Oral anticoagulants may be started at the same time and should be continued for at least three to six months, depending on the individual. The optimal duration of intravenous heparin treatment is five to seven days because this is the time needed to obtain an adequate and persistent reduction in the vitamin K dependent clotting factors with oral anticoagulants such as warfarin. Heparin can then be stopped when concomitant use with warfarin has achieved an international normalised ratio (INR) of 2-3 for at least 48 hours. In patients with large ileofemoral vein thromboses or major pulmonary embolism, heparin infusion can be continued for up to 10 days.

Heparin use for more than five to six days is associated with a rare risk of thrombocytopenia. In a recent trial, only one of 308 patients (0.32%) who received unfractionated heparin for acute pulmonary embolism developed a thrombocytopenia, whereas none of 304 patients receiving low molecular weight heparin had this problem. The thrombocytopenia is normally mild, but precipitous falls in platelet count to less than 100×10^9/l can occur. When this happens, antibody mediated injury to platelets should be suspected. As this condition may be associated with arterial or venous thromboembolism, heparin should be stopped and warfarin use delayed. Alternative anticoagulation cover should be given by danaparoid, a heparinoid, or hirudin, a thrombin inhibitor, until the platelet count rises above 100 000 and it is safe to start warfarin. Unfractionated heparin has also been reported to increase platelet activation in vivo: low molecular weight heparin had no such effect.

Antithrombotic treatment is often inadequate in the first few days, predisposing to recurrences. Anticoagulation with warfarin after discharge should continue for at least three months, possibly six months. Low molecular weight heparin is as efficacious as unfractionated heparin in prophylaxis and treatment

Initial antithrombotic therapy for deep vein thrombosis with unfractionated heparin

1. Check baseline APTT, prothrombin time, full blood count
2. Confirm there are no contraindications to heparin therapy
3. Intravenous bolus 5000 IU
4. Choose between:
   Continuous unfractionated heparin infusion—Start infusion at 18 IU/kg/hour (~30 000/24 hours in a 70 kg man)
   Check APTT every six hours for first 24 hours, then daily thereafter
   Aim for APTT 1.5-2.5 × normal
   Recheck APTT at six hours after each adjustment
   Continue infusion for five to seven days
   Subcutaneous—Start at 17 500 IU every 12 hours (or 250 IU/kg every 12 hours)
5. Check platelet count daily for thrombocytopenia
6. Warfarin therapy can be started on the first day of heparin therapy according to local protocol
7. Continue heparin for at least four to five days after starting warfarin
8. Stop heparin when INR greater than 2 for more than 48 hours
9. Continue warfarin therapy for at least three months keeping INR between 2 and 3 (target 2.5)

Low molecular weight heparin

Low molecular weight heparin has a more predictable relation between dose and response than unfractionated heparin and does not need monitoring or adjustments if the dose is based on patient body weight. Low molecular weight heparin is also associated with lower risk of thrombocytopenia. Its use in deep vein thrombosis and pulmonary embolism is now firmly established: many trials and meta-analysis have confirmed its superior efficacy, safer profile, and greater cost effectiveness over unfractionated heparins. However, all low molecular weight heparins are different, and trials for one product cannot be extrapolated to another. The introduction of low molecular weight heparin has advanced antithrombotic therapy by providing effective anticoagulation without the need for monitoring or adjustments. It also allows patients with uncomplicated deep vein thrombosis to be treated in the community, saving an average of four to five days' admission per patient.

Advantages of low molecular weight heparin over unfractionated heparin

- More reliable relation between dose and response
- Does not need monitoring
- Does not need dose adjustments
- Lower incidence of thrombocytopenia
- No excess bleeding
- Can be administered by patient at home
- Saves about five to six days' admission per patient

Coumarins

Warfarin is the most widely used oral anticoagulant for treating venous thromboembolism. It is well absorbed from the gut, metabolised in the liver, and excreted in urine. The lag time for warfarin to take effect may be related to the natural clearance of normal clotting factors from plasma. Of the vitamin K dependent clotting factors, factor II takes the longest to clear. Warfarin monitoring is performed using an INR rather than prothrombin time, which may vary between laboratories. Warfarin interacts with many other drugs and alcohol. It is also teratogenic and may induce spontaneous abortion.

A target INR range of 2-3 is standard for treatment of venous thromboembolism. Higher levels tend to increase incidence of bleeding without reducing recurrent thromboembolism and so are unnecessary. The exception to this is for patients with the antiphospholipid antibody syndrome, where the risk of recurrent venous thromboembolism is high. Here, an INR of 3-4.5 is recommended. Warfarin should be started in conjunction with heparin or low molecular weight heparin when the diagnosis of venous thromboembolism is confirmed, although local protocols may vary in their starting doses and titration schedule. As indicated, heparin should be continued concomitantly for five days and until INR is >2.

Warfarin therapy should then be maintained for at least three months in all patients. However, it has recently been established that longer treatments (such as six months) may be necessary. Patients without a readily identifiable risk factor (idiopathic venous thromboembolism) have higher rates of recurrences. These recurrences can be reduced by prolonged anticoagulation. However, there is a corresponding rise in bleeding complications with prolonged anticoagulation. Current recommendations advocate anticoagulation for at least six months for the first presentation of idiopathic venous thromboembolism. Patients with recurrent venous thromboembolism and hypercoagulable states (acquired or inherited) or with cancer (especially while receiving chemotherapy) should take anticoagulation therapy for at least a year, and perhaps indefinitely.

Duration of anticoagulation therapy for venous thromboembolism*

Three to six months
- First event with reversible† or time limited risk factor (patient may have underlying factor V Leiden or prothrombin 20210 mutation)

More than six months
- Idiopathic venous thromboembolism, first event

A year to life time
- First event‡ with cancer (until resolved), anticardiolipin antibody, antithrombin deficiency
- Recurrent event, idiopathic or with thrombophilia

*All recommendations are subject to modification by individual characteristics including patient preference, age, comorbidity, and likelihood of recurrence
†Reversible or time limited risk factors such as surgery, trauma, immobilisation, and oestrogen use
‡Proper duration of therapy is unclear in first event with homozygous factor V Leiden, homocystinaemia, deficiency of protein C or S, or multiple thrombophilias; and in recurrent events with reversible risk factors

Recent trials with the oral thrombin inhibitor, ximelagatran suggest that, in certain circumstances, this agent may be an alternative to warfarin for the management of venous thromboembolism, without the need for anticoagulation monitoring

Thrombolytic therapy

Unlike heparin and warfarin, which prevent extension and recurrences of thrombosis, thrombolytic agents (including streptokinase, urokinase, and tissue plasminogen activator) lyse the thrombi. It is therefore unsurprising that patients with

Thrombolytic regimens for pulmonary embolism

1. Check suitability of patient for thrombolysis
2. Choose between:
   Streptokinase—250 000 IU loading dose then 100 000 IU/hour for 24 hours
   Urokinase—4400 IU/kg loading dose then 2200 IU/kg/hour for 12 hours
   Alteplase—100 mg intravenously over an hour
3. Check APTT two to four hours after starting infusion:
   >10 seconds prolongation indicates active fibrinolysis
4. Start heparin at 5000-10 000 IU loading followed by 15-25 units/kg/hour when APTT < 2
5. Adjust according to local protocol to keep APTT 1.5-2.5

pulmonary embolism treated with streptokinase and urokinase are three times more likely to show clot resolution than patients taking heparin alone. Even so, thrombolytic therapy of pulmonary embolism does not dissolve the clot completely as it does with acute coronary thrombosis, and increases the risk of bleeding. Occasionally, thrombolytic therapy is administered via a catheter placed in the pulmonary artery. The catheter can be used to "disrupt" the thrombus before starting the drug.

Until there is more evidence that thrombolytic therapy reduces mortality in pulmonary embolism, this treatment should be reserved for patients with massive pulmonary embolism, cardiorespiratory compromise, and low risk of bleeding. Evidence is emerging that streptokinase can decrease swelling and pain in deep vein thrombosis. Again, further trials are needed before this can be recommended routinely.

Physical methods

Non-drug treatments include physically preventing embolisation of the thrombi and extraction of thromboemboli (usually from the pulmonary vasculature).

Inferior vena cava filters may be used when anticoagulation is contraindicated in patients at high risk of proximal deep vein thrombosis extension or embolisation. The filter is normally inserted via the internal jugular or femoral vein. It is then advanced under fluoroscopic guidance to the inferior vena cava. Filters are now available that are easy to insert, and complications are low in skilled hands. For now, this technique should be considered in patients with recurrent symptomatic pulmonary embolism and as primary prophylaxis of thromboembolism in patients at high risk of bleeding (such as patients with extensive trauma or visceral cancer), although the evidence is based on uncontrolled case series. The only randomised trial showed a reduction in pulmonary embolism but no improvement in short or long term survival, because of greater risk of recurrent deep vein thrombosis in patients who received a filter.

Other mechanical and surgical treatments are usually reserved for massive pulmonary embolism where drug treatments have failed or are contraindicated. None of these methods has shown a long term reduction in mortality, but better techniques have led to acceptable complication rates and warrant further evaluation.

Treatment during pregnancy

Unfractionated heparin and low molecular weight heparin do not cross the placenta and are probably safe for the fetus during pregnancy. Oral anticoagulants cross the placenta and can cause fetal bleeding and malformations. Pregnant women with venous thromboembolism can be treated with therapeutic doses of subcutaneous heparin or low molecular weight heparin until after delivery, when warfarin can be used safely. These issues are developed in chapter 14.

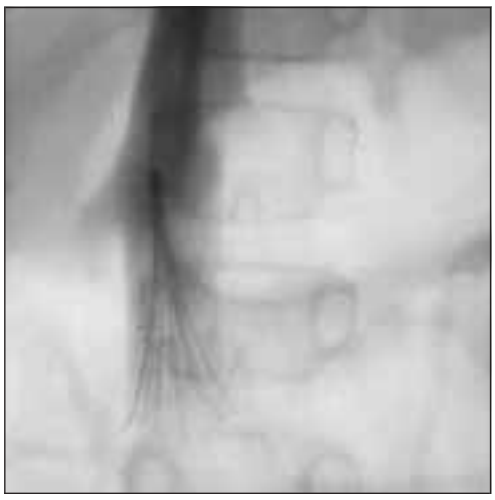
The data on duration of anticoagulation therapy for venous thromboembolism is adapted from the 6th ACCP guidelines Hyers TM, et al. Antithrombotic therapy for venous thromboembolic disease. *Chest* 2001;119:176-93S

Indications for inferior vena cava filter placement

- Patients at high risk of proximal deep vein thrombosis extension where anticoagulation is contraindicated
- Recurrent venous thromboembolism despite adequate anticoagulation
- Chronic recurrent venous thromboembolism with pulmonary hypertension
- Simultaneous surgical pulmonary embolectomy or endarterectomy

Mechanical and surgical treatment of pulmonary embolism

- Inferior vena cava filter placement
  Indications—See box above
- Pulmonary embolectomy
  Indication—Massive pulmonary embolism compromising cardiac output where thrombolysis has failed or is contraindicated
  Experienced cardiac surgical cover essential
  Where available, catheter transvenous extraction of emboli may be an alternative to pulmonary embolectomy
- Pulmonary endarterectomy
  Indication—Chronic recurrent pulmonary embolism with secondary pulmonary hypertension



Vena cavagram showing umbrella delivery device for filter inserted into the inferior vena cava through the jugular vein

Further reading

- Decousus H, Leizorovicz A, Parent F, Page Y, Tardy B, Girard P, et al. A clinical trial of vena caval filters in the prevention of pulmonary embolism in patients with proximal deep vein thrombosis. *N Engl J Med* 1998;338:409-15
- Geerts WH, Heit JA, Clagett GP, Pineo GF, Colwell CW, Anderson FA Jr, et al. Prevention of venous thromboembolism. *Chest* 2000;119:132-75S
- Heit JA, O'Fallon WM, Petterson T, Lohse CM, Silverstein MD, Mohr DN, et al. Relative impact of risk factors for deep vein thrombosis and pulmonary embolism. *Arch Intern Med* 2002;162:1245-8
- Levine M, Gent M, Hirsch J, Leclerc J, Anderson D, Weitz J, et al. A comparison of low-molecular-weight heparin administered primarily at home with unfractionated heparin administered in the hospital for proximal deep-vein thrombosis. *N Engl J Med* 1996;334:677-81
- Walker ID, Greaves M, Preston FE. Guideline: investigation and management of heritable thrombophilia. *Br J Haematol* 2001;114:512-28

5 Antithrombotic therapy for atrial fibrillation: clinical aspects

Gregory Y H Lip, Robert G Hart, Dwayne S G Conway

Atrial fibrillation is the commonest sustained disorder of cardiac rhythm. Although patients often present with symptoms caused by haemodynamic disturbance associated with the rhythm itself, the condition carries an increased risk of arterial thromboembolism and ischaemic stroke due to embolisation of thrombi that form within the left atrium of the heart. Presence of the arrhythmia confers about a fivefold increase in stroke risk, an absolute risk of about 4.5% a year, although the precise annual stroke risk ranges from $<1\%$ to $>12\%$, according to the presence or absence of certain clinical and echocardiographically identifiable risk factors.

From trial data, patients with paroxysmal atrial fibrillation seem to carry the same risk as those with persistent atrial fibrillation. The same criteria can be used to identify high risk patients, although it is unclear whether the risk is dependent on the frequency and duration of the paroxysms.

Severely damaged left atrial appendage endocardial surface with thrombotic mass in a patient with atrial fibrillation and mitral valve disease

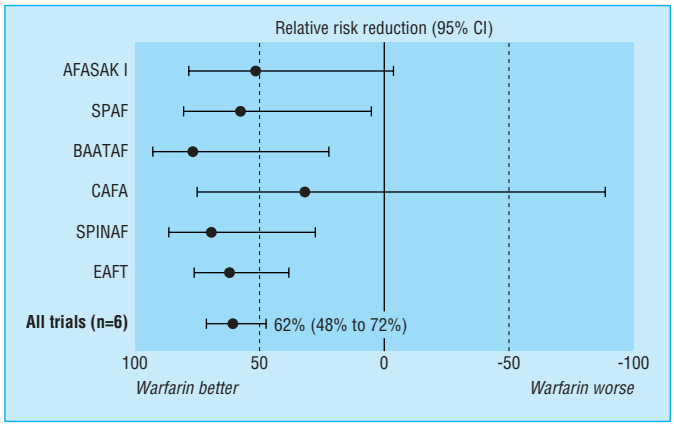
Evidence from clinical trials

It is well established that antithrombotic therapy confers thromboprophylaxis in patients with atrial fibrillation who are at risk of thromboembolism. A recent meta-analysis of antithrombotic therapy in atrial fibrillation showed that adjusted dose warfarin reduced stroke by about 60%, with absolute risk reductions of 3% a year for primary prevention and 8% a year for secondary prevention (numbers needed to treat for one year to prevent one stroke of 33 and 13, respectively). In contrast, aspirin reduced stroke by about 20%, with absolute risk reductions of 1.5% a year for primary prevention and 2.5% a year for secondary prevention (numbers needed to treat of 66 and 40, respectively). Relative to aspirin, adjusted dose warfarin reduced the risk by about 40%, and the relative risk reduction was similar for primary and secondary prevention, and for disabling and non-disabling strokes. However, these data, obtained from well planned clinical trials recruiting patients with relatively stable conditions, are unlikely to be fully extrapolable to all patients in general practice, so that some caution is advised.

Randomised controlled trials have shown the benefit of warfarin and, to a lesser extent, aspirin in reducing the incidence of stroke in patients with atrial fibrillation without greatly increasing the risk of haemorrhagic stroke and extracranial haemorrhage. However, anticoagulant therapy is still underprescribed in patients with atrial fibrillation, particularly in elderly patients, who stand to benefit most

Overall, warfarin (generally at a dose to maintain an international normalised ratio (INR) of 2-3) is significantly more effective than aspirin in treating atrial fibrillation in patients at high risk of stroke, especially in preventing disabling cardioembolic strokes. The effect of aspirin seems to be on the smaller, non-cardioembolic strokes from which elderly, and often hypertensive, patients with atrial fibrillation are not spared.

Recent clinical trials have suggested that there is no role for minidose warfarin (1 mg/day regardless of INR), alone or in combination with antiplatelet agents or aspirin, as thromboprophylaxis in atrial fibrillation. However, the role of other antiplatelet agents (such as indobufen and dipyridamole) in atrial fibrillation is still unclear. One small trial (SIFA) compared treatment with indobufen, a reversible cyclo-oxygenase inhibitor, with full dose warfarin for secondary prevention and found no statistical difference between the two groups, who were well matched for confounding risk factors. Trials of other antiplatelet and antithrombotic drugs (including low molecular weight heparin) have been performed but have generally been too small and underpowered to show significant differences. Large

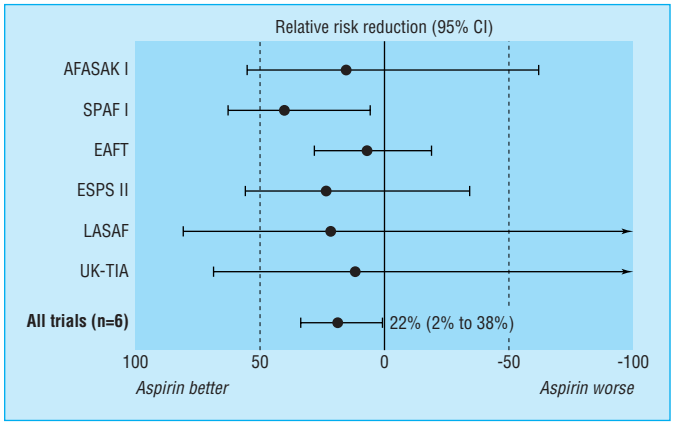


Meta-analysis of trials comparing warfarin with placebo in reducing the risk of thromboembolism in patients with atrial fibrillation
AFASAK=Copenhagen atrial fibrillation, aspirin, and anticoagulation study; BAATAF=Boston area anticoagulation trial for atrial fibrillation; CAFA=Canadian atrial fibrillation anticoagulation study; EAFT=European atrial fibrillation trial; SPAF=Stroke prevention in atrial fibrillation study; SPINAF=Stroke prevention in non-rheumatic atrial fibrillation

multinational trials comparing a direct thrombin inhibitor (ximelagatran) with adjusted dose warfarin in over 7000 patients with atrial fibrillation at high risk of stroke and thromboembolism suggest that this agent may be an alternative to warfarin, without the need for anticoagulation monitoring.

The reduction in relative risk with warfarin applies equally to primary and secondary prevention but, as history of stroke confers an increased annual stroke risk (12% *v* 4.5%), the absolute risk reduction is greater for secondary prevention. The number of patients with atrial fibrillation needing treatment with warfarin to prevent one stroke is therefore about three times greater in primary prevention (37) than in secondary prevention (12).

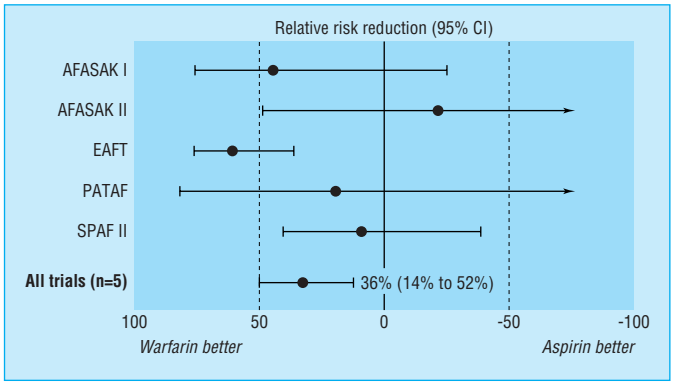
Treatment with full dose anticoagulation carries the potential risk of major bleeding, including intracranial haemorrhage. Meta-analysis of the initial five primary prevention trials plus a further secondary prevention trial suggests the risk of haemorrhagic stroke is only marginally increased from 0.1% to 0.3% a year. Higher rates of major haemorrhage were seen in elderly patients and those with higher intensity anticoagulation. Further recent trials have confirmed an increased bleeding risk in patients with INR >3.



Meta-analysis of trials comparing aspirin with placebo in reducing risk of thromboembolism in patients with atrial fibrillation
AFASAK=Copenhagen atrial fibrillation, aspirin, and anticoagulation study; EAFT=European atrial fibrillation trial; ESPS II= European stroke prevention study II; LASAF=Low-dose aspirin, stroke, and atrial fibrillation pilot study; SPAF=Stroke prevention in atrial fibrillation study; UK-TIA=United Kingdom TIA study

Antiplatelet therapy in atrial fibrillation

Several clinical trials have studied the effects of aspirin in atrial fibrillation, with doses ranging from 25 mg twice daily to 1200 mg a day. Overall, aspirin reduces the relative risk of stroke by about 20% (a figure which just reaches statistical significance) with no apparent benefit of increasing aspirin dose. Aspirin seems to carry greater benefit in reducing smaller non-disabling strokes than disabling strokes. This may be due to an effect primarily on carotid and cerebral artery platelet thrombus formation, rather than on formation of intra-atrial thrombus. A meta-analysis of trials directly comparing full dose warfarin with aspirin confirmed significant reductions in stroke risk about three times greater with warfarin. The SPAF III trial demonstrates that addition of fixed low doses of warfarin to aspirin treatment is not sufficient to achieve the benefits of full dose warfarin alone.



Meta-analysis of trials comparing warfarin with aspirin in reducing risk of thromboembolism in patients with atrial fibrillation
AFASAK=Copenhagen atrial fibrillation, aspirin, and anticoagulation study; EAFT=European atrial fibrillation trial; PATAF=Prevention of arterial thromboembolism in atrial fibrillation; SPAF=Stroke prevention in atrial fibrillation study

Putting the evidence into practice

Despite the evidence from the trials, many doctors are reluctant to start warfarin treatment for patients with atrial fibrillation. This could be due to fears (of patient and doctor) of haemorrhagic complications in an elderly population, logistical problems of INR monitoring, and a lack of consensus guidelines on which patients to treat and the ideal target INR. Such attitudes may result in otherwise avoidable stroke and arterial thromboembolism. A systematic evidence based approach needs to be encouraged, targeting appropriate antithrombotic therapy at those patients who stand to gain most benefit (those at greatest risk of thromboembolism) and using levels of anticoagulation that have been proved both effective and reasonably safe for both primary and secondary prevention of stroke, if we are to realise in clinical practice the large reduction in incidence of stroke achieved in the clinical trials.

Who to treat?

Even though there are impressive figures for relative risk reduction with warfarin, the figures for absolute risk reduction (more important in clinical practice) depend greatly on the underlying risk of stroke if untreated. Elderly patients are often denied anticoagulant therapy because of fears of increased haemorrhage risk. However, the benefits of anticoagulant therapy are greater for elderly patients because of the increased

Independent predictors of ischaemic stroke in non-valve atrial fibrillation

Consistent predictors
- Old age
- Hypertension
- Previous stroke or transient ischaemic attack
- Left ventricular dysfunction*

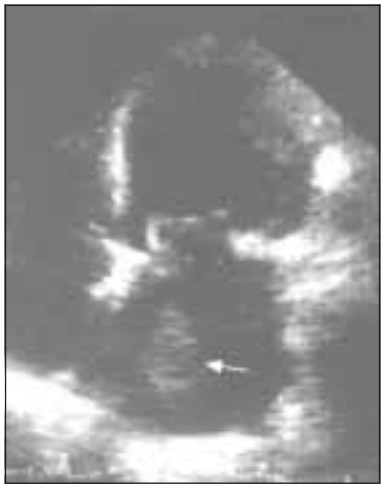
Inconsistent predictors
- Diabetes
- Systolic blood pressure >160 mm Hg†
- Women, especially older than 75 years
- Postmenopausal hormone replacement therapy
- Coronary artery disease

Factors which decrease the risk of stroke
- Moderate to severe mitral regurgitation
- Regular alcohol use (>14 drinks in two weeks)

*Recent clinical congestive cardiac failure or moderate to severe systolic dysfunction on echocardiography
†In some analyses, systolic blood pressure >160 mm Hg remained an independent predictor after adjustment for hypertension

underlying thromboembolic risk. Conversely, young patients at relatively low risk of stroke have less to gain from full dose anticoagulation as there may be little difference between the number of strokes prevented and the number of haemorrhagic complications. Risk stratification is possible using the clinical and echocardiographic parameters and can be used to target treatment at the most appropriate patients.

Risk stratification for thromboprophylaxis can be undertaken in many ways. Clinical risk factors would assist with risk stratification in most cases. Although echocardiography is not mandatory, it would help refine risk stratification in cases of uncertainty. Based on echocardiographic data on 1066 patients, the Atrial Fibrillation Investigators reported that the only independent predictor of stroke risk was moderate or severe left ventricular dysfunction on two dimensional echocardiography. Left atrial size on M mode echocardiography was not an independent predictor on multivariate analysis. Transoesophageal echocardiography is rarely needed to undertake risk stratification, but "high risk" features include the presence of dense spontaneous echocardiographic contrast



Two dimensional echocardiography showing left atrial thrombus in patient with prosthetic valve

Practical guidelines for antithrombotic therapy in non-valvar atrial fibrillation

Assess risk, and reassess regularly

High risk (annual risk of cerebrovascular accident=8-12%)
- All patients with previous transient ischaemic attack or cerebrovascular accident
- All patients aged ≥75 with diabetes or hypertension
- All patients with clinical evidence of valve disease, heart failure, thyroid disease, and impaired left ventricular function on echocardiography*

Treatment—Give warfarin (target INR 2-3) if no contraindications and possible in practice

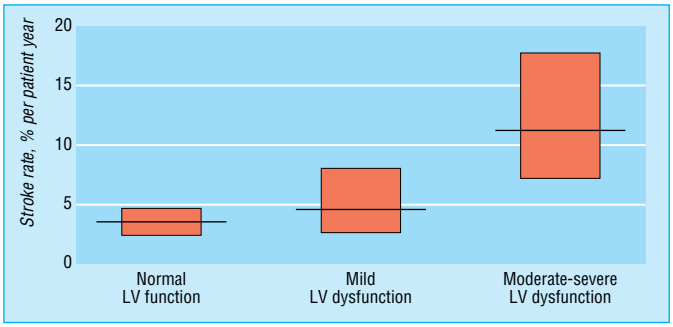
Moderate risk (annual risk of cerebrovascular accident=4%)
- All patients <65 with clinical risk factors: diabetes, hypertension, peripheral vascular disease, ischaemic heart disease
- All patients >65 not in high risk group

Treatment—Either warfarin (INR 2-3) or aspirin 75-300 mg daily. In view of insufficient clear cut evidence, treatment may be decided on individual cases. Referral and echocardiography may help

Low risk (annual risk=1%)
- All patients aged <65 with no history of embolism, hypertension, diabetes, or other clinical risk factors

Treatment—Give aspirin 75-300 mg daily

*Echocardiogram not needed for routine risk assessment but refines clinical risk stratification in case of moderate or severe left ventricular dysfunction (see figure below) and valve disease. A large atrium per se is not an independent risk factor on multivariate analysis



Effect of left ventricular funtion on stroke rate in atrial fibrillation
LV=left ventricle

Different risk stratification schemes for primary prevention of stroke in non-valvar atrial fibrillation

	Risk		
Study	**High**	**Intermediate**	**Low**
Atrial Fibrillation Investigators (1994)	High to intermediate risk: Age >65 years; History of hypertension; Diabetes		Age <65 years; No high risk features
American College of Chest Physicians Consensus (1998)	Age >75 years; History of hypertension; Left ventricular dysfunction†; >1 moderate risk factor	Age 65-75 years; Diabetes; Coronary disease (thyrotoxicosis)*	Age <65 years; No risk factors
Stroke Prevention in Atrial Fibrillation	Women aged ≥75 years; Systolic blood pressure >160 mm Hg; Left ventricular dysfunction‡	History of hypertension; No high risk features	No high risk features; No history of hypertension
Lip (1999)	Patients aged >75 years and with diabetes or hypertension; Patients with clinical evidence of heart failure, thyroid disease, and impaired left ventricular function on echocardiography§	Patients aged <65 years with clinical risk factors: diabetes, hypertension, peripheral arterial disease, ischaemic heart disease; Patients aged >65 not in high risk group	Patients aged <65 years with no risk factors

*Patients with thyrotoxicosis were excluded from participation in the test cohort
†Moderate to severe left ventricular dysfunction on echocardiography
‡Recent congestive heart failure or fractional shortening ≤25% by M mode echocardiography
§Echocardiography not needed for routine risk assessment but refines clinical risk stratification in case of impaired left ventricular function and valve disease

(often with low atrial appendage velocities, indicating stasis), the presence of thrombus of the atrial appendage, and complex aortic plaque.

Which INR range?

The evidence suggests that INR levels greater than 3 may result in an excess rate of haemorrhage, whereas low dose warfarin regimens (with INR maintained below 1.5) do not achieve the reductions in stroke of higher doses. An INR range of between 2 and 3 has been shown to be highly effective without leading to excessive haemorrhage and should therefore be recommended for all patients with atrial fibrillation treated with warfarin unless they have another indication for higher levels of anticoagulation (such as a mechanical heart valve). Although INR monitoring is often coordinated by hospital based anticoagulant clinics, general practitioners are likely to play a more important part with the development of near patient INR testing.

Particular care must be taken and INR levels closely monitored when warfarin is used in elderly patients. It has been suggested that an INR of between 1.6 and 2.5 can provide substantial, albeit partial, efficacy (estimated to be nearly 90% of the highest intensities). Given the uncertainty about the safety of INRs >2.5 for atrial fibrillation patients over 75 years, a target INR of 2 (range 1.6-2.5) may be a reasonable compromise between an increased risk of haemorrhage and a reduced risk of thrombotic stroke for some patients within this age group, in the absence of additional risk factors, pending further data about the safety of higher intensities.

The ongoing MRC sponsored Birmingham atrial fibrillation trial of anticoagulation in the aged (BAFTA) is comparing warfarin with aspirin in atrial fibrillation patients over 75 years to further define the relative benefits and risks.

DC cardioversion

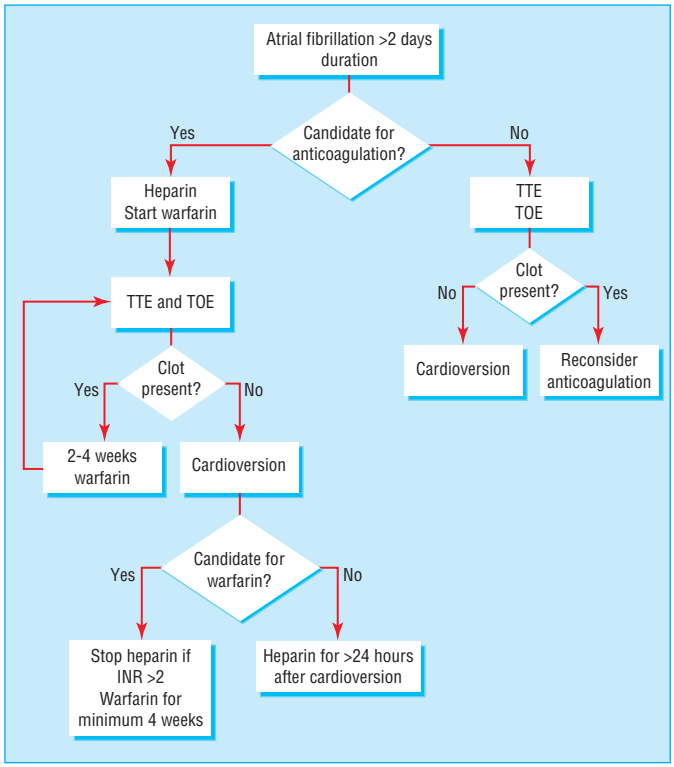
No hard evidence exists in the literature that restoration of sinus rhythm by whatever means reduces stroke risk. Transoesophageal echocardiography performed immediately before cardioversion (to exclude intra-atrial thrombus) may allow DC cardioversion to be performed without prior anticoagulation. However, as the thromboembolic risk may persist for a few weeks postprocedure, it is still recommended that patients receive warfarin for at least four weeks afterwards.

The figures showing a severely damaged left atrial appendage endocardial surface is reproduced from Goldsmith I et al, *Am Heart J* 2000;140:777-84 with permission from Mosby Inc. The figures showing results of trials comparing warfarin with placebo, aspirin with placebo, and warfarin with aspirin are adapted from Hart RG et al, *Ann Intern Med* 1999;131:492-501. The independent predictors of ischaemic stroke are adapted from Hart RG et al, *Ann Intern Med* 1999;131:688-95. The practical guidelines for antithrombotic therapy in non-valvar patients is adapted from Lip GYH, *Lancet* 1999;353:4-6. The table containing risk stratification schemes for primary prevention of stroke is adapted from Pearce LA et al, *Am J Med* 2000;109:45-51. Guidelines for transoesophageal echocardiography guided cardioversion is adapted from the ACUTE Study, *N Engl J Med* 2001;344:1411-20. The recommendations for anticoagulation for cardioversion of atrial fibrillation are based on the 6th ACCP Consensus Conference on Antithrombotic Therapy. Albers GW et al, *Chest* 2001;119:194-206S.

Recommendations for anticoagulation for cardioversion of atrial fibrillation

- For elective cardioversion of atrial fibrillation of >48 hours duration start warfarin treatment (INR 2-3) three weeks before and continue for four weeks after cardioversion
- In urgent and emergency cardioversion administer intravenous heparin followed by warfarin
- Treat atrial flutter similarly
- No anticoagulation treatment is required for supraventricular tachycardia or atrial fibrillation of <48 hours duration
- Continue anticoagulation in patients with multiple risk factors or those at high risk of recurrent thromboembolism

Based on the 6th ACCP Consensus Conference on Antithrombotic Therapy



Guidelines for transoesophageal echocardiography guided cardioversion TOE=transoesophageal echocardiography; TTE=transthoracic echocardiography

Further reading

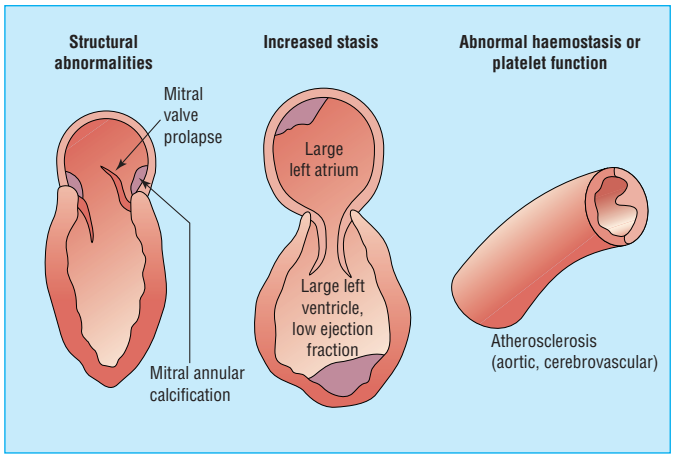
- Hart RG, Pearce LA, Rothbart RM, McAnulty JH, Asinger RW, Halperin JL. Stroke with intermittent atrial fibrillation: incidence and predictors during aspirin therapy. Stroke Prevention in Atrial Fibrillation Investigators. *J Am Coll Cardiol* 2000;35:183-7
- Lip GYH. Does atrial fibrillation confer a hypercoagulable state? *Lancet* 1995;346:1313-4
- Lip GYH. Thromboprophylaxis for atrial fibrillation. *Lancet* 1999; 353:4-6
- Lip GYH, Kamath S, Freestone B. Acute atrial fibrillation. *Clinical Evidence* June Issue 7. BMJ Publishing Group. 2002
- Straus SE, Majumdar SR, McAlister FA. New edvidence for stroke prevention: scientific review. *JAMA* 2002;288:1388
- Thomson R, Parkin D, Eccles M, Sudlow M, Robinson A. Decision analysis and guidelines for anticoagulant therapy to prevent stroke in patients with atrial fibrillation. *Lancet* 2000;355:956-62

6 Antithrombotic therapy for atrial fibrillation: pathophysiology, acute atrial fibrillation, and cardioversion

Gregory Y H Lip, Robert G Hart, Dwayne S G Conway

Pathophysiology of thromboembolism in atrial fibrillation

The pathophysiological mechanism for thrombus formation and embolism seems to be abnormalities in blood flow within the fibrillating (and possibly dilated) left atrium. These abnormalities predispose to thrombus formation and arterial embolism, especially in the presence of underlying heart disease. The latter (including valvar heart disease, hypertensive heart disease, and poor left ventricular function) substantially increases the risk for stroke and thromboembolism in patients with atrial fibrillation. For example, the thromboembolic risk of atrial fibrillation is 18 times greater if valvar heart disease is present. In addition, a history of stroke, transient ischaemic attack, or other thromboembolism substantially increases the risk of stroke in atrial fibrillation (by 2.5 times). Hypertension and diabetes are also common risk factors for stroke, increasing the risk of stroke in atrial fibrillation by nearly twofold.



Virchow's triad of thrombogenesis needs the presence of structural abnormalities (for example–atherosclerosis, valve disease), abnormal flow (stasis in left atria, heart failure), and abnormal blood constituents (for example–clotting factors, platelets, etc). All are present in "high risk" patients with atrial fibrillation

Anatomical aspects

A left atrial diameter of ≥ 4.0 cm was previously regarded as the strongest single predictor of increased risk of thromboembolisation, but atrial dilatation rarely occurs in isolation without associated clinical risk factors such as hypertension. Thus, in the most recent analysis from the Atrial Fibrillation Investigators, isolated left atrial dilatation by M mode echocardiography was not independently predictive of stroke and thromboembolism on multivariate analysis. Nevertheless, patients with lone atrial fibrillation (that is, those who have no underlying cause for their arrhythmia), have been shown to have a low risk of stroke and usually have an atrial size < 4.0 cm. Left atrial enlargement has also been associated with the presence of "spontaneous echocontrast" on transoesophageal echocardiography. With scanning electron microscopy, the endocardium of the left atrial appendage shows evidence of damage in mitral valve disease, especially in the presence of atrial fibrillation.

In patients with paroxysmal atrial fibrillation, thromboembolic events seem to cluster in the transition from atrial fibrillation to sinus rhythm, perhaps reflecting embolisation of a preformed clot

Mechanical aspects

The loss of atrial systolic function reduces stroke volume, leading to a corresponding reduction in cardiac output and increased atrial stasis. The latter results in increased propensity to thrombus formation. The prevalence of thrombus in the left atrial appendage, detected as an incidental finding during transoesophageal echocardiography, has been reported to be about 10-15% in patients admitted with acute atrial fibrillation and up to 30% in patients with atrial fibrillation and recent stroke.

Indeed, after cardioversion from atrial fibrillation to sinus rhythm, there is risk of thromboembolism of about 7% if anticoagulation is not used, with the highest risk one to two weeks after cardioversion. This may reflect mechanical reasons (embolisation of preformed thrombus), but it is more likely that this increased risk is related to atrial dysfunction after cardioversion ("stunning") and the delay of the return of atrial systolic function, which can be up to three weeks or more after cardioversion.

Thrombus in the left atrial appendage of patient with mitral valve disease at surgery

Atrial fibrillation and hypercoagulability

It has been recognised for over 150 years that abnormalities in the blood vessel wall, blood flow, and blood constituents (Virchow's triad) may increase the propensity for thrombus formation. Clinical and echocardiographic criteria can help identify the first two of Virchow's postulates for thrombogenesis—namely, abnormalities of blood flow and vessels, such as valvar heart disease and cardiac impairment. Patients with atrial fibrillation also show abnormalities of haemostatic and platelet markers that are unrelated to aetiology and underlying structural heart disease (and alter with antithrombotic therapy and cardioversion), which point towards the presence of a hypercoagulable state in this common arrhythmia. Thus, atrial fibrillation has been described as an arrhythmia which confers a prothrombotic or hypercoaguable state.

Anticoagulation for atrial fibrillation in special circumstances

Acute atrial fibrillation

In patients presenting with de novo atrial fibrillation, a clear history of arrhythmia onset is needed to guide appropriate antithrombotic therapy and timing of cardioversion.

Although no randomised trials have specifically addressed the issue, there is evidence that cardioversion may be safely performed without anticoagulation if the arrhythmia has been present for <48 hours. However, in one series intra-atrial thrombus was detected by transoesophageal echocardiography in about 15% of patients presenting with acute atrial fibrillation (apparent duration <48 hours), raising the possibility that the development of intra-atrial thrombus may be faster than previously suspected, or that in many apparent cases of acute atrial fibrillation the arrhythmia developed asymptomatically >48 hours before. Thus, in cases of uncertainty, anticoagulation is needed. Again, no randomised prospective studies have addressed the use of intravenous unfractionated heparin or subcutaneous low molecular weight heparin derivatives, but both drugs have been used with good results in the acute and pericardioversion periods.

Atrial fibrillation patients presenting with acute stroke

The role of antiplatelet drugs after acute stroke in sinus rhythm is well proved, but there is less certainty about the potential benefits and hazards of anticoagulant treatment in patients with atrial fibrillation, particularly the timing of administration. Although the benefits of secondary stroke prevention using warfarin in atrial fibrillation patients are dramatic, it must be certain that there is no ongoing intracerebral haemorrhage (or risk of new intracerebral haemorrhage) before starting the drug.

Previous consensus guidelines from the American College of Chest Physicians state that before any antithrombotic drug is started computed tomography or magnetic resonance imaging scan should be done to confirm the absence of intracranial haemorrhage and to assess the size of any cerebral infarction. In atrial fibrillation patients with no evidence of haemorrhage and small infarct size (or no evidence of infarction) warfarin (INR 2.0-3.0) can be given with minimal risk, provided patients are normotensive. In atrial fibrillation patients with large areas of cerebral infarction, the start of warfarin treatment should be delayed for two weeks because of the potential risk of haemorrhagic transformation. The presence of intracranial haemorrhage is an absolute contraindication to the immediate and future use of anticoagulation for stroke prevention in atrial fibrillation. The mortality benefits of aspirin treatment in acute stroke seen in the international stroke trial and Chinese acute

The development of intra-atrial thrombus, and thus the immediate risk of thromboembolism, is thought to be temporally related to the duration of the arrhythmia, with minimal risk if the arrhythmia has been present for <48 hours

Benefits of anticoagulant treatment in patients with non-rheumatic atrial fibrillation in preventing stroke

Stroke risk	NNT (95% CI) to prevent one stroke
Low: Age <65 years, no major risk factors (including previous stroke, systemic embolism, or transient ischaemic attack; hypertension; and poor left ventricular function as determined by a clinical history of heart failure or left ventricular ejection fraction <50%)	Aspirin 227 (132 to 2500)
Low moderate: Age 65-75 years, no major risk factors	Aspirin 152 (88 to 1667) Warfarin 54 (46 to 69)
High moderate: Age 65-75 years, no major risk factors but either diabetes or coronary heart disease	Warfarin 32 (28 to 42)
High: Age <75 years with hypertension, left ventricular dysfunction, or both, or age >75 without other risk factors	Warfarin 14 (12 to 17)
Very high: Age >75 years with hypertension, left ventricular dysfunction, or both, or any age and previous stroke, transient ischaemic attack, or systemic embolism	Warfarin 8 (7 to 10)

NNT = number needed to treat

stroke trial were less marked in patients with atrial fibrillation, presumably because of the presence of preformed intra-atrial thrombus rather than new localised platelet thrombus adhering to carotid and cerebral artery atheroma.

Cardioversion of persistent atrial fibrillation

Although there are no randomised studies to show that successful cardioversion of atrial fibrillation reduces the number of subsequent thromboembolic events, the improvement in haemodynamic function and observed reduction in indices of clotting suggest that this may be the case. However, cardioversion is known to increase the short term risk of thromboembolism, and thus, unless the arrhythmia has been present for less than 48 hours, thromboprophylactic measures are needed. The mechanism behind pericardioversion thromboembolism is complex and not entirely understood, but it is likely to be associated with the return of atrial systole, temporary "stunning" of the left atrium before return of systolic function, and possibly an increase in thrombotic tendency caused by the procedure itself. The increase in thromboembolic risk may therefore persist for two weeks or more after successful cardioversion.

The American College of Chest Physicians (ACCP) sixth recommendations for pericardioversion anticoagulation have been summarised in the previous chapter. However, the recent ACUTE study found that by excluding thrombus on transoesophageal echocardiography before cardioversion, the need for prior anticoagulation could be safely avoided. Patients treated in this manner had similar rates of thromboembolism as those treated with the standard antithrombotic regimen but their haemorrhage rates were reduced. Transoesophageal echocardiography guided technique also allowed faster cardioversion of patients and resulted in higher initial success rates, although by eight weeks there was no substantial difference in death rates, maintenance of sinus rhythm, or in functional status between the two groups. Transoesophageal echocardiography guided cardioversion is now regarded by many as the optimum approach to cardioversion and is recognised as a suitable alternative to standard practice by the ACCP. This point is included in the recent American Heart Association (AHA)/American College of Cardiology (ACC)/European Society of Cardiology (ESC) guideline recommendations.

Recent trials comparing a "rate control" strategy with a "rhythm control" strategy for persistent atrial fibrillation showed an excess of thromboembolism in the patients randomised to rhythm control (that is—cardioversion), as such events happened in patients successfully cardioverted, the anticoagulation stopped and on recurrence of atrial fibrillation, thromboembolism occurred. Thus, anticoagulation should be considered long term in patients postcardioversion at high risk of stroke and thromboembolim, or high arrhythmia recurrence risk after cardioversion.

The box showing the benefits of anticoagulant treatment in patients with non-rheumatic atrial fibrillation in preventing stroke is adapted from Straus SE, et al. *JAMA* 2002;288:1388. The table showing the rate versus rhythm in atrial fibrillation: ischaemic strokes is adapted from Verheught FWA et al. *J Am Coll Cardiol* 2003;41:130A. The table showing ischaemic stroke in the AFFIRM study is adapted from the AFFIRM Investigators *New Engl J Med* 2002;347:1825-33. The box showing the recommendations for antithrombotic therapy to prevent ischaemic stroke and systemic embolism in patients with atrial fibrillation undergoing cardioversion is adapted from the ACC/AHA/ESC guidelines *Eur Heart J* 2001;22:1852-93

Rate versus rhythm in atrial fibrillation: ischaemic strokes

Study	n	Rate control (%)	Rhythm control	Relative ratio (95% CI)	p
AFFIRM	4917	5.7	7.3	1.28 (0.95 to1.72)	0.12
RACE	522	5.5	7.9	1.44 (0.75 to 2.78)	0.44
STAF	266	1.0	3.0	3.01 (0.35 to 25.30)	0.52
PIAF	252	0.8	0.8	1.02 (0.73 to 2.16)	0.49
Total	5957	5.0	6.5	1.28 (0.98 to 1.66)	0.08

Ischaemic stroke in the AFFIRM study

	Rhythm control	Rate control
Ischaemic stroke	84 (7.3%)*	79 (5.7%)*
INR ⩾2.0	18 (22%)	24 (30%)
INR <2.0	17 (20%)	28 (35%)
Not taking warfarin	48 (58%)	26 (33%)
Atrial fibrillation at time of event	25 (36%)	45 (69%)

*Event rates derived from Kaplan Meier analysis, p = 0.680

Electrical cardioversion of atrial fibrillation

Recommendations for antithrombotic therapy to prevent ischaemic stroke and sytemic embolism in patients with atrial fibrillation undergoing cardioversion*

Class 1

- Administer anticoagulation therapy regardless of the method (electrical or pharmacological) used to restore sinus rhythm
- Anticoagulate patients with atrial fibrillation lasting more that 48 hours or of unknown duration for at least three to four weeks before and after cardioversion (INR 2.0-3.0)
- Perform immediate cardioversion in patients with acute (recent onset) atrial fibrillation accompanied by symptoms or signs of haemodynamic instability resulting in angina pectoris, myocardial infarction, shock, or pulmonary oedema, without waiting for prior anticoagulation
  - —If not contraindicated, administer heparin concurrently by an initial intravenous bolus injection followed by a continuous infusion in a dose adjusted to prolong the activated partial thromboplastin time at 1.5-2.0 times the reference control value
  - —Next, provide oral anticoagulation (INR 2.0-3.0) for at least three to four weeks, as for patients who are undergoing elective cardioversion
  - —Limited data from recent studies support subcutaneous administration of low molecular weight heparin in this indication
- Screening for the precence of thrombus in the left atrium or left atrial appendage by transoesophageal echocardiography is an alternative to routine preanticoagulation in candidates for cardioversion of atrial fibrillation
  - —Anticoagulate patients in whom no thrombus is identified with intravenous unfractionated heparin by an initial bolus injection before cardioversion, followed by a continuous infusion in a dose adjusted to prolong the activated partial thromboplastin time at 1.5-2.0 times the reference control value
  - —Next, provide oral anticoagulation (INR 2.0-3.0) for at least three to four weeks, as for patients who are undergoing elective cardioversion
  - —Limited data from recent studies support subcutaneous administration of low molecular weight heparin in this indication
  - —Treat patients in whom thrombus is identified by transoesophageal echocardiography with oral anticoagulation (INR 2.0-3.0) for at least three to four weeks before and after restoration of sinus rhythm

Class IIb

- Cardioversion without transoesophageal echocardiography guidance during the first 48 hours after the onset of atrial fibrillation
  - —In these cases, anticoagulation before and after cardioversion is optional, depending on assessment of risk
- Anticoagulate patients with atrial flutter undergoing cardioversion in the same way as for patients with atrial fibrillation

ACC/AHA Classification

Class I —Conditions for which there is evidence or general agreement or both that a given procedure or treatment if useful and effective
Class II —Conditions for which there is conflicting evidence or a divergence of opinion about the usefulness and efficacy of a procedure or treatment
Class IIa—Weight of evidence and opinion is in favour of usefulness and efficacy
ClassIIb—Usefulness and efficacy is less well established by evidence and opinion
Class III —Conditions for which there is evidence or general agreement or both that the procedure or treatment is not useful and in some cases may be harmful

*Data from Fuster V et al. *J Am Cardiol* 2001;38:1231.

Further reading

- Rate Control Versus Electrical Cardioversion for Persistent Atrial Fibrillation Study Group. A comparison of rate control and rhythm control in patients with recurrent persistent atrial fibrillation. *New Engl J Med* 2002;347:1834-40
- AFFIRM Investigators. Survival in patients presenting with atrial fibrillation follow up investigation of rhythm management study. *New Engl J Med* 2002;347:1825-33
- Lip GYH. The prothrombotic state in atrial fibrillation: new insights, more questions, and clear answers needed. *Am Heart J* 2000;140:348-50
- Fuster V, Ryden LE, Asinger RW, Cannom DS, Crijns HJ, Frye RL, et al. ACC/AHA/ESC guidelines for the management of patients with atrial fibrillation. A report of the American College of Cardiology, American Heart Association Task Force on practice guidelines and the European society of Cardiology Committee for Practice Guidelines and Policy Conferences developed in collaboration with the North American Society of Pacing and Electrophysiology. *Eur Heart J* 2001;22:1852-923

7 Antithrombotic therapy in peripheral vascular disease

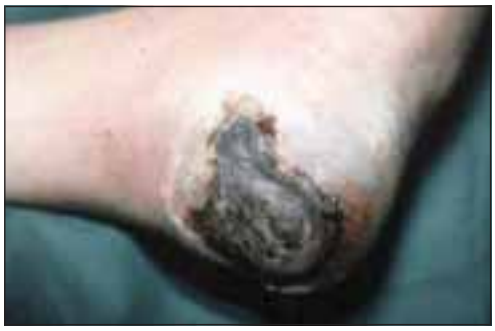
Andrew J Makin, Stanley H Silverman, Gregory Y H Lip

Atherosclerotic peripheral vascular disease is symptomatic as intermittent claudication in 2-3% of men and 1-2 % of women aged over 60 years. However, the prevalence of asymptomatic peripheral vascular disease, generally shown by a reduced ankle to brachial systolic pressure index, is three to four times greater. Peripheral vascular disease is also a significant cause of hospital admission, and is an important predicator of cardiovascular mortality. Pain at rest and critical ischaemia are usually the result of progression of atherosclerotic disease, leading to multilevel arterial occlusion. Other causes of arterial insufficiency—including fibromuscular dysplasia, inflammatory conditions, and congenital malformations—are much rarer. Therapeutic objectives in peripheral vascular disease include relieving symptoms and preventing the disease, and any associated events, progressing.

The symptoms of peripheral vascular disease are progressive. A claudicating patient encouraged to exercise tends to report a symptomatic improvement. This effect is generally not accepted to be an improvement in the diseased segment of blood vessel, but the formation of collateral vessels perfusing the ischaemic tissue.

Vasodilating agents, such as naftidrofuryl, have little value in managing claudication and peripheral vascular disease as their effect is small and does not stop progression of the disease. Cilostazol has been shown to increase absolute walking distance in some patients by up to 47%. However, it has no clear antithrombotic effect and has not been shown to stop disease progression.

Unfortunately, not all progression is amenable to improvement and, without the appropriate risk factor management, progression to rest pain and necrosis can be rapid.



Ischaemic ulcer on foot

Antithrombotic therapy in peripheral vascular disease

Clinical problem	Antithrombotic therapy of choice
Intermittent claudication	Aspirin (to reduce risk of stroke and myocardial infarction) Clopidogrel
Diabetes	Aspirin (to reduce risk of stroke and myocardial infarction) Clopidogrel
Embolic arterial occlusion	Intravenous heparin and emergency surgical intervention
Acute on chronic arterial occlusion	Heparin and angioplasty, intra-arterial thrombolysis or early surgery
Intraoperative anticoagulation during vascular surgery	Heparin
Infrainguinal vein bypass and infrainguinal prosthetic bypass	Aspirin (to reduce risk of stroke and myocardial infarction) Clopidogrel (if unable to take aspirin)
Infrainguinal bypass at high thrombotic risk	Aspirin or consider warfarin
Carotid endarterectomy	Aspirin or clopidogrel
Symptomatic carotid stenosis and too unwell for surgery	Consider warfarin or aspirin plus dipyridamole

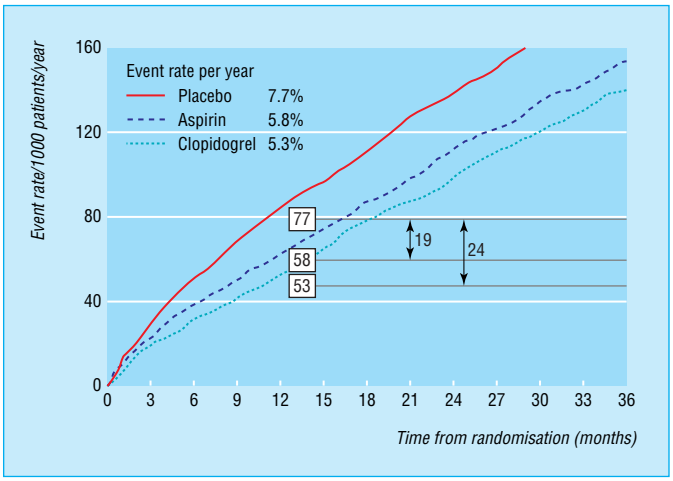
Intermittent claudication

The role of aspirin as an antiplatelet agent has been shown to be beneficial beyond doubt. In peripheral vascular disease it reduces the frequency of thrombotic events in the peripheral arteries and reduces overall cardiovascular mortality in claudicating patients. The dose of aspirin has been the subject of some debate, but 81-325 mg daily has been shown to be of value. Larger doses have no apparent additional benefit but increase the risk of adverse effects. Aspirin has been shown to reduce the progression of atherosclerosis in a few trials, but this remains unsubstantiated.

The role of dipyridamole remains controversial. Several small studies have shown the benefit of giving it in conjunction with aspirin, but it is uncertain if dipyridamole alone is superior to aspirin.

In aspirin intolerant patients there is now a clear role for clopidogrel 75 mg once a day. This is as effective as aspirin in preventing cardiovascular events. If a thrombotic event has occurred (whether the patient is taking aspirin or not) there may be an advantage in using clopidogrel to prevent further events, especially in peripheral vascular disease.

In non-critical peripheral ischaemia, there is no indication for warfarin treatment as the complexities of management and bleeding risks seem to far outweigh the benefits, unless the



Survival curve from CAPRIE study showing the benefits of aspirin and clopidogrel on vascular events, with placebo rates from the Antiplatelet Trialists' Collaboration

patient has concomitant problems needing anticoagulation such as atrial fibrillation.

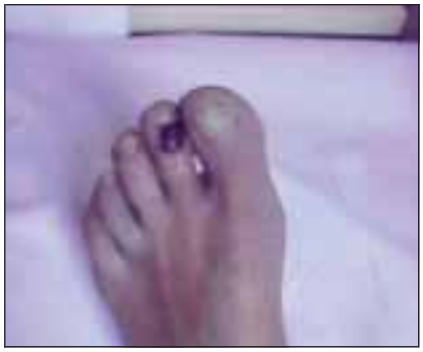
Critical ischaemia

Rest pain and gangrene are markers of critical ischaemia. This is nearly always the result of extensive vessel occlusion with absent pedal pulses. The patient will almost certainly be immobile because of pain and arterial insufficiency making walking impossible. These patients need prophylaxis against venous thromboembolism.

Patients giving a short history of rest pain of sudden onset require full, immediate anticoagulation with low molecular weight heparin or intravenous unfractionated heparin (the latter with a target activated partial prothrombin time (APTT) ratio of 1.5-2.5). Warfarin should be avoided initially until investigations and possible interventions are complete.

Patients with chronic, progressive pain at rest also need full anticoagulation. Although the evidence is limited, these patients are often treated with warfarin to prevent progression, especially if remedial surgery is not possible. The international normalised ratio (INR) should be kept in the range of 2-3.

Acute thromboembolic occlusion of the peripheral arteries requires immediate anticoagulation with intravenous unfractionated heparin to prevent propagation of the thrombus and to guard against further embolism. Surgical intervention or, less commonly, thrombolytic therapy is indicated. Once the embolus has been cleared, the source needs to be investigated and this usually requires treatment with warfarin long term.



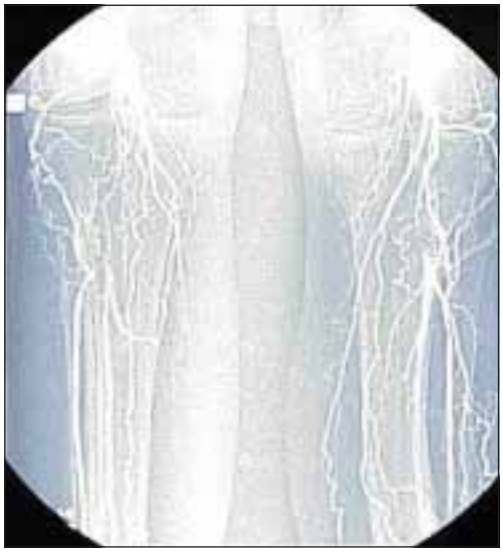
Gangrenous toe indicating critical ischaemia

Peripheral artery revascularisation

When the ischaemia reaches a state where peripheral artery revascularisation or reconstruction is necessary, the requirements for antithrombotic therapy change.

Neointimal hyperplasia is a considerable problem in the long term survival of a graft as its consequences (reduced blood flow caused by reduced lumen) in some respects mimic those of the original disease. Hyperplasia of smooth muscle cells can occur along the entire length of a vein graft, but particularly does so at the anastomoses of prosthetic grafts.

Aspirin has no apparent effect on graft survival in humans. One trial showed that low molecular weight heparin had a profound beneficial effect on graft patency, when compared with aspirin and dipyridamole over three months, suggesting that early treatment with low molecular weight heparin suppresses neointimal hyperplasia. In the United Kingdom most low molecular weight heparins have licences only for 14 days'



Peripheral angiogram showing chronic occlusions with multiple collateral vessels

Risk of thrombosis with different vein grafts

Site of proximal anastomosis	Site of distal anastomosis	Graft material	Other factors	Thrombotic risk	Recommended antithrombotic therapy
Aorta	Iliac or femoral	Prosthetic		Low	Antiplatelet*
Axilla	Femoral	Prosthetic		Medium	Antiplatelet
Femoral	Popliteal (above knee joint)	Vein		Low	Antiplatelet*
		Prosthetic		Low	Antiplatelet*
	Distal (below knee)	Vein	Good flow (> 100 ml/min) and good distal arteries	Medium	Antiplatelet
			Poor flow (< 50 ml/min) or poor distal arteries	High	Antiplatelet Consider warfarin
		Prosthetic		High	Antiplatelet Consider warfarin

*Antiplatelet therapy not indicated for graft survival but recommended as prophylaxis against cardiovascular events

treatment, and, until more data are available, the prolonged use of low molecular weight heparin cannot be recommended.

Infrainguinal bypass

Antiplatelet treatment has no beneficial role for graft patency in short (femoral-popliteal) bypass with native vein grafts because these are high flow and non-thrombogenic. None the less, aspirin has been shown to reduce all cardiovascular end points in patients with peripheral vascular disease, and so should be continued. Anticoagulation with warfarin has not been shown to be of benefit.

Patients with prosthetic femoral-popliteal bypass are a different consideration. Taking aspirin with dipyridamole reduces platelet accumulation at the anastomosis. Starting antiplatelet treatment preoperatively leads to improved patency rates, especially in "high risk" (low flow, prosthetic) grafts once the increased complication rate of postoperative wound haematoma has passed. Again, aspirin (with or without dipyridamole) is recommended.

High risk grafts need to be dealt with cautiously. All patients should continue taking aspirin (or clopidogrel). The use of warfarin needs to be judged carefully. In cases of poor run off, marginal quality vein, and previous graft failure, oral anticoagulation has been shown to improve primary patency and limb salvage rates with a target INR of 2-3. If this is being considered then full heparinisation should begin immediately after the operation while oral anticoagulation is started. Naturally, older patients are more likely to have bleeding complications, including intracranial haemorrhage, and this should be considered.

Aortoiliac and aortofemoral grafts

Large aortoiliac and aortofemoral grafts are at low risk of thrombosis. Primary patency rates of 80-90% can be expected at five to ten years. Thus, specific antithrombotic therapy is not indicated. However, once again, the presence of peripheral vascular disease needs antiplatelet therapy to reduce all cardiovascular end points.

Percutaneous transluminal angioplasty

Almost all patients undergoing percutaneous transluminal angioplasty have atherosclerotic peripheral vascular disease. As such, they should all be treated with aspirin or clopidogrel.

Studies with radiolabelled platelets have found substantial platelet accumulation at the sites of angioplasty, and antiplatelet treatment reduces this. In coronary angiography, this treatment has been shown to reduce the incidence of new thrombus at the site of the angioplasty. However, in similar coronary artery studies, antiplatelet treatment has no effect on restenosis compared with placebo. It is unclear how these results will extrapolate to peripheral angioplasty, and there are insufficient data to make recommendations in peripheral vascular disease.

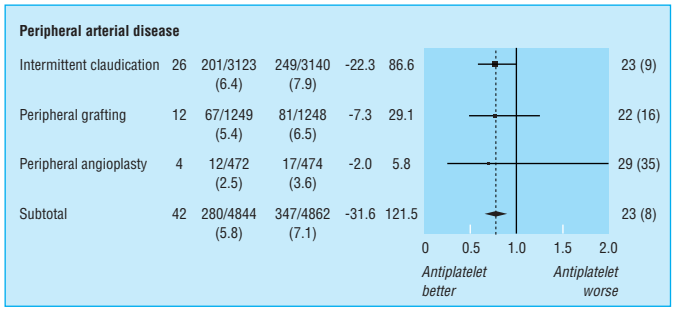
Similarly there have not been enough studies to recommend the use of dipyridamole, ticlopidine, or clopidogrel as an adjunct to aspirin. Although the long term use of antiplatelet drugs is not known to have any long term effect on restenosis, the drug should be used to prevent cardiovascular mortality in patients undergoing percutaneous transluminal angioplasty.

Carotid stenosis

Evidence for treatment of asymptomatic carotid stenosis of greater than 50% is unclear. One trial showed no reduction in stroke rate in patients treated with aspirin for two to three years. However, it is increasingly accepted that atherosclerosis affects all arteries to a greater or lesser extent. With this in mind, and

Problems in patients undergoing infrainguinal bypass

- The thrombogenic characteristics of prosthetic graft materials
- The poor flow states associated with some grafts, for example, long bypasses passing over the knee joint
- The medium to long term complication of neointimal hyperplasia



Meta-analysis from the Antithrombotic Trialists' Collaboration showing the benefits of antiplatelet treatment in patients with peripheral vascular disease

Among high risk patients, antiplatelet treatment reduces the combined outcome of any serious vascular event by about a quarter, non-fatal myocardial infarction by a third, non-fatal stroke by a quarter, and vascular mortality by a sixth (with no apparent adverse effect on other deaths)

Suggestions from the Antithrombotic Trialists' Collaboration

- Clopidogrel reduced serious vascular events by 10% compared with aspirin, which was similar to the 12% reduction observed with its analogue ticlopidine
- Addition of dipyridamole to aspirin produced no significant further reduction in vascular events compared with aspirin alone

the evidence for using aspirin in lower limb atherosclerosis, it is still recommended that these patients have antiplatelet treatment to prevent all cardiovascular events.

Treating patients who have had a transient ischaemic attack or known ischaemic stroke with aspirin has clear benefit as shown by the Antiplatelet Trialists' Collaboration. The dose is not clear, but 81-325 mg should be effective without unacceptable bleeding risk. Clopidogrel (75 mg daily) is recommended for aspirin intolerant patients. Limited evidence shows that the combination of aspirin and dipyridamole (400 mg daily) may be more beneficial to these patients than aspirin alone.

Inadequate data exist on the use of warfarin in symptomatic carotid stenosis, and so this cannot be recommended because of possible bleeding complications.

Carotid endarterectomy is the treatment of choice for all symptomatic carotid stenosis. Aspirin treatment should be continued in the perioperative period to prevent platelet deposition at the site of the endarterectomy and thus reduce intraoperative and postoperative stroke. Platelet deposition is known to start immediately after the operation, and aspirin started in the first few postoperative days seems to provide much less benefit. In patients with symptomatic disease who are not undergoing endarterectomy antiplatelet therapy is essential to reduce the incidence of ischaemic stroke. Again, warfarin should not be used as not enough evidence exists. In all patients with cerebrovascular or carotid disease, antiplatelet therapy is recommended at all stages to decrease the risk of cardiovascular events.

Further reading

- Jackson MR, Clagett GP. Antithrombotic therapy in peripheral arterial occlusive disease. *Chest* 2001;119:293-9S
- CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *Lancet* 1996;348:1329-39
- Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy I: prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *BMJ* 1994;308:81-106
- Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy II: maintenance of vascular graft or arterial patency by antiplatelet therapy. *BMJ* 1994;308:159-68
- Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction and stroke in high risk patients. *BMJ* 2002;324:71-86

The survival curve from the CAPRIE study is adapted from CAPRIE Steering Committee, *Lancet* 1996;348:1329-39. The table showing the graft risk of thrombosis and the table of antithrombotic therapy in peripheral vascular disease are adapted from Jackson MR, Clagett GP, *Chest* 2001;119: 293-9S. The meta-analysis showing the benefits of antiplatelet treatment in patients with peripheral vascular disease is adapted from the Antithrombotic Trialist's Collaboration, *BMJ* 2002;324:71-86.

8 Antithrombotic therapy for cerebrovascular disorders

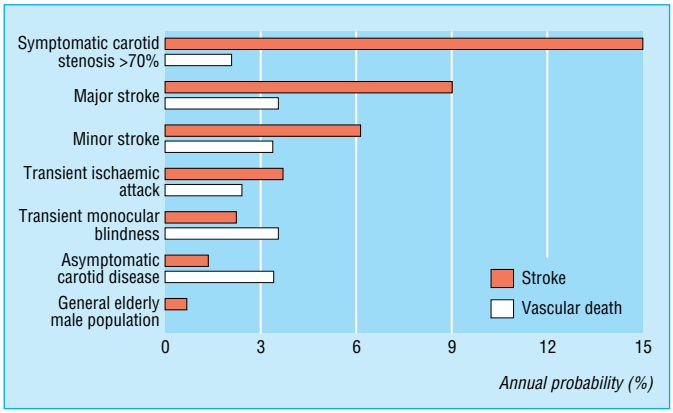
Gregory Y H Lip, Sridhar Kamath, Robert G Hart

Stroke remains one of the leading causes of death and disability throughout the world. It is the third commonest cause of death in developed countries, exceeded only by coronary artery disease and cancer.

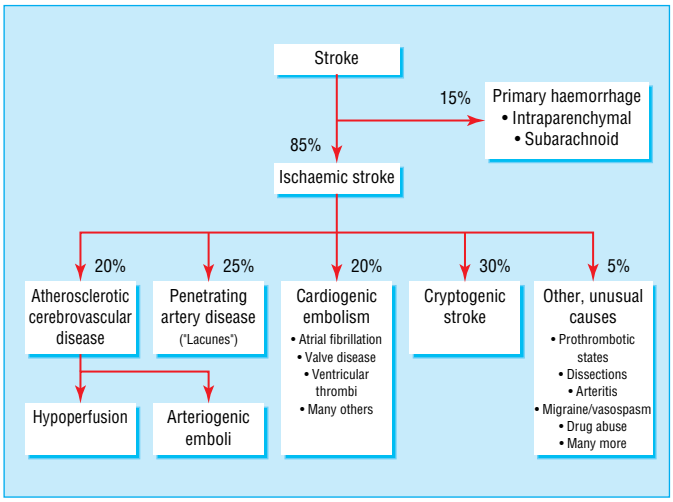
The incidence of stroke is 1-2 cases in 1000 people a year in the Western world, and is probably slightly higher among African-Caribbeans than other ethnic groups. Cerebrovascular disorders are uncommon in people aged <40 years, but there is a definite increase with age, with an incidence of 10 cases in 1000 people aged >75 in a year. Stroke is slightly more common in men, but women tend to have a poorer prognosis because of a higher mean age at onset. The incidence of stroke has been declining in recent decades in many Western countries because of better population control of hypertension, smoking, and other risk factors. However, the absolute number of strokes continues to increase because of the ageing population, which is predicted to peak in 2015. Thus, the present annual incidence of 700 000 strokes in the United States is expected to rise to 1 100 000 in 2015, without further advances in prevention.

About 80-85% of the strokes are ischaemic, with the rest primarily haemorrhagic. Even among patients with ischaemic stroke, there is much heterogeneity in aetiological and pathophysiological factors contributing to the disease.

Atherosclerosis of the major cerebral vessels probably accounts for most ischaemic strokes, either as thrombotic occlusion at the site of atherosclerotic plaques or atherogenic embolism. Embolism from a source in the heart (cardioembolic stroke) and lipohyalinosis of the penetrating small cerebral vessels (lacunar stroke) account for a substantial proportion of ischaemic strokes. In many patients the aetiology remains unknown. The major risk factors for ischaemic stroke include old age, male sex, obesity, hypertension, diabetes, and tobacco smoking.



Annual risk of stroke or vascular death among patients in various high risk subgroups



Classification of stroke by mechanism, with estimates of the frequency of various categories of abnormalities

Management of acute ischaemic stroke

The principles of management of patients with ischaemic stroke include slowing the progression of stroke, decreasing the recurrence of stroke, decreasing death and disability, preventing deep vein thrombosis and pulmonary embolism, and suppressing fever, managing hypertension and controlling glucose levels.

Antiplatelet treatment

Aspirin is the only antiplatelet drug evaluated for the treatment of acute ischaemic stroke and is recommended early in the management at a dose of 160-325 mg daily. Two major randomised trials (the international stroke trial (IST) and the Chinese acute stroke trial (CAST)) have shown that starting daily aspirin promptly (<48 hours after the start rather than the end of the hospital stay) in patients with suspected acute ischaemic stroke reduces the immediate risk of further stroke or death in hospital, and the overall risk of death and dependency at six months later. About 10 deaths or recurrent strokes are avoided in every 1000 patients treated with aspirin in the first few weeks after an ischaemic stroke.

The benefit of aspirin is seen in a wide range of patients irrespective of age, sex, atrial fibrillation, blood pressure, stroke subtype, and computed tomographic findings. In IST 300 mg of

Pathophysiological classification of stroke

Thrombosis
- Atherosclerosis
- Vasculitis
- Thrombophilic disorders
- Drug abuse such as cocaine, amphetamines

Embolism
- From the heart
- From the major cerebrovascular vessels
- Unknown source

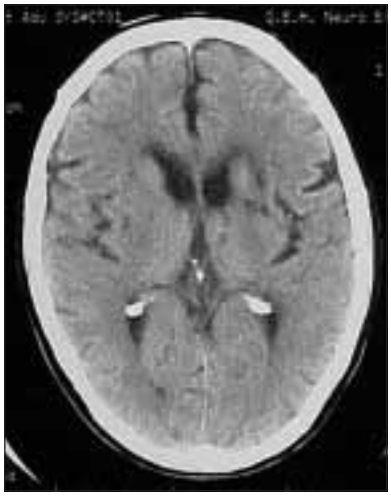
Lipohyalinolysis
- Small penetrating arteries

Vasospasm
- Migraine
- Subarachnoid haemorrhage

Dissection
- Spontaneous
- Traumatic

aspirin was used and in CAST 160 mg. Thus, the two studies show that giving aspirin early in acute stroke is safe, although side effects should always be considered. Other trials have shown that continuing treatment with low dose aspirin gives protection in the longer term. Until further evidence is available, however, aspirin should be withheld from patients receiving other forms of anticoagulant (except low dose heparin (5000 IU twice daily)) or thrombolytic treatment (and for 24 hours after finishing treatment).

The results of the IST and CAST studies apply chiefly to patients who had a computed tomography scan to exclude intracranial haemorrhage. A meta-analysis of subgroups from the trials showed that aspirin was safe and beneficial. Even among patients who did not have a computed tomogram and patients with haemorrhagic stroke, aspirin treatment did not result in net hazard. Thus, aspirin can be started in patients with suspected ischaemic stroke even when computed tomography is not available immediately.

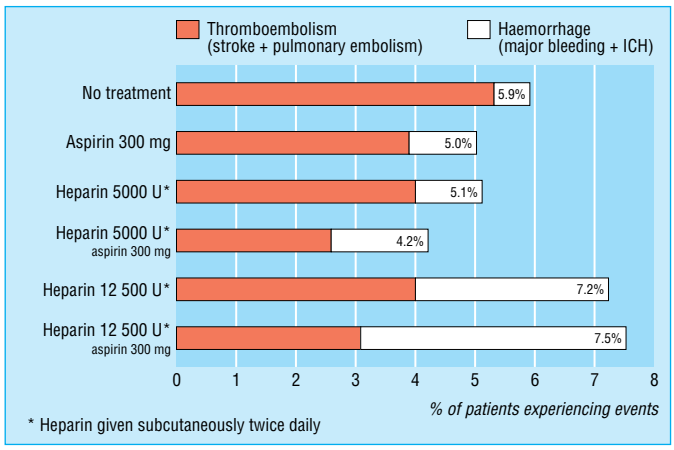


Computed tomogram of the brain showing lacunar infarcts in the anterior limb of the left internal capsule

Anticoagulation treatment

Heparin is not routinely recommended for patients with acute ischaemic stroke. There are no randomised trials supporting the use of standard doses of heparin (for example >10 000 IU daily) even in patients with acute stroke and risk factors for recurrent events. The risk:benefit ratio of heparin administration is narrow, ill defined, and probably depends on the pathophysiological subtype of stroke and the factors that predispose to haemorrhage. For patients with atrial fibrillation and acute ischaemic stroke, there seems to be no net benefit from standard dose heparin (aspirin should be given immediately, then warfarin (INR 2.0-3.0) started for secondary prevention as soon as the patient is medically stable). However, a subgroup analysis from IST showed that in acute ischaemic stroke low dose heparin (5000 IU twice daily) reduced death and recurrence, especially if combined with aspirin, and it is indicated if appreciable leg weakness is present for prevention of venous thromboembolism.

No particular benefit was observed in ischaemic stroke in the vertebrobasilar region with anticoagulation at six months. Trials with low molecular weight heparins or heparinoids have yielded contradictory (but generally negative) results, and they are not recommended for use at the moment.



Thromboembolic and major haemorrhagic events in the International Stroke Trial. ICH=intracranial haemorrhage

Thrombolytic treatment

Thrombolytic treatment for acute ischaemic stroke has been in vogue since its immense benefit was seen with myocardial infarction. The National Institute of Neurological Disorders and Stroke (NINDS) rtPA Study Group trial showed that recombinant tissue plasminogen activator administered within three hours of onset of acute cerebral infarct at a dose of 0.9 mg/kg (maximum 90 mg) given over an hour under strict treatment protocols increased the likelihood of minimal or no disability at three months by at least 30%. This benefit was seen in all stroke patients. Recombinant tissue plasminogen activator is licensed for treating acute cerebral infarct in several countries. However, the risk:benefit ratio is narrow because of substantial risk of intracerebral haemorrhage, and the need to start treatment (after computed tomographic assessment) within three hours of stroke onset severely restricts the number of patients who can be treated.

Streptokinase is not approved for use in acute cerebral infarct because of the results of three large trials, which were terminated early due to excessive bleeding. These trials used streptokinase at a dose of 1.5 million units given more than three hours after stroke onset. Intra-arterial thrombolytic treatment for patients with large artery occlusions (such as of

Cardiac disorders predisposing to stroke

Major risk
- Atrial fibrillation
- Prosthetic mechanical heart valve
- Mitral stenosis
- Severe left ventricular dysfunction with mobile left ventricular thrombus
- Recent myocardial infarction
- Infective endocarditis

Minor risk*
- Mitral annular calcification
- Mitral valve prolapse
- Patent foramen ovale
- Calcific aortic stenosis
- Atrial septal aneurysm

*Occasionally can cause cardioembolic stroke, but the risk of initial stroke is low and often unrelated when identified during the evaluation of patients with cerebral ischaemia

the internal carotid artery, middle cerebral artery, or basilarartery) remains investigational. In the United Kingdom, thrombolytic treatment is not licensed for treatment of stroke, pending results from ongoing clinical trials (for example, IST-3).

Stroke prevention

In broad terms, antiplatelet agents are more effective in cerebrovascular atherogenic strokes, and anticoagulants are more effective in primary and secondary prophylaxis against cardioembolic stroke.

Antiplatelet agents

Among patients with vascular disorders (such as coronary artery disease, previous stroke or transient ischaemic attack, and peripheral vascular disease) antiplatelet agents substantially reduce the incidence of non-fatal stroke, non-fatal myocardial infarction, vascular mortality, and composite end point of stroke, myocardial infarction, and vascular death.

A variety of antiplatelet drugs with varying mechanisms of action are used to minimise stroke in patients at high risk. These include aspirin (irreversible inhibitor of cyclo-oxygenase), clopidogrel (which inhibits adenosine diphosphate induced platelet aggregation) and dipyridamole (precise mechanism of action not yet clear). Aspirin remains the most commonly used antiplatelet drug, partly because of its cost effectiveness.

Aspirin is effective for stroke prevention in doses ranging from 30 mg/day to 1300 mg/day. Its beneficial effect is seen in all age groups and sexes. The European stroke prevention study II (ESPS II) showed that a combination of aspirin and dipyridamole (sustained release 200 mg tablets twice daily) significantly reduced the risk of stroke and all vascular events compared with aspirin alone. An important ongoing trial (ESPRIT) is seeking to replicate these results.

Clopidogrel is a newer thienopyridine derivative without the adverse effect profile of ticlopidine. The CAPRIE (clopidogrel versus asprin in patients at risk of ischaemic events) study showed that clopidogrel is slightly more effective than aspirin in reducing the combined outcome of stroke, myocardial infarction, and vascular death among patients with atherosclerotic vascular disease. Although clopidogrel seems to be as safe as aspirin, it is considerably more expensive, and it remains to be seen whether its use in routine practice is cost effective. Its use is justified in patients who are intolerant to aspirin or who develop a stroke while taking aspirin.

Anticoagulation treatment

Anticoagulation in the form of warfarin has a role in a variety of cardiac disorders in primary and secondary prevention of stroke. Cardiac disorders that predispose to stroke and unequivocally seem to benefit from anticoagulation therapy include atrial fibrillation (with additional risk factors putting patients at moderate to high risk), mitral stenosis (with or without atrial fibrillation), and mechanical valve prosthesis. In contrast, recent randomised trials (SPIRIT, WARSS) did not show advantages of warfarin over aspirin for secondary prevention of non-cardioembolic brain ischaemia. At present, warfarin should not be used routinely for patients with common causes of non-cardioembolic stroke, pending results from ongoing randomised trials.

Risk factors for haemorrhagic transformation of ischaemic stroke

- Hypertension
- Concomitant use of two or more antiplatelet or antithrombotic therapies
- Major early infarct signs on pretreatment computed tomography, including brain oedema and mass effect
- Severe neurological deficit

Details of the second European stroke prevention study (ESPS-II)

- Randomised, placebo controlled, double blind trial of aspirin (50 mg), dipyridamole (400 mg), or both versus placebo
- Two year follow up of >6600 patients (secondary prevention of stroke)

Placebo compared with:	Aspirin alone	Dipyridamole alone	Aspirin + dipyridamole
Reduction of stroke risk	18% (P=0.013)	16% (P=0.015)	37% (P<0.001)
Reduction of risk of stroke or death	13% (P=0.016)	15% (P=0.015)	24% (P<0.001)

- Clear additive benefit in stroke reduction (36%) when aspirin and dipyridamole were used in combination

Further reading

- Adams HP. Emergency use of anticoagulation for treatment of patients with ischemic stroke. *Stroke* 2002;33:856-61
- Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high-risk patients. *BMJ* 2002;324:71-86
- Albers GW, Amarenco P, Easton JD, Sacco RL, Teal P. Antithrombotic and thrombolytic therapy for ischemic stroke. *Chest* 2001;119:300–20S
- Atkinson RP, DeLemos C. Acute ischemic stroke management. *Thromb Res* 2000;98:V97-111
- CAPRIE Steering Committee. A randomised, blinded trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *Lancet* 1996;348:1329-39
- CAST Collaborative Group. CAST: randomised placebo-controlled trial of early aspirin use in 20 000 patients with acute ischaemic stroke. *Lancet* 1997;349:1641-9
- International Stroke Trials Collaborative Group. IST: a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *Lancet* 1997;349: 1569-81
- Mohr JP, Thompson JLP, Lazar RM, Levin B, Sacco RL, Furie KL, et al for the Warfarin-Aspirin Recurrent Stroke Study Group. A comparison of warfarin and aspirin for the prevention of recurrent ischemic stroke. *N Engl J Med* 2002;345:1444-51

The flow diagram showing a classification of stroke by mechanism with estimates of the frequency of various categories of abnormalities is adapted from Albers GW et al, *Chest* 2001;119:300-20. Annual risk of stroke or vascular death among patients in various high risk subgroups is adapted from Wilterdink and Easton, *Arch Neurol* 1992;49:857-63. The figure showing thromboembolic and major haemorrhagic events in the IST is adapted from IST Collaborative Group. *Lancet* 1997;349:1569-81.

9 Valvar heart disease and prosthetic heart valves

Ira Goldsmith, Alexander G G Turpie, Gregory Y H Lip

Thromboembolism and anticoagulant related bleeding are major life threatening complications in patients with valvar heart disease and those with prosthetic heart valves. In these patients effective and safe antithrombotic therapy is indicated to reduce the risks of thromboembolism while keeping bleeding complications to a minimum.

Assessment

Risk factors that increase the incidence of systemic embolism must be considered when defining the need for starting antithrombotic therapy in patients with cardiac valvar disease and prosthetic heart valves. These factors include age, smoking, hypertension, diabetes, hyperlipidaemia, type and severity of valve lesion, presence of atrial fibrillation, heart failure or low cardiac output, size of the left atrium (over 50 mm on echocardiography), previous thromboembolism, and abnormalities of the coagulation system including hepatic failure.

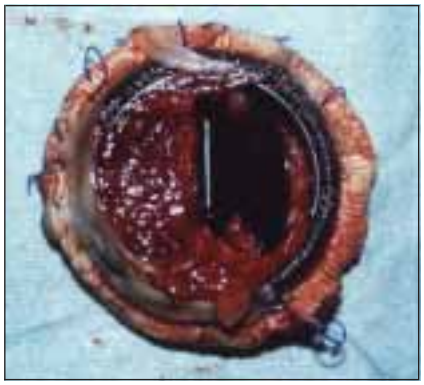
Secondly, the type, number, and location of prostheses implanted must be considered. For example, mechanical prostheses are more thrombogenic than bioprostheses or homografts, and hence patients with mechanical valves require lifelong anticoagulant therapy. However, the intensity of treatment varies according to the type of mechanical prosthesis implanted. First generation mechanical valves, namely the Starr-Edwards caged ball valve and Bjork-Shiley standard valves, have a high thromboembolic risk; single tilting disc valves have an intermediate thromboembolic risk; and the newer (second and third generation) bileaflet valves have low thromboembolic risks.

In patients with a bioprosthesis in sinus rhythm, antithrombotic therapy with an antiplatelet drug may suffice, whereas patients with homografts in sinus rhythm may not need any antithrombotic therapy. Thromboembolic events are commoner with prosthetic mitral valves than aortic valves and in patients with double replacement valves compared with those with single replaced valves. Moreover, the risk of thromboembolic events is greatest in the first three months after implantation.

Choice of antithrombotic agent

Warfarin is the most used oral anticoagulant, and its dose is guided by achieving a target international normalised ratio (INR) range. The use of heparin is confined to short periods when anticoagulant cover is needed and oral anticoagulants are stopped. The dose of heparin is adjusted to achieve at least twice normal level of activated partial thromboplastin time (APTT) regardless of cardiac rhythm and type or position of prosthesis. Fixed weight-adjusted low molecular weight heparin may be used as an alternative to unfractionated heparin. Antiplatelet drugs, such as low dose aspirin or dipyridamole, are used in patients with bioprosthesis in sinus rhythm and in addition to anticoagulants in the high risk patients with mechanical valves.

Patients with mechanical valves and those with bioprostheses and associated risk factors require lifelong anticoagulant cover. In patients with a bioprosthetic valve in sinus rhythm anticoagulant cover with warfarin for the first three postoperative months may suffice, followed by low dose aspirin treatment for life. Alternatively, some surgeons give only low dose aspirin after



Valve thrombosis of a bileaflet prosthetic mitral valve

Considerations for antithrombotic therapy in patients with valve disease

- Assessment of risk for thromboembolic events, which may be patient related or valve prosthesis related
- Indications for starting treatment
- Choice of antithrombotic agent
- Duration of treatment and optimal therapeutic range
- Antithrombotic therapy in special circumstances (surgical procedures, pregnancy, and resistance to oral anticoagulants)
- Management of treatment failures and complications

Types of prosthetic valves and thrombogenicity

Type of valve	Model	Thrombogenicity
Mechanical		
Caged ball	Starr-Edwards	+ + + +
Single tilting disc	Bjork-Shiley, Medtronic Hall	+ + +
Bileaflet	St Jude Medical, Sorin Bicarbon, Carbomedics	+ +
Bioprosthetic		
Heterografts	Carpentier-Edwards, Tissue Med (Aspire), Hancock II	+ to + +
Homografts		+

Risk factors for patients with bioprostheses include previous thromboembolic events, atrial fibrillation, enlarged left atrial cavity, and severe cardiac failure

surgery in patients with bioprostheses in sinus rhythm (providing aspirin is not contraindicated). Patients with homografts usually do not require any antithrombotic therapy.

Indications for antithrombotic therapy

Native valve disease

Oral anticoagulant treatment is indicated in all patients who have established or paroxysmal atrial fibrillation with native valve disease regardless of the nature or severity of the valve disease. In patients with mitral stenosis in sinus rhythm, treatment is guided by the severity of stenosis, the patient's age, size of the left atrium, and the presence of spontaneous echocontrast or echocardiographic evidence of left atrial appendage thrombus. In these patients a target INR of 2.5 (range 2-3) is recommended. Similarly, in patients with mitral regurgitation treatment is indicated in the presence of congestive cardiac failure, marked cardiomegaly with low cardiac output, and an enlarged left atrium. In the absence of cardiac failure, previous thromboemboli, or heart failure, antithrombotic therapy is not indicated in patients with isolated aortic or tricuspid valve disease.

Mitral valve prolapse per se does not require anticoagulant cover, although sometimes aspirin is recommended because of the association with cerebrovascular events.

Percutaneous balloon valvuloplasty

In patients with mitral stenosis, the presence or absence of left atrial thrombus is first confirmed by transoesophageal echocardiography. In the presence of thrombus, valvuloplasty is deferred and anticoagulant treatment started for two months before the procedure, with a target INR range of 2-3. In the absence of atrial thrombus but in the presence of risk factors—namely, previous thromboembolism, enlarged left atrium, spontaneous echocontrast, or atrial fibrillation—oral anticoagulant treatment should be started a month before the procedure.

During the procedure, intravenous heparin (2000-5000 IU bolus) should be given to all patients immediately after trans-septal catheterisation. After the procedure, subcutaneous heparin should be given for 24 hours and oral anticoagulant treatment restarted 24 hours after the procedure in patients with risk factors, especially in the presence of atrial fibrillation or spontaneous echocontrast.

Patients in sinus rhythm who are undergoing aortic valvuloplasty do not need long term anticoagulant treatment. However, treatment with heparin during the procedure is required.

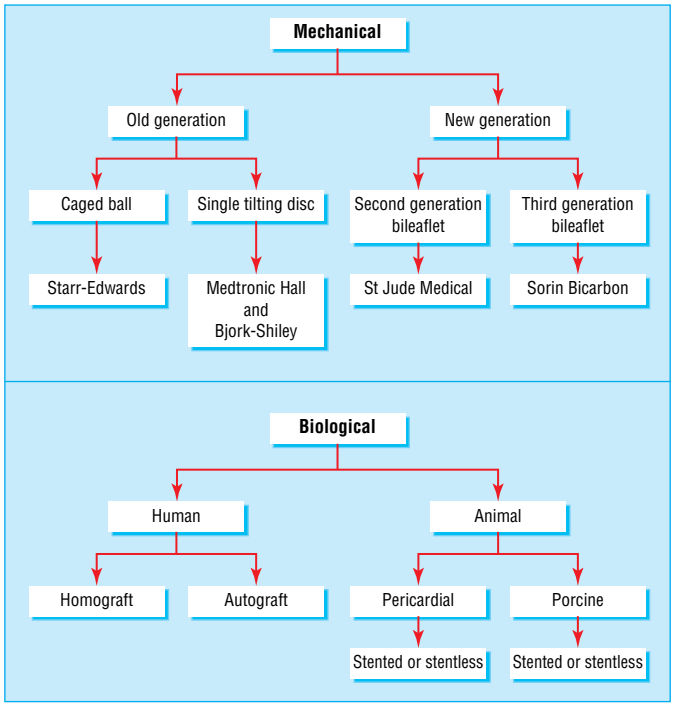
Mitral valve repair

After mitral valve repair, oral anticoagulation (target INR 2.5) is needed for the first six weeks to three months, and thereafter treatment is guided by the presence or absence of risk factors such as atrial fibrillation, heart failure, and enlarged left atrium.

Heart valve replacement

Antithrombotic therapy in patients with replaced heart valves is guided by the type of prosthesis implanted (mechanical or biological), position of the implant, associated risk factors (such as atrial fibrillation), previous thromboembolism, bleeding risk, and the patient's age.

Patients with porcine or pericardial bioprostheses in sinus rhythm may be started on lifelong antiplatelet treatment with low dose aspirin as soon as they can swallow the drugs.



Types of heart valve prostheses

Comparison of mechanical and biological valve prostheses

Mechanical	Biological
Durable—valves lasting 20-30 years	Limited life span—10% of homografts and 30% of heterografts fail within 10-15 years
Thrombogenic—patients require lifelong anticoagulant therapy	Low thrombogenic potential—lifelong anticoagulation is not required
Preferred in younger patients with >10-15 years life expectancy	Preferred in older patients with <10-15 years life expectancy
Preferred in patients who require lifelong anticoagulant therapy	Preferred in those who cannot (or will not) take lifelong anticoagulant therapy

Examples of biological and mechanical valve protheses: (left) stentless porcine valve, (middle) stented porcine valve, (right) Sorin Bicarbon valve

However, many centres start oral anticoagulant treatment the day after implantation, maintaining an INR range of 2-3 for the first three months. Lifelong anticoagulant treatment is recommended for patients with associated risk factors. These factors are previous thromboembolism, left atrial thrombus, marked cardiomegaly, heart failure, dilated left atrium, or spontaneous echocontrast.

Patients with mechanical heart valves require lifelong anticoagulant treatment, and patients with first generation valves (with the highest thromboembolic risk) need a higher target INR than patients with single tilting disc prostheses (intermediate thromboembolic risk) or the newer bileaflet prosthesis (lower thromboembolic risk).

Most centres start (or restart) oral anticoagulant treatment the day after implantation, with or without heparinisation. As the thromboembolic risk is highest in the early postoperative period, it is advisable to give heparin and to continue it until the oral anticoagulant treatment achieves the target INR. The dose of heparin should be adjusted to achieve twice the normal level of APTT regardless of cardiac rhythm and type or position of the valve.

The European and North American guidelines have minor differences. The duration of antithrombotic therapy also varies according to a number of factors. Lifelong anticoagulant treatment is indicated for patients with mechanical valves and those with bioprosthetic valves or native valve disease with additional risk factors.

Intensity of anticoagulation guidelines for Europe

	European Society of Cardiology 1995 INR range	British Society of Haematology 1998 INR target
*Mechanical valves**		
Aortic:		
First generation	3.0-4.5	3.5
Second generation	2.5-3.0	3.5†
Third generation	2.5-3.0	3.5†
Mitral	3.0-3.5	3.5
Bioprosthetic valves		
In sinus rhythm:		
Aortic	2.5-3.0 for three months	No anticoagulation‡
Mitral	3.0-3.5 for three months. No anticoagulation after three months	2.5 for three months. No anticoagulation after three months
In atrial fibrillation:		
Rheumatic valvar heart disease	3.0-4.5	2.5
Patients with recurrent emboli under adequate anticoagulation	3.0-4.5 + 100 mg aspirin	—
Non-valvar atrial fibrillation with risk factors	2.0-3.0	2.5

*First generation valves include Starr-Edwards and Bjork-Shiley; second generation valves include St Jude Medical and Medtronic Hall; and third generation valves include the Sorin Bicarbon bileaflet valve
†For second and third generation mechanical aortic valves a target INR of 2.5 is used
‡Low dose aspirin is used by most centres in the United Kingdom

Antithrombotic therapy in special circumstances

Modification of anticoagulant treatment may be required in patients who have prosthetic valves and are undergoing non-cardiac surgical procedures, who are are pregnant, or who have resistance to oral anticoagulants.

Surgical procedures

For minor procedures, such as certain dental surgery or cryotherapy, where blood loss is expected to be minimal and easily manageable, anticoagulant treatment may be continued. After dental extraction bleeding can be stopped with oral tranexamic acid (4.8%) mouth wash. However, before a planned minor surgical procedure, the INR should be adjusted to between 1.5 and 2.0. This can be achieved by stopping or adjusting oral anticoagulant treatment one to three days before the procedure depending on the drug used. In most cases, resumption of oral anticoagulant treatment is possible on the same day as the procedure, and interim heparin treatment is not needed. Patients undergoing endoscopic procedures and in whom an endoscopic biopsy is anticipated should be managed in the same way as patients needing major non-cardiac surgical procedures.

For major non-cardiac surgical procedures, in which there is a substantial risk of bleeding, anticoagulation should be discontinued for several days (generally four to five days) before surgery and the INR should be normalised at 1.0. The risk of thromboembolism increases, and so interim heparin treatment should be given in a dose that prolongs the APTT to twice the control value. However, heparin should be stopped in time to bring the APTT down to near normal at the time of operation and resumed as soon as possible postoperatively. An alternative approach would be to use therapeutic fixed weight-adjusted doses of low molecular weight heparin.

Intensity of anticoagulation guidelines for North America

	AHA and ACC 1998 INR range	ACCP 2001 INR (target range)
Mechanical valves		
First, second, and third generation valves:		
Aortic	2.0-3.0	2.5 (2.0-3.0)
Mitral	2.5-3.5	3.0 (2.5-3.5)
Bioprosthetic valves		
In sinus rhythm:		
Aortic	Aspirin 80-100 mg/day	2.5 (2.0-3.0) for three months
Mitral	Aspirin 80-100 mg/day. No anticoagulation after three months	2.5 (2.0-3.0) for three months
In atrial fibrillation:		
Aortic	2.0-3.0	2.5 (2.0-3.0)
Mitral	2.5-3.5	2.5 (2.0-3.0)

AHA = American Heart Association; ACC = American College of Cardiology; ACCP = American College of Chest Physicians

Pregnancy

In pregnant women with prosthetic valves, the incidence of thromboembolic complications is increased. Hence, adequate antithrombotic therapy is particularly important. Warfarin use in the first trimester of pregnancy is associated with a substantical risk of embryopathy and fetal death, and so warfarin should be stopped when a patient is trying to become pregnant or when pregnancy is detected. Instead, twice daily subcutaneous unfractionated heparin should be given to prolong the APTT to twice the control value, and this treatment may be continued until delivery. Alternatively, unfractionated heparin may be given until the thirteenth week of pregnancy, then a switch made to warfarin treatment until the middle of the third trimester. Then warfarin can be stopped and unfractionated heparin resumed until delivery. Because low dose aspirin is safe for mother and child, it may be used in conjunction with anticoagulant treatment in women at high risk of thromboembolism. However, low molecular weight heparin (which does not cross the placental barrier) may be an alternative to unfractionated heparin in this setting, although there are limited data on its efficacy or safety in pregnancy. Other pregnancy related issues are discussed in chapter 14.

Management of temporary interruption of oral anticoagulants

- Discontinue oral anticoagulation five days before the procedure
- Measure INR three days before procedure
  If INR <2 start low molecular weight heparin in therapeutic doses
  If INR >2.5 consider giving Vitamin K_1 1-2 mg orally and start low molecular weight heparin in therapeutic doses. Repeat INR measurement the day before procedure
- Continue low molecular weight heparin until evening before procedure (last injection not less than 12 hours preprocedure)
- Restart warfarin night of or day after procedure
- Restart low molecular weight heparin 12-24 hours after procedure and when haemostasis is established

Indications for lifelong oral anticoagulation in valve disease

- Mechanical prostheses
- Chronic or paroxysmal atrial fibrillation in the presence of native valve disease, bioprosthesis, valve repair, or valvuloplasty
- Native valve disease and previous thromboembolism
- Mitral valve stenosis, irrespective of rhythm, in association with high transmitral valve gradient, left atrial thrombus, spontaneous echocontrast, large left atrium (>50 mm), low cardiac output, or congestive heart failure

Further reading

- Bonow RO, Carobello D, de Leon AC, Edmunds LH Jr, Fedderly BJ, Freed MD, et al. ACC/AHA guidelines for the management of patients with valvular heart disease. *J Am Coll Cardiol* 1998; 32:1486-8
- Gohlke-Barwölf C. Anticoagulation in valvular heart disease: new aspects and management during non-cardiac surgery. *Heart* 2000;84:567-72
- Haemostasis and Thrombosis Task Force of the British Society for Haematology. Guidelines for prevention of thromboembolic events in valvular heart disease. *Eur Heart J* 1995;16:1320-30

10 Antithrombotic therapy in myocardial infarction and stable angina

Gregory Y H Lip, Bernard S P Chin, Neeraj Prasad

Acute Q wave myocardial infarction

The use of thrombolytic treatment in acute myocardial infarction is now established beyond doubt. However, primary angioplasty is now proved to be an effective alternative and is used increasingly in preference to thrombolysis in many centres worldwide.

Thrombolytic treatment

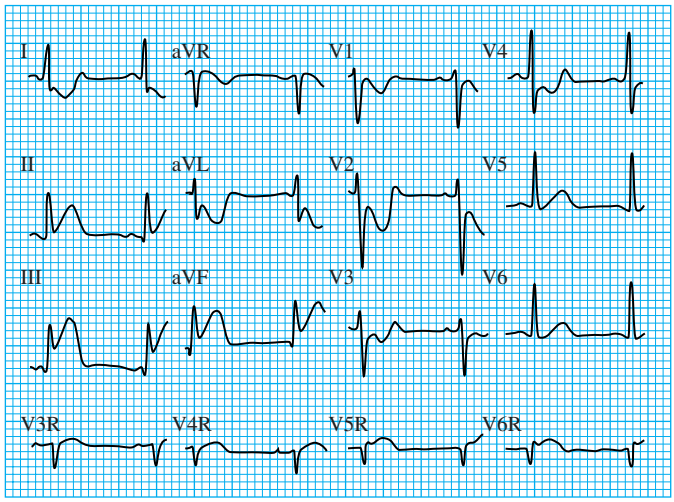
Current key issues relate to the clinical situations in which thrombolysis may be beneficial or contraindicated. For example, all patients with a history suggesting cardiac ischaemia and accompanying electrocardiographic changes indicating acute myocardial infarction should be considered for thrombolysis. However, patients with only ST segment depression on an electrocardiogram or with a normal electrocardiogram do not benefit from thrombolysis, and treatment should therefore be withheld. Exceptions to this are when there is evidence of new development of left bundle branch block or a true posterior myocardial infarction (shown by ST segment depression with dominant R waves present in leads V1 and V2). These situations require thrombolytic treatment.

Thrombolytic treatment should be given within six hours of the onset of symptoms and electrocardiographic changes for patients to derive full benefit. Patients with persisiting pain and ST segment elevation may still benefit from thrombolysis up to 12 hours from the onset of symptoms. Beyond that, few patients will benefit, and there is no clear evidence of whether this benefit outweighs the risk of haemorrhage.

Thrombolytic treatment should be offered to all eligible patients presenting with an acute myocardial infarction regardless of age, sex, or site of infarct. In general, patients over 75 years and those with anterior myocardial infarction or previous heart attack have a higher mortality. Therefore, the absolute reduction in mortality in these patients will be greater. Many of the accepted contraindications (absolute and relative) come from observational studies only. Some conditions, such as diabetic proliferative retinopathy and menstruation, are no longer considered to be absolute contraindications.

Reperfusion of the artery affected by infarction occasionally fails with thrombolytic treatment. If this happens patients will have ongoing chest pain or acute electrocardiographic changes. In these instances the optimal management is still uncertain, although readministration of an alternative thrombolytic agent ("rescue thrombolysis") or emergency percutaneous transluminal coronary angioplasty ("rescue" or "salvage" percutaneous transluminal coronary angioplasty) has been advocated. Rescue thrombolysis more than doubles the bleeding complications. Also, the limited data available showed benefit only in cases where plasma fibrinogen concentration was >1.0 g/l and where recombinant tissue plasminogen activator was given if initial streptokinase did not achieve 25% reduction of maximal ST elevation on the pretreatment electrocardiogram.

All thrombolytic agents are plasminogen activators. Streptokinase is the cheapest widely available agent. However, it is highly antigenic, and neutralising antibodies preclude use of this agent more than once in a patient. Thus, it should not be



Electrocardiogram indicating acute inferior myocardial infarction

Indications and contraindications for thrombolysis in acute myocardial infarction

Indications

- Clinical history and presentation strongly suggestive of myocardial infarction within 6 hours plus one or more of:
  1 mm ST elevation in two or more contiguous limb leads
  2 mm ST elevation in two or more contiguous chest leads
  New left bundle branch block
  2 mm ST depression in V1-4 suggestive of true posterior myocardial infarction
- Patients presenting with above within 7-12 hours of onset with persisting chest pains and ST segment elevation
- Patients aged <75 years presenting within 6 hours of anterior wall myocardial infarction should be considered for recombinant tissue plasminogen activator

Contraindications

Absolute

- Aortic dissection
- Previous cerebral haemorrhage
- Known history of cerebral aneurysm or arteriovenous malformation
- Known intracranial neoplasm
- Recent (within past 6 months) thromboembolic stroke
- Active internal bleeding (excluding menstruation)
- Patients previously treated with streptokinase or anisolated plasminogen streptokinase activator complex (APSAC or anistreplase) should receive recombinant tissue plasminogen activator, reteplase, or tenecteplase

Relative

- Severe uncontrolled hypertension (blood pressure $>180/110$ mm Hg) on presentation or chronic severe hypertension
- Current use of anticoagulants or known bleeding diathesis
- Recent (within past 2-4 weeks) trauma including head injury or traumatic or prolonged (>10 minutes) cardiopulmonary resuscitation
- Recent (within 3 weeks) major surgery, organ biopsy, or puncture of non-compressible vessel
- Recent (within past 6 months) gastrointestinal or genitourinary or other internal bleeding
- Pregnancy
- Active peptic ulcer disease

readministered after 48 hours from the initial infusion. Tenecteplase, reteplase, and recombinant tissue plasminogen activator (alteplase) have been shown to be as good as streptokinase in reducing mortality after acute myocardial infarction. They are suitable alternatives if a patient has already received streptokinase. In the GUSTO trial, recombinant tissue plasminogen activator produced greater mortality reduction than streptokinase, especially in patients aged under 75 years who presented within six hours of onset of anterior myocardial infarction. Patients presenting within six hours of inferior myocardial infarction accompanied by right ventricular infarct, haemodynamic compromise, or anterior wall extension may also benefit. Tenecteplase has the advantage of being easily administered as a single bolus injection.

Antiplatelet treatment

The concurrent use of aspirin with a thrombolytic drug reduces mortality far more than either drug alone. In the ISIS-2 trial, streptokinase reduced mortality by 25%, aspirin by 23%, and the combination of aspirin with streptokinase by 42%. In addition, there was no increase in incidence of stroke or major bleeding by giving aspirin and streptokinase in combination.

Anticoagulant treatment

The use of "full dose" heparin, either intravenously or by subcutaneous injections, is not warranted routinely after streptokinase (or anistreplase) infusion, with no difference in mortality during hospitalisation and an increased risk of cerebral haemorrhage and other major bleeding. Patients who may benefit from heparin treatment after streptokinase (or anistreplase) are those at high risk of developing thromboembolism. These patients include those with large infarctions, atrial fibrillation, or congestive cardiac failure.

On the other hand, recombinant tissue plasminogen activator, tenecteplase, and reteplase have short half lives and thus have only small systemic fibrinolytic effects. The high reocclusion rates seen in patients given recombinant tissue plasminogen activator may be stopped by concomitant use of full dose heparin (for at least 24 hours). Several trials of angiographic patency have also reported a favourable synergistic effect of heparin after recombinant tissue plasminogen activator.

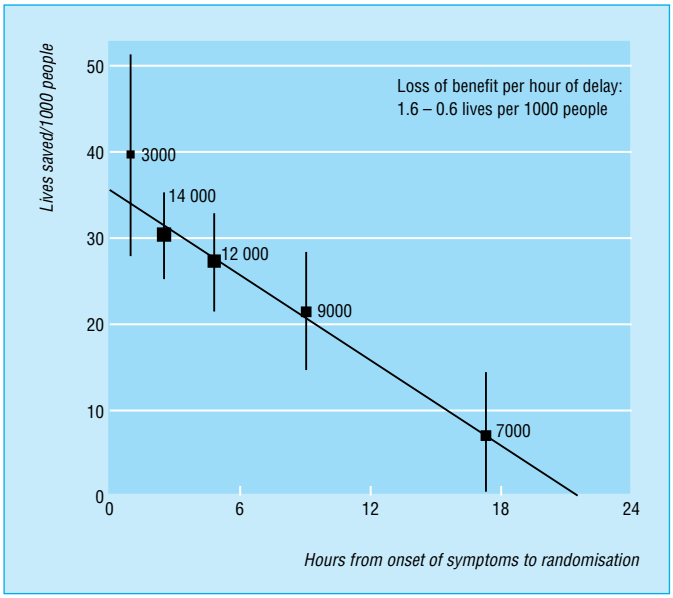
For patients with persistent risk factors for systemic embolisation, consideration should be given to starting oral warfarin or continuing heparin treatment as subcutaneous injections beyond 48 hours. All other patients should receive prophylactic heparin (unfractionated or low molecular weight heparin) until they are sufficiently mobile to minimise venous thromboembolism. Finally, trials comparing the use of hirudin and heparin after recombinant tissue plasminogen activator showed hirudin to be no better than heparin at reducing cardiovascular death or reinfarction at 30 days.

Postmyocardial infarction and stable coronary artery disease

Antiplatelet treatment

Data from the Antiplatelet Trialists' Collaboration, which analysed more than 100 trials and 100 000 patients, including 20 000 with acute myocardial infarction, confirmed that antiplatelet treatment reduced cardiovascular events in acute myocardial infarction by 25%, representing a two year treatment benefit in 36 out of 1000 patients treated.

Aspirin is the most widely used antiplatelet drug and is effective at doses from 75 to 300 mg daily in patients who have



Lives saved per thousand people in relation to time of administration of thrombolytic treatment from onset of symptoms of chest pain. Numbers along the curve are the number of people treated at different times

Antithrombotic therapy in acute Q wave myocardial infarction and after myocardial infarction

All patients should receive
- Aspirin 300 mg orally as soon as possible and 75-300 mg daily thereafter
- Consideration for thrombolysis
- β blockers, nitrates, and other standard antianginal drugs as appropriate

Choice of thrombolytic agents
- Streptokinase 1.5 MU intravenously over an hour
- Recombinant tissue plasminogen activator 15 mg intravenous bolus followed by 0.75 mg/kg (maximum 50 mg) infusion over 30 minutes, then at 0.5 mg/kg (maximum 35 mg) over 60 minutes
- Reteplase 10 MU intravenous bolus, repeated once after 30 minutes
- Tenecteplase 30-50 mg (according to body weight) intravenously over 10 seconds

Adjuvant heparin treatment
- In all patients receiving recombinant tissue plasminogen activator—75 U/kg intravenous bolus with recombinant tissue plasminogen activator infusion, followed by 1000-1200 U/hour to maintain APTT ratio 1.5-2.0 for 48 hours
- In all patients receiving reteplase or tenecteplase—75 U/kg intravenous bolus with first reteplase bolus, followed by 1000-1200 U/hour to maintain APTT ratio 1.5-2.0 for 48 hours

Prevention of systemic and venous thromboembolism
- In all patients with acute myocardial infarction—low dose low molecular weight heparin until ambulatory
- In patients receiving streptokinase at high risk of systemic or venous thromboembolism—measure APTT from four hours after thrombolysis. Start intravenous heparin at 1000-1200 U/hour once APTT ratio has fallen to less than 2.0. Continue for 48 hours, maintaining APTT at ratio 1.5-2.0. Alternatively, use low molecular weight heparin
- In all patients at high risk of systemic or venous thromboembolism—heparin infusion may be continued beyond 48 hours or converted to 15 000 U subcutaneously twice daily (alternatively, use low molecular weight heparin) or to warfarin (INR 2-3) for up to three months
- In patients with atrial fibrillation—warfarin treatment after heparin infusion should continue indefinitely

had myocardial infarction and those with stable coronary artery disease. Although there is no substantial difference in efficacy between lower and higher doses of aspirin within the stated range, higher doses are associated with greater side effects.

Dipyridamole and ticlopidine have both been compared with aspirin. Dipyridamole showed no benefit over aspirin in the PARIS trials. Ticlopidine may be slightly better than aspirin treatment but is associated with undesirable side effects such as neutropenia and thrombocytopenia. In the CAPRIE study, which compared clopidogrel with aspirin over two years in patients with vascular disease (ischaemic stroke, myocardial infarction, peripheral vascular disease), clopidogrel was slightly better in reducing the number of vascular events (5.32% *v* 5.83%, P = 0.04). Importantly, clopidogrel was as well tolerated as aspirin. Therefore, it would be reasonable to give patients clopidogrel after acute myocardial infarction if aspirin were contraindicated or not tolerated.

The glycoprotein IIb/IIIa antagonists have been tried in conjunction with thrombolysis in acute myocardial infarction, but the various regimens used in recent trials did not confer any additional benefit over conventional treatment. However, there was some evidence of more rapid and complete reperfusion, and these agents warrant further evaluation and refinement.

Anticoagulant treatment

Long term anticoagulation with heparin followed by warfarin is not needed routinely except in patients at higher risk of venous or systemic thromboembolism.

Intracardiac thrombi usually occur within 48 hours after acute myocardial infarction and tend to embolise within the first few weeks. Low dose dalteparin has been shown to reduce the incidence of intramural thrombus (21.9% *v* 14.2%, P = 0.03) in patients given thrombolytic treatments, although this is at a risk of small increase in minor bleeding complications. Thus, in patients at high risk of mural thrombus formation, dalteparin should be started as soon as possible after the diagnosis of acute myocardial infarction.

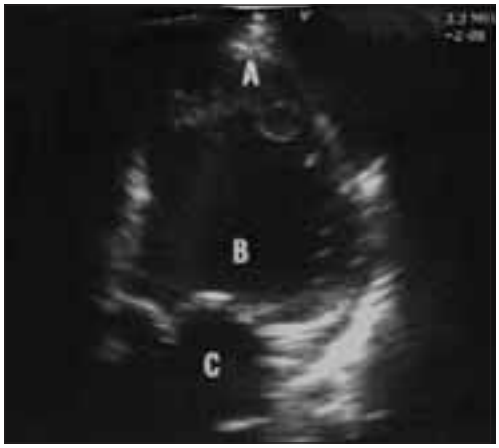
Warfarin should be continued for two to three months, except in the case of atrial fibrillation, when it may be maintained indefinitely. While a patient is taking warfarin, aspirin use may increase the risk of bleeding, but, pending further evidence, many clinicians still continue to use low dose aspirin for its antiplatelet effect. Although thrombus is commonly associated with left ventricular aneurysm (up to 60%), systemic emboli are uncommon (4-5%), and long term anticoagulation does not seem to further reduce the risk of systemic embolisation; thus, anticoagulant treatment is not currently indicated in these patients in the long term.

Venous thromboembolism is often associated with acute myocardial infarction, although its incidence has fallen since the introduction of thrombolytic treatment. Although no trials have compared the efficacy of low molecular weight heparin with unfractionated heparin in preventing venous thromboembolism after acute myocardial infarction per se, it is likely that these agents are equally effective, and are increasingly used in clinical practice.

The box showing antithrombotic therapy in acute Q wave myocardial infarction and after myocardial infarction is adapted from the 6th ACCP consensus conference on antithrombotic therapy. The figure showing lives saved in relation to time of administration of thrombolytic treatment from onset of symptoms of chest pain is adapted from Collins R, et al, *N Engl J Med* 1997;336:847-60.

Risk factors for systemic embolisation when anticoagulation should be considered

- Large anterior wall myocardial infarction
- Myocardial infarction complicated by severe left ventricular dysfunction
- Congestive heart failure
- Echocardiographic evidence of mural thrombus or left ventricular aneurysm
- Previous emboli
- Atrial fibrillation



Echocardiogram showing thrombus at left ventricular apex in patient with dilated cardiomyopathy (A=thrombus, B=left ventricle, C=left atrium)

Further reading

- Cairns JA, Theroux P, Lewis D, Ezekowitz M, Meade TW. Antithrombotic agents in coronary artery disease. *Chest* 2001;119:228–52S
- Collins R, MacMahon S, Flather M, Baigent C, Remvig L, Mortensen S, et al. Clinical effects of anticoagulant therapy in suspected acute myocardial infarction: systematic overview of randomised trials. *BMJ* 1996;313:652-9
- ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* 1988;II:349-60
- Oldroyd KG. Identifying failure to achieve complete (TIMI 3) reperfusion following thrombolytic treatment: how to do it, when to do it, and why it's worth doing. *Heart* 2000;84:113-5
- Mounsey JP, Skinner JS, Hawkins T, MacDermott AF, Furniss SS, Adams PC, et al. Rescue thrombolysis: alteplase as adjuvant treatment after streptokinase in acute myocardial infarction. *Br Heart J* 1995;74:348-53
- The GUSTO Investigators. An international randomized trial comparing 4 thrombolytic strategies for acute myocardial infarction. *N Eng J Med* 1993;329:673-82
- National Institute for Clinical Excellence. *Technology appraisal guideline no 52. Guidance on the use of drugs for early thrombolysis in the treatment of acute myocaardial infarction.* London: NICE, 2002
- Ohman EM, Harrington RA, Cannon CP, Agnelli G, Cairns JA, Kennedy JW. Intravenous thrombolysis in acute myocardial infarction. *Chest* 2001;119:253-77S

11 Antithrombotic therapy in acute coronary syndromes

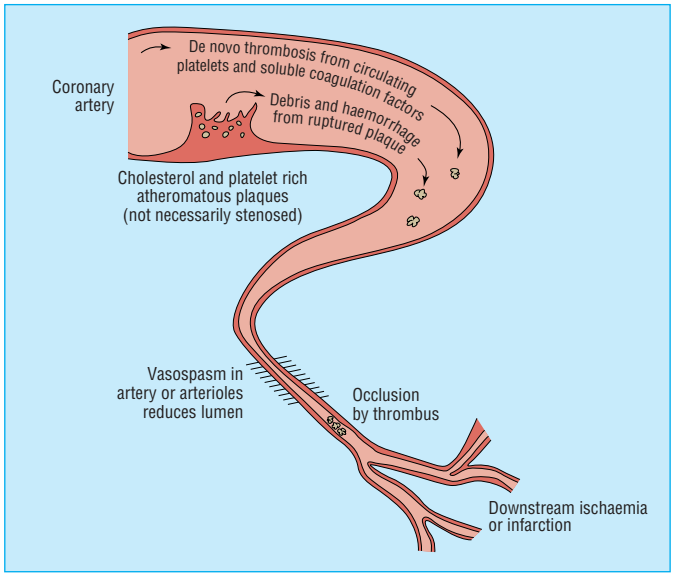
Robert D S Watson, Bernard S P Chin, Gregory Y H Lip

The use of antithrombotic therapy in acute coronary syndromes has reduced the incidence of death and Q wave myocardial infarction dramatically in recent years. Antithrombotic drugs in routine use include antiplatelet drugs (aspirin, clopidogrel, and glycoprotein IIb/IIIa receptor antagonists) and anticoagulants (unfractionated and low molecular weight heparin, warfarin, and direct thrombin inhibitors).

Pathogenesis

Thrombosis is the basic pathophysiological process underlying the acute coronary syndromes. Thus, antithrombotic therapy is the cornerstone of management, and the appropriate choice of antithrombotic drugs to reduce platelet aggregation or interfere with the clotting process can be critical.

Rupture of the fibrous cap of an atheromatous plaque exposes the lipid core, which is highly thrombogenic and contains an abundance of procoagulant tissue factor. Plaque rupture (exposing surface binding glycoproteins) allows platelets to adhere to the plaque, become activated, and release thromboxane A2, which causes further platelet aggregation and vasoconstriction. As the platelets aggregate around the ruptured plaque, membrane glycoprotein IIb/IIIa receptors undergo a configuration change to bind fibrinogen and form a complex platelet linkage. Further incorporation of fibrin and red blood cells within this platelet-rich thrombus results in a partial or total occlusion of the coronary artery. Alternatively, thrombus may break off from a ruptured plaque and occlude a downstream vessel. Occlusion may also follow from trapping of circulating thrombi formed elsewhere in the circulation.



Thrombosis in relation to acute coronary syndromes

Antithrombotic drugs

Antiplatelet drugs

Aspirin

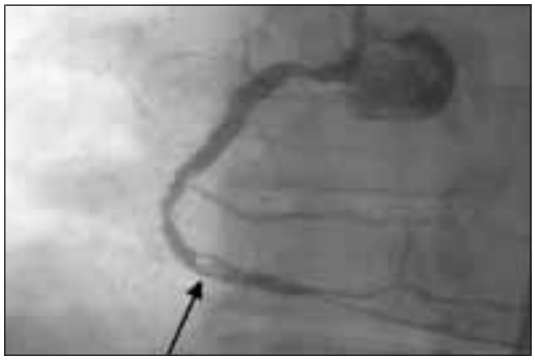
Aspirin has been in use for more than 150 years and is cheap and effective. It has been shown to reduce the risk of fatal and non-fatal myocardial infarction by at least 50% in patients with unstable angina. Aspirin blocks cyclo-oxygenase and formation of thromboxane A2, thus reducing platelet aggregation induced via this pathway.

Aspirin is the cornerstone of treatment in acute coronary syndromes and chronic coronary artery disease. The beneficial effects of aspirin seem to be sustained for at least two years and regardless of the dose used. However, 75-150 mg daily may have a lower incidence of gastrointestinal side effects than the higher doses used in some randomised studies.

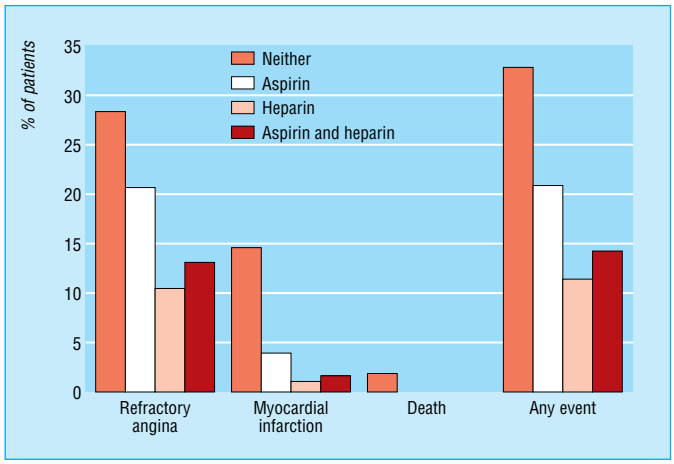
ADP receptor antagonists

Ticlopidine and clopidogrel are ADP inhibitors. Evidence exists that ticlopidine reduces mortality, recurrent infarction, stroke, and angina at least to six months after myocardial infarction or unstable angina. Ticlodipine has fewer gastrointestinal effects than aspirin but may cause reversible neutropenia and thrombocytopenia (< 1% of patients), which dictates therapeutic monitoring with regular blood counts.

Clopidogrel is a derivative of ticlopidine that seems to be six times more effective than its predecessor in preventing platelet



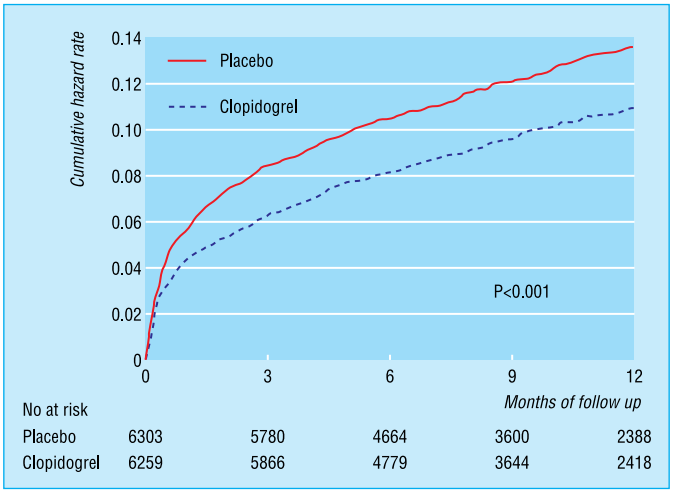
Thrombus within right coronary artery (arrow) in a patient with unstable angina



Reduction of adverse events in patients treated with aspirin, heparin, or both compared with neither drug

aggregation. Clopidogrel has shown better tolerability with less bleeding than aspirin and fewer haematological side effects than ticlopidine. In the CURE trial—a randomised, double blind, parallel group study of 12 562 patients with acute coronary syndrome or non-Q wave myocardial infarction—patients received aspirin 75-325 mg and then were randomly assigned to additional clopidogrel (300 mg load followed by 75 mg daily) or placebo for three months to a year. Additional clopidogrel resulted in a 20% relative risk reduction in the primary end point (cardiovascular death, myocardial infarction, or stroke) ($P < 0.0001$), mainly caused by a 23% relative reduction in myocardial infarction. However, there was a 34% excess of major bleeding (3.6% *v* 2.7% in placebo; $P = 0.003$).

These observations raise the question of whether combination antiplatelet treatment (such as aspirin with clopidogrel) is preferable to other treatments (such as heparin with glycoprotein IIb/IIIa receptor inhibitors) in the acute phase of acute coronary syndromes and suggest that prolonged antiplatelet treatment is better for high risk patients.



Effects of clopidogrel plus aspirin in patients with acute coronary syndromes without ST segment elevation. From the clopidogrel in unstable angina to prevent recurrent events (CURE) trial investigators

Glycoprotein IIb/IIIa receptor inhibitors

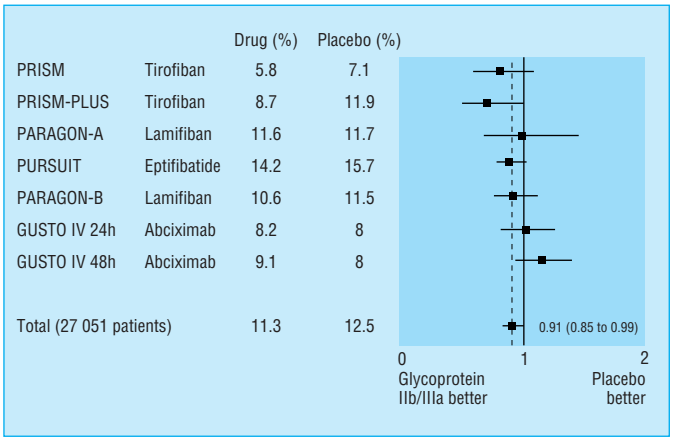
Abciximab is a (large molecule) monoclonal antibody and the first glycoprotein IIb/IIIa receptor antagonist to be developed. Eptifibatide is a peptide receptor antagonist, whereas tirofiban is a non-peptide receptor antagonist. Both eptifibatide and tirofiban are small molecules, apparently non-immunogenic, and therefore suitable for repeat infusions. They have a shorter half life (90-120 minutes) than abciximab (12 hours). Because they are mainly renally cleared, their doses should be adjusted in patients with renal impairment.

In trials with patients with acute coronary syndromes the rate of death, reinfarction, and refractory angina was reduced when glycoprotein IIb/IIIa inhibitors were added to aspirin and heparin. In the largest study, PURSUIT, a bolus injection of eptifibatide followed by a 72 hour infusion resulted in a 9.6% relative risk reduction in death and myocardial infarction when given to patients with acute coronary syndromes already receiving aspirin and heparin. In the PRISM and PRISM-PLUS studies, tirofiban, when given in addition to aspirin and heparin, achieved a 43% relative risk reduction in mortality and reinfarction at seven days. This benefit was sustained at 30 days. Patients taking tirofiban and aspirin without heparin had an excess mortality rate, and this treatment arm was stopped early.

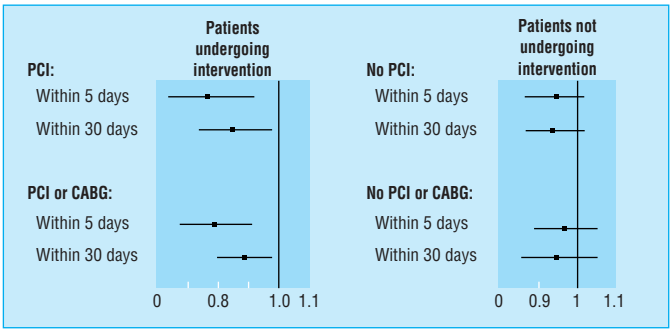
The CAPTURE trial studied patients with acute coronary syndromes scheduled for percutaneous coronary angioplasty. The use of abciximab for about 24 hours before the procedure substantially reduced the risk of mortality, myocardial infarction, or need to proceed to other revascularisation.

The benefits are greatest in patients with elevated levels of troponin T or I, indicating that assessment of subtle indices of cardiac damage predicts patients at higher risk and those most likely to benefit from treatment. Economic evaluations of the costs of using intravenous glycoprotein IIb/IIIa inhibitors suggest that, for patients with elevated troponin levels, 11 are needed to be treated to prevent one death or acute myocardial infarction at 30 days. The equivalent cost effectiveness is about £5000 an outcome.

Using glycoprotein IIb/IIIa inhibitors for acute coronary syndrome in addition to conventional antithrombotic therapy is approved by the British National Institute for Clinical Excellence (NICE). However, interpretation of trial evidence is complicated by patient heterogeneity. Although adjunctive treatment before revascularisation procedures in acute coronary syndromes shows clear benefit, there is some uncertainty of benefit if these drugs are used only as "medical" management without revascularisation.



Glycoprotein IIb/IIIa inhibitors *v* conventional treatment in six trials. Results show odds ratio (95% confidence interval) for death and myocardial infarction at 30 day follow up



Glycoprotein IIb/IIIa inhibitors in patients with acute coronary syndromes: patients undergoing percutaneous coronary intervention and patients not undergoing percutaneous coronary intervention. Results show odds ratio (95% confidence interval) for death or myocardial infarction (CABG=coronary artery bypass graft surgery, PCI=percutaneous coronary intervention)

Anticoagulant treatment

Low molecular weight heparin

Low molecular weight heparins possess more anti-Xa activity than unfractionated heparins. They have a more predictable anticoagulant effect and cause less thrombocytopenia. Titrated appropriately against body weight, low molecular weight heparin provides effective anticoagulation and does not need regular monitoring of the activated partial thromboplastin time. However, platelet count monitoring to detect thrombocytopenia is recommended if treatment is extended beyond a few days. Studies suggest similar safety profiles to unfractionated heparin when used with glycoprotein IIb/IIIa inhibitors.

The therapeutic benefits of low molecular weight heparins are now clear. Recent randomised trials comparing the efficacy of various low molecular weight heparins showed enoxaparin and nadroparin to be more effective in reducing mortality in unstable angina than either aspirin or unfractionated heparin alone. In the ESSENCE study, enoxaparin reduced the incidence of death, myocardial infarction, and reccurent angina at 14 days from 19.6% to 16.6% when compared with unfractionated heparin. This benefit was maintained at 30 days and a year.

Dalteparin in addition to aspirin was more effective than aspirin alone in the FRISC study. It reduced cardiac events and death (1.8% *v* 4.8%) when used in acute coronary syndromes but was no better than adjusted dose unfractionated heparin.

Low molecular weight heparin is, therefore, at least as good as unfractionated heparin in managing unstable angina (with trials showing enoxaparin and nadroparin to be even better). It has practical advantages because it has more consistent antithrombin effects and is easier to administer, and frequent assessment of antithrombotic effect (activated partial thromboplastin time monitoring) is not neccesary. In view of these findings, low molecular weight heparins (enoxaparin and nadroparin) should be used routinely to treat unstable angina concurrently with aspirin in place of unfractionated heparin.

Unfractionated heparin

An uncontrolled study showed that heparin treatment in unstable angina reduced the risk of progression to myocardial infarction by 80%. Other studies show a definite reduction in the incidence of refractory angina and myocardial infarction with the use of unfractionated heparin compared with placebo (risk reduction 0.29). Treatment with heparin and aspirin seems to be more effective than either heparin or aspirin alone. In a meta-analysis on the effect of adding heparin to aspirin in patients with unstable angina, combination treatment resulted in 33% reduction in deaths or myocardial infarction.

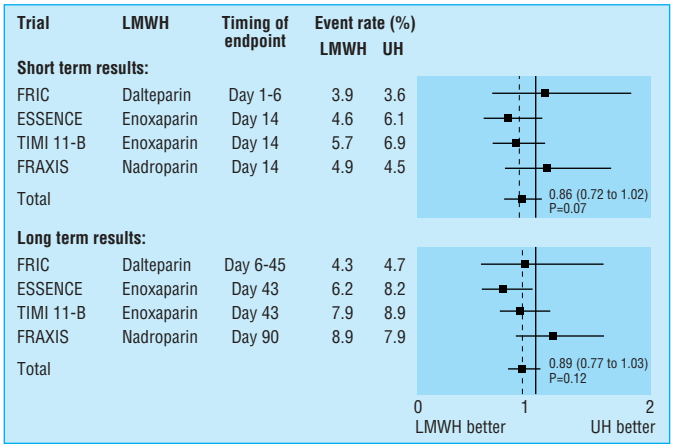
Direct thrombin inhibitors

Newer and powerful anticoagulants, such as the direct thrombin inhibitor hirudin, are being investigated. Several trials compared the effects of hirudin with unfractionated heparin and showed a slight reduction in primary end points, and a pooled analysis of these trials showed a 22% relative risk reduction in myocardial infarction or deaths at 72 hours. Bleeding, however, may be a problem. Hirudin is now approved for patients with heparin induced thrombocytopenia but is not yet licensed for treatment of acute coronary syndromes. The direct thrombin inhibitors melagatran and ximelagatran are currently under investigation for use in patients with acute coronary syndromes.

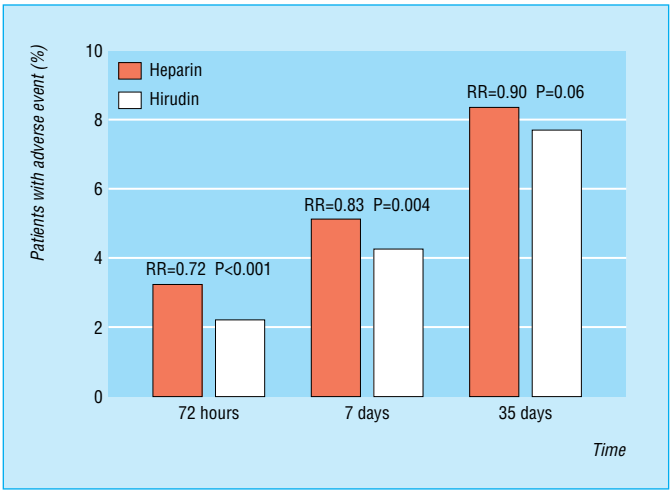
Warfarin

Several trials have investigated using warfarin for unstable angina. In many studies, warfarin was given to patients two days after unfractionated heparin infusion and continued for ten weeks to six months with an international normalised ratio (INR) adjusted

Antithrombotic therapy is the cornerstone of management in acute coronary syndromes



Comparison of low molecular weight heparins with unfractionated heparins in patients with acute coronary syndromes. Results show odds ratio (95% confidence interval) for death and myocardial infarction at long and short term follow up (LMWH=low molecular weight heparin, UH=unfractionated heparin)



Comparison of hirudin therapy with unfractionated heparin in patients with acute coronary syndromes (RR=relative risk)

to between 2.0 and 3.0. Earlier studies showed low fixed doses of warfarin gave no benefit. However, limited evidence suggests that in patients with acute coronary syndromes, combination treatment with aspirin and warfarin adjusted to INR 2.0-3.0 may reduce event rates and hospitalisations.

Pending further trials and in view of potential bleeding risks and need for INR monitoring, warfarin is best reserved for patients with other indications for warfarin use such as coexistent atrial fibrillation, known left ventricular aneurysm with mural thrombus formation, or previous recurrent thromboembolic stroke. Warfarin should not be used as the sole antithrombotic drug in acute cases because of its delayed onset of action.

Thrombolytic treatment

Thrombolytic treatment is not recommended for unstable angina and non-Q wave infarction. It has not been shown to reduce cardiovascular events and death and may worsen clinical outcomes by increasing bleeding complications, as well as increasing bleeding within a ruptured atherosclerotic plaque. This worsens obstruction to coronary flow and exposes clot-bound thrombin to flowing blood, creating an even more prothrombotic environment. In addition, the thrombus in unstable angina is often platelet-rich, unlike the fibrin-rich thrombus seen in infarction, and thus may be less responsive to thrombolytic treatment.

The figure showing reduction of adverse events in patients treated with aspirin, heparin sodium, or both compared with neither drug is adapted from Theroux P et al, *N Engl J Med* 1988:319;1105-11. The effects of clopidogrel plus aspirin in patients with acute coronary syndromes is adapted from the CURE trial investigators, *N Engl J Med* 2001;345: 494-502. The figures showing glycoprotein IIb/IIIa inhibitors *v* conventional treatment, comparing the effects glycoprotein IIb/IIIa inhibitors in acute coronary syndrome patients who are undergoing percutaneous coronary intervention with those who are not, and comparing low molecular weight heparin with unfractionated heparin in patients with acute coronary syndrome are all adapted from Bertrand M, et al, *Eur Heart J* 2002;23:1809-40. The figure comparing hirudin treatment with unfractionated heparin in patients with acute coronary syndromes is adapted from Glenn N et al, *Arch Intern Med* 2001;161:937-48

Antithrombotic therapy in acute coronary syndromes: a summary

Aspirin
- Acute treatment with aspirin is used in all patients with suspected acute coronary syndromes in the absence of contraindications and for long term treatment thereafter

Clopidogrel
- In acute coronary syndrome patients clopidogrel is used for acute treatment and for longer term treatment for at least 9-12 months. Beyond this, treatment will depend on the risk status of the patient and individual clinical judgment
- Clopidogrel is used for immediate and long term treatment in patients who do not tolerate aspirin and in patients receiving a stent
- Clopidogrel is given to acute coronary syndrome patients scheduled for angiography unless there is a likelihood that the patient will proceed to urgent surgery (within 5 days)

Low molecular weight heparin
- In aspirin treated patients, low molecular weight heparin is better than placebo. There are also data in favour of low molecular weight heparin (enoxaparin) over unfractionated heparin when administered as an acute regimen

Glycoprotein IIb/IIIa receptor inhibitors
- Treatment with a glycoprotein IIb/IIIa receptor blocker is recommended in all patients with acute coronary syndromes undergoing percutaneous coronary interventions. The infusion should be continued for 12 hours (abciximab) or 24 hours (eptifibatide, tirofiban) after the procedure
- Medical treatment with a glycoprotein IIb/IIIa receptor blocker during the first days after admission, followed by percutaneous coronary intervention or bypass surgery, yields a significant reduction in death and non-fatal myocardial infarction at 72 hours, from 4.3 to 2.9%
- Diabetic patients with acute coronary syndrome derive particular benefit from glycoprotein IIb/IIIa receptor inhibitors

Fibrinolytic treatment
- In contrast to acute coronary syndromes *with* ST segment elevation, thrombolytic therapy is not recommended for patients with acute coronary syndromes *without* persistent ST segment elevation

Further reading

- Anand SS, Yusuf S, Pogue J, Weitz JI, Flather M. Long term oral anticoagulant therapy in patients with unstable angina or suspected non-Q wave myocardial infarction. The OASIS (warfarin) substudy. *Circulation* 1998;98:1064-70
- British Cardiac Society Guidelines and Medical Practice Committee, and Royal College of Physicians Clinical Effectiveness and Evaluation Unit. Guideline for the management of patients with acute coronary syndrome without persistent ECG ST segment elevation. *Heart* 2001;85:133-42
- The CAPTURE Investigators. Randomised placebo controlled trial of abciximab before and during intervention in refractory unstable angina: The CAPTURE study. *Lancet* 1997;349:1429-35
- Cohen M, Demers C, Gurfinkel EP, Turpie AG, Fromell GJ, Goodman S, et al. A comparison of low molecular weight heparin with unfractionated heparin for unstable coronary artery disease (ESSENCE). *N Engl J Med* 1997;337:447-52
- Fragmin and fast revascularisation during instability in coronary artery disease (FRISC-II) Investigators. Long term low molecular mass heparin in unstable coronary artery disease: FRISC-II prospective multicentre randomised study. *Lancet* 1999;354:701-7
- PRISM-PLUS Investigators. Inhibiton of the platelet glycoprotein IIb/IIIa receptor with tirofiban in unstable angina and non Q-wave myocardial infarction. *N Engl J Med* 1998;338:1488-97
- The PURSUIT Trial Investigators. Inhibiton of platelet glycoprotein IIb/IIIa with eptifibatide in patients with acute coronary syndromes. *N Engl J Med* 1998;339:436-43
- National Institute for Clinical Excellence (NICE). Guidance on the use of glycoprotein IIb/IIIa inhibitors in the treatment of acute coronary syndromes. NICE Technology Appraisal Guidance No 12. London: NICE, 2000
- Task force on the management of acute coronary syndromes of the European Society of Cardiology. Management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. *Eur Heart J* 2002;23:1809-40

12 Antithrombotic strategies in acute coronary syndromes and percutaneous coronary interventions

Derek L Connolly, Gregory Y H Lip, Bernard S P Chin

Acute coronary syndromes

All patients suspected of having acute coronary syndrome should be managed as medical emergencies and monitored in the critical care unit. Baseline tests must include 12 lead electrocardiography, chest *x* ray examination, and venous blood samples for analyses of haemoglobin and markers of myocardial damage, preferably cardiac troponin T or I.

Initial management

Assessment

Patients with persistent ST segment elevation on 12 lead electrocardiography need immediate reperfusion treatment (thrombolysis or intervention). Patients with ST segment depression, inverted T waves, or pseudonormalisation of T waves on the electrocardiogram, but with a clinical history suggesting cardiac ischaemia should receive initial treatment for angina.

This would include aspirin 300 mg followed by a low dose of 75-150 mg daily. In cases of aspirin intolerance, clopidogrel should be used. β Blockers and nitrates should also be given. Rate limiting calcium antagonists can be used if β blockers are contraindicated or are already being used. Ideally, patients should be given low molecular weight heparin (such as enoxaparin) according to their weight. If low molecular weight heparin is unavailable, unfractionated heparin may be used. A bolus of 5000 U is given, followed by an infusion adjusted to get an activated partial thromboplastin time (APTT) ratio of 1.8 to 2.5. In light of data from the CURE and PCI-CURE study, clopidogrel (given for at least one month and up to nine months) should be considered in addition to aspirin when an early non-interventional approach is planned. The optimal dose of aspirin to limit bleeding is probably 75 mg, particularly with clopidogrel. A glycoprotein IIb/IIIa receptor inhibitor should be added to aspirin and heparin for patients in whom catheterisation and percutaneous coronary intervention are planned. In these patients clopidogrel could be considered if they are not at high risk for bleeding.

Observation

Patients should be observed over the next eight to 12 hours. Patients at high risk of progressing to acute myocardial infarction or death should receive a glycoprotein IIb/IIIa receptor inhibitor (eptifibatide or tirofiban) in addition to heparin and aspirin or clopidogrel (alone or with asprin). Abciximab would be used in high risk patients undergoing percutaneous coronary intervention. There is no role for thrombolytic therapy in patients without acute ST segment elevation, except in the situations of a true posterior myocardial infarction, or a presumed new left bundle branch block.

Subsequent management

When patients have been free from symptoms and ischaemic electrocardiographic changes for >48 hours, and any intravenous treatments and heparin have been stopped for >24 hours, risk assessment with stress testing should be performed unless contraindicated. Stress testing for risk assessment is unnecessary if a patient is already in a high risk category for which coronary angiography is indicated.

High and low risk patients with suspected acute coronary syndromes

High risk

- Recurrent or persistent chest pains with associated electrocardiographic changes (ST segment depression or transient ST elevation) despite anti-ischaemic treatment
- Elevated troponin concentrations
- Age >65 years
- Comorbidity, especially diabetes
- Development of pulmonary oedema or haemodynamic instability within observation period
- Development of major arrhythmias (repetitive ventricular tachycardia or ventricular fibrillation)
- Early postinfarction unstable angina

Low risk

- No recurrence of chest pains within observation period
- Troponin or other markers of myocardial damage not elevated
- No ST segment depression or elevation on electrocardiogram (T wave inversion is classified as intermediate risk)

Management strategies for patients with suspected acute coronary syndromes, with risk stratification by troponin and stress tests*

Low risk

Results of tests

- Cardiac troponin result is negative or low (troponin T <0.01 μg/l or troponin I equivalent) on two occasions
- Stress test indicates a low risk category

Action

- If free from cardiac symptoms, no more cardiac interventions needed
- Subsequent outpatient review appropriate for further investigations and adjustment of drug treatment

Intermediate risk

Results of tests

- Impaired left ventricular function, or haemodynamic abnormalities or arrhythmia during the acute phase *but*
- Normal cardiac troponin result (troponin T <0.01 μg/l, or troponin I equivalent), with a stress test indicating intermediate risk *or*
- Moderately elevated cardiac troponin (troponin T 0.01-0.1 μg/l, or troponin I equivalent) with stress test indicating low risk category

Action

- Many cardiologists perform coronary angiography on these patients, but clear evidence of benefit is lacking

High risk

Results of tests

- Maximal cardiac troponin result is high (troponin T >0.1 μg/l, or troponin I equivalent) *or*
- Stress test indicates high risk category

Action

- Coronary angiography should be arranged, unless contraindicated, and performed urgently, before discharge from hospital
- Patients with suitable lesions for percutaneous coronary intervention should be given clopidogrel, which should also be given to patients with coronary lesions not suitable for any revascularisation

*If patient is unable to perform an exercise electrocardiogram, an alternative non-exercise (pharmacological) stress test, such as a stress echocardiograph or isotope myocardial stress perfusion study, should be arranged unless contraindicated. In all cases, aggressive risk factor management and regular aspirin treatment (or clopidogrel, or both, depending on clinical situation) is necessary

Antithrombotic treatment

Low molecular weight heparin should be given for at least two days, and for up to eight days or longer in cases of recurrent ischaemia or where myocardial revascularisation is delayed or contraindicated. Patients requiring a bypass operation may have their glycoprotein IIb/IIIa receptor antagonist infusion stopped before or at the time of cardiac surgery, although clopidogrel should be withheld for five to seven days.

Risk stratification and antithrombotic strategies

Recent trials have shown that patients with elevated troponin benefit from treatment with low molecular weight heparin, glycoprotein IIb/IIIa blockers, or an invasive strategy, whereas patients without troponin elevation showed no such benefit.

In these high risk patients, angiography with a view to revascularisation should be performed on the same admission. Infusion of glycoprotein IIb/IIIa receptor inhibitors should be started while waiting and preparing for angiography and continued for 12 hours (abciximab) or 24 hours (tirofiban) after angioplasty is performed.

Low risk patients can be mobilised and discharged if (at least 12 hours after the onset of symptoms of a suspected acute coronary syndrome) the symptoms have not recurred, cardiac troponin concentrations are normal, electrocardiograms remain normal (or unchanged compared with a recording from before the current presentation), and cardiac enzyme activities are not raised. Risk assessment with stress testing should be performed before a patient is discharged unless contraindicated.

Pros and cons of invasive strategy

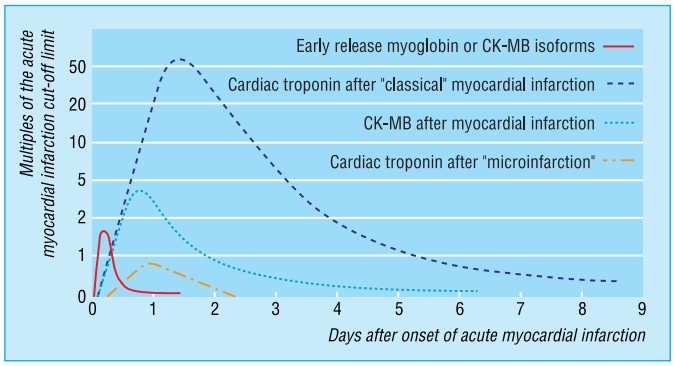
There are arguments for and against an invasive approach to acute coronary syndromes. In the FRISC-II trial an invasive strategy had, after a year, saved 1.7 lives in 100 patients and prevented 2.0 non-fatal myocardial infarctions and 20 readmissions. It provided earlier and better symptom relief at the cost of 15 more patients with coronary artery bypass grafting and 21 more with percutaneous transluminal coronary angioplasty, and these results were independent of treatment with dalteparin or placebo. In the BHF RITA3 trial of patients with unstable angina, myocardial infarction, or non-ST segment elevation, an invasive strategy reduced refractory or severe angina, with no increase of death or myocardial infarction, compared with a conservative strategy. Against these benefits is the need to have adequate provision of facilities and trained staff to undertake such procedures.

Percutaneous coronary intervention

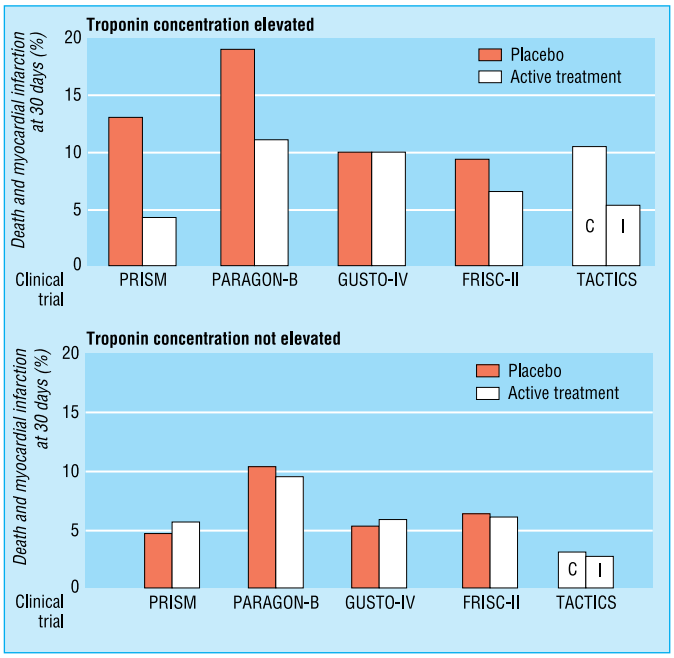
Arterial thrombi occur soon after percutaneous coronary interventions for coronary artery disease, usually at the site of the dilated segment. Arterial thrombi are rich in platelets, red blood cells, fibrin, and leucocytes and may contribute to vessel reocclusion with the consequent need for revascularisation. The risk of reocclusion depends on the extent of segment dilatation and vessel injury, as well as local shear forces. Antiplatelet and antithrombin drugs generally reduce the risk of occlusion or the need for further intervention but are not perfect. Where facilities are available, percutaneous coronary angioplasty is an alternative to thrombolytic treatment as a means of reperfusion in acute myocardial infarction.

Antiplatelet treatment

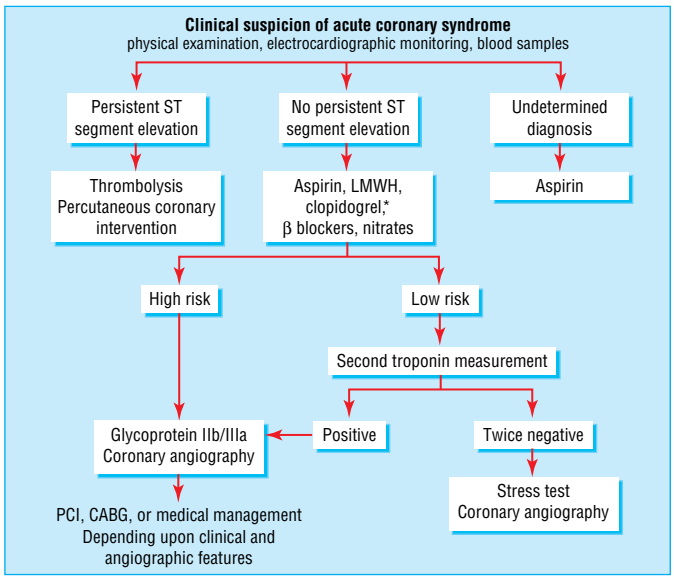
Patients with coronary artery disease undergoing angioplasty should continue taking antiplatelet drugs as usual. For patients not receiving regular antiplatelet treatment, aspirin 100-325 mg should be given orally at least two hours before angioplasty.



Time course of different cardiac biochemical markers



Death or myocardial infarction in patients with elevated troponin concentration or negative troponin result in contemporary trials. The FRISC-II trial also used low molecular weight heparin, and the bars for the TACTICS trial show the strategies used (I=invasive strategy, C=conservative strategy)



European Society of Cardiology recommended strategy for acute coronary syndromes (CABG=coronary artery bypass graft, LMWH=low molecular weight heparin, PCI=percutaneous coronary intervention. *Omit clopidogrel if patient likely to go to CABG within 5 days)

Aspirin substantially reduces the rate of intracoronary thrombus formation at the treatment site and restenoses. The addition of dipyridamole to aspirin adds little extra benefit and is not recommended. Ticlopidine alone has not been shown to be more effective than aspirin alone in patients undergoing percutaneous interventions. Although clopidogrel is marginally better than aspirin in patients with atherosclerotic vascular disease (CAPRIE study), direct comparisons between aspirin and clopidogrel in coronary intervention have not revealed marked differences. Thus, ticlopidine and clopidogrel are useful alternatives for patients scheduled for percutaneous coronary angioplasty who are unable to take aspirin.

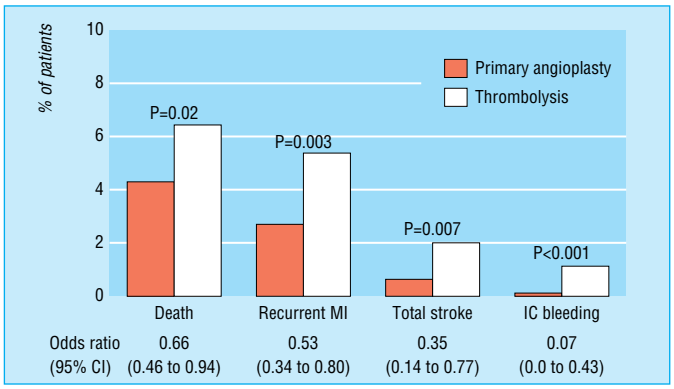
After intervention, antiplatelet combination treatment (aspirin plus ticlopidine) is superior to aspirin alone at reducing ischaemic complications and cardiac events, particularly after intracoronary stent placements. The stent anticoagulation restenosis investigators (SARI) trial, compared aspirin-ticlopidine combination treatment with aspirin-warfarin combination treatment and aspirin alone and showed that patients taking aspirin-ticlopidine had the best 30 day mortality (0.5% *v* 2.7% *v* 3.6% respectively, P = 0.001). Total bleeding complications occurred in 5.5% of those taking aspirin-ticlopidine, quite high when compared with 1.8% in those taking aspirin only (P < 0.001), and the incidence of neutropenia was not significantly different. The CLASSICS trial showed clopidogrel-aspirin combination treatment to be as effective as ticlopidine-aspirin combination treatment after angioplasty and stent placement.

Glycoprotein IIb/IIIa receptor antagonists

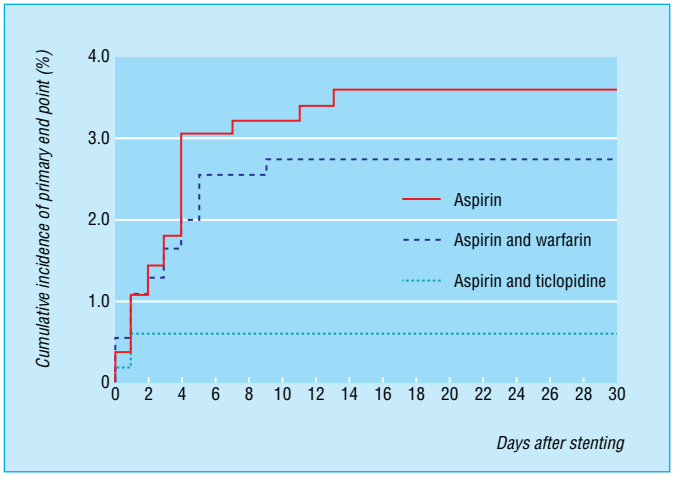
Despite adequate treatment with antiplatelet drugs, platelet activation still continues along other pathways not blocked by these agents. This is where the glycoprotein IIb/IIIa receptor antagonists, which block the final common pathways of platelet aggregation, have contributed most to the management of angioplasty and stenting.

Abciximab, eptifibatide, and tirofiban have all been shown to reduce reocclusion and cardiovascular events, including deaths and myocardial infarctions, at 30 days when used in patients undergoing elective and urgent angioplasty and stenting. These benefits are additional to those achieved with antiplatelet drugs, and the effects were most prominent with abciximab. The EPIC, EPILOG, and CAPTURE trials all showed that abciximab infusion reduced major complication rates during balloon angioplasty, a benefit that was sustained at 30 days' follow up. The EPISTENT trial showed that abciximab reduces major complications during stent placement and was superior to a combination of abciximab and balloon angioplasty. In the only direct comparison of glycoprotein IIb/IIIa antagonists, the TARGET trial randomised 5308 patients to tirofiban or abciximab before percutaneous transluminal coronary angioplasty or stent, or both: by six months the primary end point was similar in both treatment arms (14.9% with tirofiban *v* 14.3% with abciximab, P > 0.05), as was mortality (1.9% *v* 1.7%, P > 0.05). Patients with unstable angina, acute myocardial infarction, and other risk factors (such as diabetes) for postprocedure in stent thrombosis or restenosis stand to benefit most from glycoprotein IIb/IIIa receptor antagonists.

In light of this, glycoprotien IIb/IIIa drugs should be considered in all patients at risk of developing in stent stenosis or with acute coronary syndrome scheduled for percutaneous coronary interventions. If percutaneous coronary intervention is planned in unstable angina, glycoprotein IIb/IIIa receptor antagonist infusions should be started before intervention and continued for 12 hours (abciximab) or 24 hours (tirofiban, eptifibatide) after the procedure.



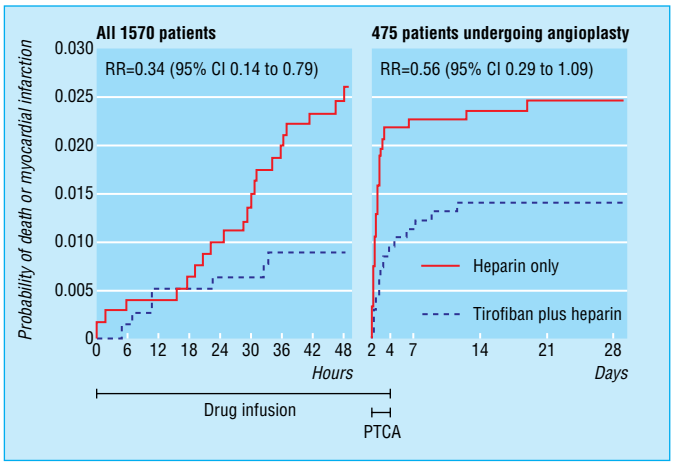
Meta-analysis of 10 randomised trials that compared thrombolytic treatment with primary angioplasty in acute myocardial infarction (MI=myocardial infarction, IC=intracranial)



Cumulative incidence of primary end points (mortality, target lesion revascularisation or thrombosis, non-fatal myocardial infarction) in patients treated with aspirin alone (557 patients), aspirin and warfarin (550 patients), or aspirin and ticlopidine (546 patients) after coronary artery stenting

Some factors prediposing to in stent thrombosis after placement

- Underdilation of the stent
- Proximal and distal dissections
- Vessel diameter < 3 mm
- Poor inflow
- Outflow obstruction



PRISM-PLUS results showing cumulative incidence of death or myocardial infarction (PTCA=percutaneous transluminal coronary angioplasty, RR=relative risk)

Several economic evaluations have found that routine use of glycoprotein IIb/IIIa drugs after percutaneous coronary interventions is extremely cost effective for patients at high risk of myocardial infarction or death. In such patients, the number needed to treat to save one life or prevent one acute myocardial infarction at 30 days may be as low as 30, or about £5000 per outcome.

Anticoagulant treatment

Most patients with acute coronary syndromes undergoing angioplasty would have been pretreated with heparin. Several small studies have shown that patients with unstable angina who receive heparin before intervention have a higher rate of success and lower postprocedure reocclusion rates.

Increasing numbers of patients with unstable angina are now being treated with low molecular weight heparin. However, such drugs tend to have a longer half life than unfractionated heparin and their effects are not completely reversed by protamine sulphate if necessary. Low molecular weight heparins can be safely substituted for unfractionated heparin as a procedural anticoagulant during percutaneous coronary intervention.

During percutaneous coronary interventions, heparin should be given to avoid postprocedure complications. The dose of unfractionated heparin given should be sufficient to increase the activated clotting time (ACT) to 250-300 seconds as measured with the HemoTec device (or 300-350 seconds with the Hemochron device). Unfractionated heparin dose may need to be adjusted for weight or sex. If glycoprotein IIb/IIIa receptor agonists are being used, then unfractionated heparin boluses should be reduced to achieve a target ACT of about 200 seconds. Although the traditional means of assessing heparin anticoagulation has been with the APTT, the ACT is an assay of whole blood clotting time that can be performed rapidly at the bedside and catheterisation laboratory.

Routine use of unfractionated heparin (either as infusion or subcutaneously) after angioplasty is probably not indicated for uncomplicated procedures. Studies have shown excess bleeding complications with heparin treatment without a reduction in the number of cardiac ischaemic events. Patients who do not receive heparin treatment after coronary interventions can have their femoral sheaths removed earlier, resulting in shorter hospital stays, fewer bleeding complications, at the risk of a similar incidence of cardiac end points including reocclusion.

With the advent of glycoprotein IIb/IIIa receptor antagonists, heparin infusions postprocedure should not be necessary routinely. Femoral sheaths should be removed once the ACT has fallen to less than 150-180 seconds. Adjunctive treatment with low molecular weight heparin or unfractionated heparin may still be warranted after angioplasty and stent implantation in patients at high risk of in stent thrombosis.

Full anticoagulation with heparin followed by warfarin in patients undergoing angioplasty with stenting is no better at reducing the number of adverse effects than combination treatment with aspirin and ticlopidine, but at increased risk of bleeding with warfarin. Use of hirudin, the direct thrombin inhibitor, was associated with a reduction of early cardiac events and restenosis at 96 hours but was no different from the heparin treatment arm at seven months.

Antithrombotic therapy in coronary angioplasty and stent placement procedures

Before procedure

- Aspirin 80-325 mg once daily at least 2 hours before procedure. Ticlopidine 250 mg twice daily or clopidogrel 75 mg once daily started 24 hours before procedure if aspirin contraindicated
- Glycoprotein IIb/IIIa receptor antagonists should be considered in high risk patients with acute coronary syndromes

During and after procedure

- Heparin* bolus to achieve activated clotting time (ACT) ~300 seconds. Give 70-150 U/kg or 7000 U for women and 8000 U for men. If ACT not achieved give extra bolus of 2500-5000 U. Reduce heparin bolus to achieve ACT ~200 seconds if glycoprotein IIb/IIIa receptor agonist is to be used
- In high risk patients, abciximab as bolus and infusion should be given at least 10 minutes before angioplasty and stent placement and continued for 12 hours after procedure

After procedure

- Start clopidogrel 300 mg orally, followed by 75 mg daily for 4 weeks
- Remove femoral sheath as soon as ACT falls below 150-180 seconds
- Heparin infusion is not routinely necessary after uncomplicated angioplasty

Heparin infusion after a procedure is indicated if

- Femoral sheath to be retained—Heparin infusion 1000-1200 U/hour until 4 hours before sheath is to be removed. Check ACT and remove sheath when ACT $<$ 150 seconds
- Patients at high risk for in stent thrombosis
- Patients with other indications for anticoagulation, such as atrial fibrillation or mechanical heart valves

*Details given for unfractionated heparin, but low molecular weight heparin can be used as an alternative in percutaneous coronary interventions

Further reading

- The Task Force on the management of acute coronary syndromes of the European Society of Cardiology. Management of acute coronary syndromes in patients presenting without persistent ST segment elevation. *Eur Heart J* 2002;23:1809-40
- Braunwald E, Antman EM, Beasley JW, Califf RM, Cheitlin MD, Hochman JS, et al. ACC/AHA 2002 guideline update for the management of patients with unstable angina and non-ST segment elevation myocardial infarction. *J Am Coll Cardiol* 2002;40:1366-74
- Fox KA, Poole-Wilson PA, Henderson RA, Clayton TC, Chamberlin DA, Shaw TR, et al for the Randomized Intervention Trial of Unstable Angina (RITA) Investigators. Interventional versus conservative treatment for patients with unstable angina or non-ST-elevation myocardial infarction: the British Heart Foundation RITA3 randomised trial. *Lancet* 2002;360:743-51
- Leon MB, Baim DS, Popma JJ, Gordon PC, Cutlip DE, Ho KK, et al for the Stent Anticoagulation Restenosis Study Investigators. A clinical trial comparing three antithrombotic-drug regimens after coronary-artery stenting. *N Engl J Med* 1998;339:1665-71
- Roffi M, Moliterno DJ, Meier B, Powers ER, Grines CL, DiBattiste PM, et al. Impact of different platelet glycoprotein IIb/IIIa receptor inhibitors among diabetic patients undergoing percutaneous coronary intervention: do tirofiban and Reopro give similar efficacy outcomes trial (TARGET) 1 year follow up? *Circulation* 2002;105:2730-6

The meta-analysis of trials comparing thrombolytic treatment with primary angioplasty is adapted from Weaver WD, et al, *JAMA* 1997;278:2093. The figure showing incidence of primary end point in patients treated with aspirin alone, aspirin and warfarin, or aspirin and ticlopidine after coronary artery stenting is adapted from Leon MB, et al, *N Engl J Med* 1998;339:1665-71. The figure showing the results from the PRISM-PLUS study is adapted from PRISM-PLUS Study Investigators. *N Engl J Med* 1998;338:1488-97. The figure showing the time course of cardiac biochemical markers is adapted from Wu AH, et al, *Clin Chem* 1999;45:1104-21. The figures of death or myocardial infarcton in patients with elevated troponins or negative troponin result, and the strategy for acute coronary syndromes are adapted from Bertrand ME, et al, *Eur Heart J* 2002;23:1809-40.

13 Antithrombotic therapy in chronic heart failure in sinus rhythm

Gregory Y H Lip, Bernard S P Chin

Chronic heart failure is one of the few remaining areas in cardiovascular medicine where the use of antithrombotic therapy remains controversial. This is largely because of conflicting outcomes from existing studies, a lack of appropriately conducted randomised clinical trials, and difficulty in defining the syndrome of heart failure and its thromboembolic complications.

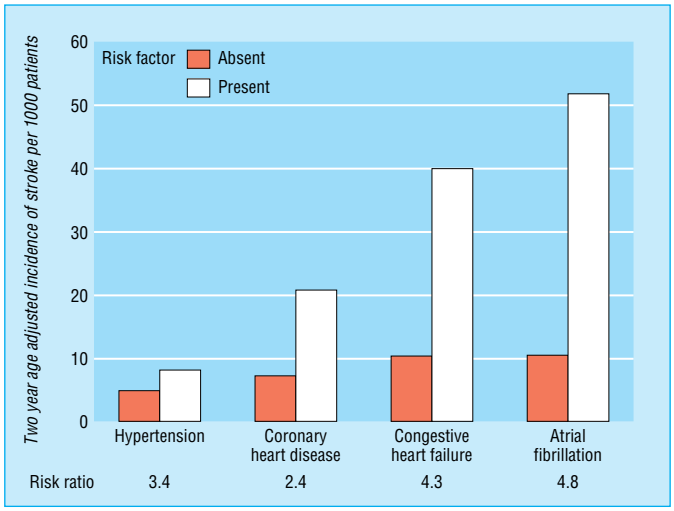
Stroke and systemic embolism in heart failure

Left ventricular dysfunction increases the risk of thrombosis and systemic embolism, and these thromboembolic events add to the high morbidity associated with this condition. In addition, ischaemic and thromboembolic events—particularly stroke, myocardial ischaemia, and myocardial infarction—contribute to the high hospital admission rates in these patients.

The incidence of thromboembolism and the factors associated with a high thromboembolic risk have been addressed in many studies, but the reported incidence of these events seems to vary between studies, depending on study methodology and populations. Retrospective analyses of recent heart failure trials have estimated this risk to be between 1.3% and 4.6% depending on the severity of heart failure. For example, mild to moderate chronic heart failure seems to be associated with an annual stroke risk of about 1.5% (V-HeFT and SOLVD studies), compared with an annual stroke risk in the general population of less than 0.5%. The annual risk of stroke increases to almost 5% in severe chronic heart failure. Evidence from the Framingham study shows that chronic heart failure is a major risk factor for stroke, second only to atrial fibrillation.

Long term oral anticoagulation is beneficial in certain groups of patients with chronic heart failure, but the role of anticoagulation for patients with chronic heart failure in general is less clear. For example, oral anticoagulation is extremely effective in reducing stroke risk and other embolic events in patients with atrial fibrillation and chronic heart failure. Indeed, there is a wide variation in the use of oral anticoagulants in patients with chronic heart failure.

Although oral anticoagulation reduces thromboembolic events in various cardiovascular diseases, the potential risks of bleeding must also be considered. Importantly, the control of



Two year age adjusted incidence of stroke for every thousand patients in the Framingham study

The risk of recurrent strokes in heart failure patients is higher than with initial events. In heart failure patients, second and recurrent stroke rates may be as high as 9% every year

Rates of stroke, pulmonary embolism, myocardial infarction, and total mortality in recent heart failure trials

						Mean (SD) annual risk of events (%)			
Study (NYHA)	*NYHA*	*Mean ejection fraction*	*Mean age*	*Follow up (years)*	*Prevalence of atrial fibrillation (%)*	*Stroke*	*Pulmonary embolism*	*Myocardial infarction*	*Death*
SOLVD	I-III	0.25	60	3.3	6	3.8 (1.3)	5.3 (1.6)	9.6 (2.9)	23.7 (7.2)
V-HeFT-I	II-III	0.30	58	2.3	15	4.1 (1.8)	5.6 (2.5)	—	43.0 (18.7)
V-HeFT-II	II-III	0.29	61	2.6	15	4.7 (1.8)	5.7 (2.2)	5.2 (2.0)	34.7 (13.3)
CONSENSUS	III-IV	—	71	0.5	50	2.3 (4.6)	—	—	46.6 (44)
PROMISE	III-IV	0.21	64	0.5	—	2.0 (3.5)	—	—	27.1 (54)
SAVE	—	0.30	59	3.5	—	4.6 (1.5)	—	13.6 (3.9)	18.9 (5.4)

anticoagulation is reported to be more difficult, and bleeding complications commoner, in chronic heart failure because of hepatic congestion and potential drug interactions that may occur.

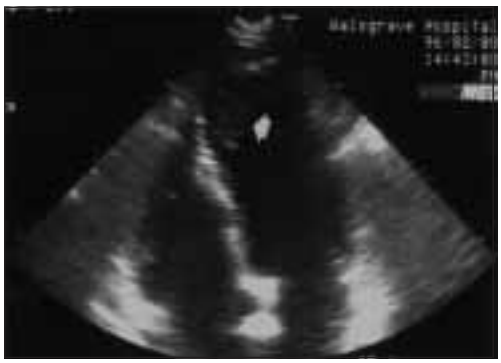
Studies have reported stroke incidences in heart failure of up to 11%. However, many of these trials were small, non-randomised and biased towards more severe disease. Importantly, atrial fibrillation was common among participants, many of whom did not receive warfarin during the period of follow up. Nevertheless, given that the risk of stroke in the general population aged 50-75 years is less than 0.5% a year, an estimated stroke incidence of 2% in patients with chronic heart failure represents a fourfold increased risk. With increasing age, the absolute risk of stroke rises, while its risk relative to the general population falls.

The commonest vascular occlusive event in heart failure is not stroke but myocardial infarction. Sudden cardiac death is common in heart failure and has been attributed to fatal arrhythmia. However, pathological studies of sudden cardiac deaths have detected fresh coronary thrombi in many cases, indicating that acute coronary occlusion may have been the primary event in the death. Many heart failure patients with myocardial infarction also die before reaching hospital. Hence, myocardial infarction and acute coronary occlusion may be more common in heart failure than estimated.

Sedentary patients with chronic heart failure patients are also at increased risk of developing venous thromboembolism, particularly in the legs. Pulmonary embolism can originate from deep vein thrombosis of the legs or, rarely, from a thrombus in the right ventricle.

Stroke risk and heart failure

	Annual risk
Stroke in general population aged 50-75 years	0.5%
Stroke in mild chronic heart failure—NYHA II-III	1.3-1.8%
Stroke in severe chronic heart failure—NYHA III-IV	3.5-4.6%
Recurrent stroke rates in chronic heart failure	9%
Stroke in atrial fibrillation	4.5%
Stroke in atrial fibrillation and chronic heart failure	8-12%



Apical thrombus in patient with poor left ventricular function

Left ventricular thrombus formation

Chronic heart failure patients in sinus rhythm are at high risk of ventricular thrombus formation because they fulfil Virchow's triad for thrombogenesis. These patients have:

- "Abnormal blood flow " with stasis, with low cardiac output, dilated cardiac chambers, and poor contractility of the heart;
- "Abnormal vessel walls", with endothelial damage or dysfunction or both;
- "Abnormal blood constituents", with abnormalities of haemostasis, platelets, and coagulation.

Autopsy and surgical studies have detected ventricular mural thrombi in 25-50% of heart failure patients. Such mural thrombi have been associated with increased mortality. In chronic heart failure with left ventricular dysfunction, abnormal wall motion leads to alterations in regional blood flow and inflow velocity, whereas impaired ventricular contractility results in further intracavitary blood stasis. Evidence exists for local platelet activation and cytokine recruitment, such as tumour necrosis factor, which can trigger the clotting cascade. This is particularly so after acute myocardial infarction, when high levels of catecholamines are circulating freely. Markers of endothelial injury and dysfunction are also elevated in heart failure, although in many cases these may be caused by underlying atherosclerosis. Endothelial damage promotes monocyte and platelet adhesion to the endothelium, predisposing to thrombosis in situ. The interplay of altered local flow characteristics, heightened clotting factors, and endothelial cell dysfunction gives rise to left ventricular thrombus formation.

Certain conditions are more likely to predispose to left ventricular thrombus formation. Patients with a dyskinetic left ventricular segment or left ventricular aneurysm and those who have had acute and extensive myocardial infarction involving the anterior wall are at highest risk. Left ventricular dilatation and severity of left ventricular dysfunction (measured by

Factors predisposing to prothrombotic state in heart failure

Endothelial dysfunction
- Impaired nitric oxide release
- Reduced anticoagulant status

Altered flow characteristic
- Left ventricular dilatation
- Abnormal left ventricular wall motion
- Left ventricular aneurysm
- Reduced pump action

Altered coagulation
- Increased platelet activation and aggregation
- Increased clotting
- Reduced fibrinolytic activity

Neurohormonal and inflammatory activation
- Increased catecholamines
- Increased inflammatory cytokines

ejection fraction) also correlate well with risk of thrombus formation. However, mitral regurgitation may have some protective effect.

Risk factors for systemic embolisation

Given the high prevalence of left ventricular thrombus formation in patients with left ventricular aneurysm and heart failure, it is perhaps surprising that the rate of stroke in patients with documented left ventricular thrombi is low. This could be because of the difficulty in determining the presence of thrombus clinically. The best method is by cross sectional echocardiography, but this can only detect thrombi greater than 0.5 cm in diameter. Microthrombi that forms early in myocardial infarction and which pose a serious threat of embolism would not therefore be detected. Conversely, an organised chronic thrombus, which is more easily detectable on echocardiography is less likely to embolise.

A newly formed left ventricular thrombus is likely to embolise as are thrombi that are mobile, protruding, or pedunculated with a narrow stalk. Left ventricular thrombi present for more than three months would have undergone fibrous organisation and endothelialisation and so are more stable.

The only parameters shown by studies to achieve statistical significance as a predictor of thromboembolic events is peak oxygen uptake during symptom limited maximal exercise testing and the severity of heart failure (measured by left ventricular ejection fraction). The left ventricular ejection fraction is a powerful predictor of stroke in patients who have had a myocardial infarction. Patients with left ventricular ejection fraction less than 28% at particularly high risk. Furthermore, for every absolute decrease of 5% in left ventricular ejection fraction, the risk of stroke increases by 18%.

Patients with idiopathic dilated cardiomyopathy tend to have a higher rate of systemic thromboembolism than patients with ischaemic cardiomyocapthy. Only women with periparturm cardiomyopathy have higher risks of thromboembolism than patients with idiopathic dilated cardiomyopathy. Atrial fibrillation, age, and previous thromboembolic events are independent risk factors for stroke but confer further risks to patients with heart failure. Atrial fibrillation is associated with a stroke risk of about 5% every year.

Factors predisposing to left ventricular thrombus formation and embolisation

Left ventricular thrombus formation
- After acute extensive myocardial infarction
- Acute anterior myocardial infarction
- Left ventricular aneurysm
- Left ventricular dilatation

Systemic embolisation of left ventricular thrombi
- New or acute thrombus (within two weeks of formation)
- Protruding segment
- Normal adjacent wall function

Identifying risk factors that predispose to systemic embolisation of left ventricular thrombi is important because it allows patients at highest risk only to be treated. Currently, the identification of risk factors has been inferred from observational data of heart failure trials and smaller non-randomised studies

Antithrombotic treatment

Warfarin

The need for oral anticoagulation in chronic heart failure has been inferred from mainly observational and retrospective studies of mortality from heart failure. In the PROMISE study, lower incidences of stroke were reported with anticoagulation (1.9% *v* 2.5%). In a follow up study of patients with idiopathic dilated cardiomyopathy, 18% of patients not taking warfarin had a stroke, whereas none occurred among the patients receiving warfarin over 11 years of follow up. Several studies of patients who had myocardial infarction showed that long term warfarin treatment to reduce rates of death, recurrent myocardial infarction, and stroke was effective.

Other studies have also shown that warfarin reduces mortality from sudden coronary death and recurrent myocardial infarction. Both the CONSENSUS and SOLVD studies showed that patients receiving warfarin had a lower mortality than those patients receiving antiplatelet treatment or those without antithrombotic therapy. In a retrospective survey of the SOLVD participants, anticoagulant monotherapy reduced the risk of sudden cardiac death by 32%. Among participants with non-ischaemic heart failure, this risk fell by

Major risk factors for cardioembolic stroke in chronic heart failure
- Atrial fibrillation
- Mitral stenosis
- Prosthetic mechanical valves
- Presence of left ventricular mural thrombus
- Previous thromboembolism (stroke, pulmonary embolism, deep vein thrombosis)
- Poor left ventricular ejection fraction (<28%)
- Acute left ventricular wall aneurysm
- Recent myocardial infarction
- Idiopathic dilated cardiomyopathy
- Infective endocarditis
- Atrial myxoma
- Reduced peak oxygen uptake at maximal exercise

Studies comparing the effect of antithrombotic therapy in heart failure

Study	Treatment (anticoagulation *v* antiplatelet agents *v* placebo)	Stroke incidence every 100 patient years	Thromboembolic incidence every 100 patient years	Mortality
SOLVD	Aspirin *v* placebo	No effect	No effect	Lower with aspirin but benefit of enalapril blunted
SOLVD	Warfarin *v* placebo	No effect	No effect	Lower with warfarin
V-HeFT-I	Aspirin *v* warfarin *v* placebo	0.5 *v* 1.9 *v* 2.0	0.5 *v* 2.9 *v* 2.7	
CONSENSUS	Warfarin *v* placebo	ND	ND	Lower with warfarin
PROMISE	Warfarin *v* placebo	1.9 *v* 2.9	ND	
Fuster et al	Anticoagulants *v* placebo	ND	None *v* 3.5	No effect
Katz et al	Aspirin *v* warfarin *v* placebo	1.1 *v* 7.5 *v* 0.8	ND	Lower with aspirin. No effect with warfarin

ND = no data provided
Full references in Lip GYH, Gibbs CR *Quart J Med Cochrane Reviews* (see Further reading)

nearly 70%. The benefit of warfarin use is not uniform, as for example, the V-HeFT-1 Trial did not show any substantial benefit with warfarin use. Data from the SOLVD and V-HeFT studies were observational, without randomisation or control with respect to oral anticoagulation. In addition, interpretation of these data is potentially confounded—patients who were considered to be at highest risk of thromboembolism may have been treated with warfarin, and this would substantially reduce their long term risk of thromboembolic events.

Balanced against its potential benefits, warfarin increases the risk of intracranial haemorrhage by 0.3%. This complication increases with higher therapeutic targets (for example, INR 3.0-4.5). In heart failure patients with sinus rhythm, assuming a 2% risk of ischaemic stroke, warfarin use must reduce the risk by at least 20% to outweigh the threat of intracranial haemorrhages. Recurrent stroke rates in heart failure patients are also high (about 9% a year), and a reduction of only 10% would outweigh the risk of bleeding complications with warfarin. Secondary prevention of ischaemic strokes by warfarin in patients with heart failure should be considered.

Similarly, not all patients with documented left ventricular thrombi carry a high risk of embolisation. "High risk" patients such as those who have had an extensive acute anterior myocardial infarction or who show a new mural thrombus on cross sectional echocardiography should be considered for anticoagulation. Long term therapy (more than three months) is generally not recommended because of low embolisation rates from chronic left ventricular thrombi (unless the mass is mobile and pedunculated or other high risk factors exist).

Despite the limited evidence, some authorities recommend anticoagulation for patients with idiopathic dilated cardiomyopathy, whereas others using the postmyocardial infarction studies and data from SOLVD and CONSENSUS, advocate the use of warfarin in patients with ischaemic cardiomyopathy. A recent Cochrane systematic review does not recommend warfarin routinely for all heart failure patients in sinus rhythm because of conflicting conclusions from retrospective analyses and case series. More evidence is needed from randomised trials, such as the WATCH study (see later).

Indications for warfarin treatment in chronic heart failure

Strongly recommended
- Atrial fibrillation
- Previous ischaemic strokes
- New left ventricular mural thrombus formation
- Unstable, mobile left ventricular thrombus

Individual consideration to be given
- Idopathic dilated cardiomyopathies
- Poor left ventricular ejection fraction (< 28%)
- Acute left ventricular aneurysm

Not recommended
- Sinus rhythm in absence of other risk factors
- Chronic left ventricular aneurysm
- Presence of chronic organised left ventricular mural thrombus

Aspirin

Use of aspirin in heart failure is controversial, but it is commonly used in patients with chronic heart failure who are in sinus rhythm, because of its general efficacy as an antithrombotic agent in vascular disease. In addition, aspirin in general reduces the incidence of stroke and death in patients with recurrent episodes of cerebral ischaemia. Furthermore, aspirin has been shown to be moderately effective in reducing venous thrombosis and thromboembolism in patients undergoing hip surgery.

Aspirin has been shown to reduce the incidence of myocardial infarction and death in men and women over 50 years, patients with unstable angina and myocardial infarction, and in patients with atherosclerotic cerebrovascular disease, whereas aspirin improves the patency rates of saphenous-vein aortocoronary bypass grafts

Most trials of angiotensin converting enzyme (ACE) inhibitors, bar the SAVE study, have shown possible attenuation of their protective effects in heart failure after myocardial infarction when combined with aspirin. Several reasons have been put forward, not least because aspirin and ACE inhibitors exert effects on the same prostaglandin pathways. Concomitant aspirin can also limit sodium excretion and impair renal function in patients with chronic heart failure. The justification for using aspirin in chronic heart failure is strongest where an ischaemic aetiology is suspected. Indeed, aspirin reduces rates of death, recurrent myocardial infarction, and stroke when used in or shortly after acute myocardial infarction.

Although observational trials have shown a beneficial reduction of stroke with aspirin in patients with chronic heart failure, in general this reduction is less than that seen with warfarin. Two studies have shown aspirin to be more beneficial than warfarin in preventing strokes in heart failure. Both studies however were not randomised and it is possible that participants in the warfarin arm were more seriously ill or at higher risk of systemic embolism (for example, associated atrial fibrillation). Aspirin had no effect on mortality or risk of systemic embolism in patients with documented left ventriuclar thrombi.

A recent Cochrane systematic review concluded that there was conflicting evidence to support the use of antiplatelet drugs to reduce the incidence of thromboembolism in patients with chronic heart failure who are in sinus rhythm. There was also no direct evidence to indicate superior effects from oral anticoagulation, when compared with aspirin, in patients with chronic heart failure. Pending further evidence, aspirin cannot be recommended routinely for all patients with chronic heart failure (in sinus rhythm or atrial fibrillation) with or without left ventricular thrombi for the prevention of stroke and thromboembolism. Current guidelines should be tailored to individual risks and benefits.

Other antiplatelet drugs have not been represented in previous trials. Clopidogrel, which inhibits platelet function by inhibiting adenosine-induced platelet aggregation does not inhibit cyclo-oxygenase. Thus, it should not attenuate the beneficial actions of ACE inhibitors in the manner of aspirin. The large, ongoing randomised controlled WATCH study using clopidogrel, aspirin, and warfarin seeks to identify the optimal antithrombotic agent with the best risk:benefit ratio in the prevention of stroke and thromboembolism in heart failure.

Acute heart failure

Patients with acute and decompensated heart failure not only have high levels of circulating catecholamines that may further activate the clotting cascade, but also tend to be less ambulatory and can be confined to bed or chair. The risk of venous thromboembolism is therefore increased and in fact, thromboembolic complications add to the burden of prolonged hospitalisation and mortality in these patients. Giving subcutaneous injections of unfractionated or low molecular weight heparin as prophylaxis can reduce the risk of venous thromboembolism, but trial evidence in chronic heart failure is lacking. No recent clinical trials show that warfarin or aspirin are effective in the primary prevention of venous thromboembolism or systemic complications in patients with acute heart failure who are in sinus rhythm.

Further reading

- Al-Khadra AS, Salem DN, Rand WM, Udelson JE, Smith JJ, Konstam MA. Warfarin anticoagulation and survival: a cohort analysis from the Studies of Left Ventricular Dysfunction. *J Am Coll Cardiol* 1998; 31:749-53
- Cleland JG. Anticoagulant and antiplatelet therapy in heart failure. *Curr Opin Cardiol* 1997 12:276-87
- Dries DL, Rosenberg YD, Waclawiw MA, Domanski MJ. Ejection fraction and risk of thromboembolic events in patients with systolic dysfunction and sinus rhythm: evidence for gender differences in the studies of left ventricular dysfunction trials. *J Am Coll Cardiol* 1997; 29:1074-80
- Dunkman WB, Johnson GR, Carson PE, Bhat G, Farrell L, Cohn JN, for the V-HeFT VA Cooperative Studies Group. Incidence of thromboembolic events in congestive heart failure. *Circulation* 1993;87:VI194-101
- Lip GYH, Gibbs CR. Does heart failure confer a hypercoagulable state? Virchow's triad revisited. *J Am Coll Cardiol* 1999;5:1424-6
- Lip GYH, Gibbs CR. Anticoagulation for heart failure in sinus rhythm: a systematic Cochrane review. *Quart J Med* 2002;95:451-9
- Lip GYH, Gibbs CR. Antiplatelet agents versus control or anticoagulation for heart failure in sinus rhythm: a Cochrane systematic review. *Quart J Med* 2002;95:461-8
- Lip GYH. Intracardiac thrombus formation in cardiac impairment: the role of anticoagulant therapy. *Postgrad Med J* 1996;72:731-8
- Loh E, St. John Sutton M, Wun CC, et al. Ventricular dysfunction and the risk of stroke after myocardial infarction. *New Engl J Med* 1997;336:251-257
- Uretsky BF, Thygesen K, Armstrong PW, Cleland JG, Horowitz JD, Massie BM, et al. Acute coronary findings at autopsy in heart failure patients with sudden death: results from the assessment of treatment with lisinopril and survival (ATLAS) trial. *Circulation* 2000;102:611-6

We thank Professor JG Cleland (University of Hull, United Kingdom), for helpful comments during the preparation of this review

The figure showing the two year adjusted incidence of stroke in the Framingham study is adapted from Wolfe CD et al. *Stroke* 1991;22:983

14 Antithrombotic therapy in special circumstances. I–pregnancy and cancer

Bernd Jilma, Sridhar Kamath, Gregory Y H Lip

Antithrombotic therapy during pregnancy

Pregnancy predisposes to venous thromboembolism for several reasons. These include a change in the balance between procoagulant and anticoagulant factors in the blood. Any conditions that predispose a woman to thromboembolism when she is not pregnant will also predispose her to thromboembolism when she is pregnant.

Generally, antithrombotic therapy started in a non-pregnant patient for a particular disorder needs to be continued during the pregnancy and in the puerperium. The use and type of antithrombotic therapy depends on the risk:benefit ratio, taking into consideration the potential harm to the mother and the fetus.

The potential risks of antithrombotic therapy during pregnancy can be divided into maternal and fetal risks, and include teratogenicity and bleeding. Unfractionated heparin and low molecular weight heparins do not cross the placenta and are probably safe for the fetus, although bleeding at the uteroplacental junction is possible. Nevertheless, data are sparse for low molecular weight heparin, with no reliable comparative trials or convincing dose assessment.

In contrast to heparin, coumarin derivatives cross the placenta and can cause both bleeding in the fetus and teratogenicity. Coumarin derivatives can cause an embryopathy (which consists of nasal hypoplasia or stippled epiphyses or both) after in utero exposure during the first trimester of pregnancy. In addition, central nervous system abnormalities can occur after exposure to such drugs during any trimester. The main risk of embryopathy occurs if coumarin derivatives are taken between six weeks and 12 weeks of gestation. At the time of delivery, the anticoagulant effect in the fetus can lead to bleeding in the neonate.

Heparin and low molecular weight heparins are not secreted into breast milk and can probably be given safely to nursing mothers. High dose aspirin should be avoided, as it could (theoretically) impair platelet function and produce hypoprothrombinaemia in the infant, if neonatal vitamin K stores are low, as well as cause Reye's syndrome. Warfarin does not induce an anticoagulant effect in an infant who is breast fed and therefore could be used safely in the postpartum period; thus, patients who are receiving long term heparin treatment could be switched over to warfarin post partum if and when considered appropriate. With regard to other agents, phenindione should be avoided, and acenocoumarol requires prophylactic vitamin K for the infant.

Venous thrombosis and pulmonary embolism

Antithrombotic prophylaxis for the prevention of venous thromboembolic disorders in pregnancy is indicated when a patient has experienced a previous thromboembolic episode or is considered to be at particularly high risk because of a predisposing condition.

Unfractionated heparin 5000 IU twice daily is generally adequate in non-pregnant women. Heparin requirements can be highly variable in pregnancy. A once daily dose of low

Disorders during pregnancy for which antithrombotic therapy is commonly considered

- Prophylaxis and treatment of venous thromboembolism
- Prophylaxis in patients with valvar disease (for example, mitral stenosis)
- Prophylaxis in patients with mechanical prosthetic valves
- Antiphospholipid syndrome
- Prophylaxis against pregnancy induced hypertension and intrauterine growth retardation

Potential risks of antithrombotic therapy during pregnancy

Maternal disadvantages and risks

Unfractionated heparin
- Haemorrhage (uteroplacental, especially during labour)
- Heparin induced thrombocytopenia
- Osteoporosis
- Regular monitoring

Low molecular weight heparin
- Bleeding risk, especially during labour

Warfarin
- Bleeding
- Regular monitoring

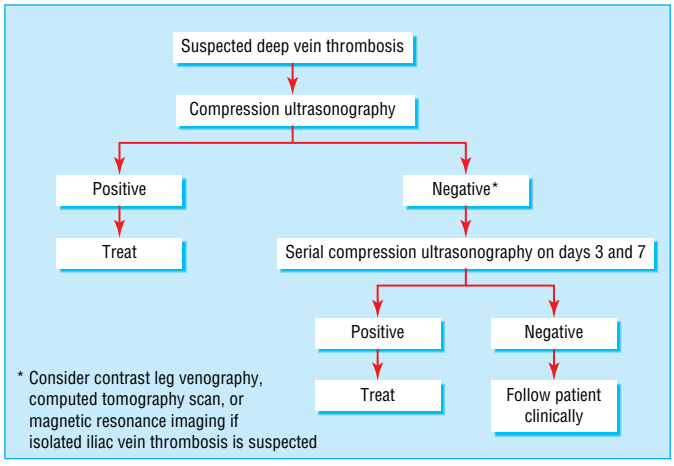
Risk to the fetus or child

Heparin
- Seems to be safe

Low molecular weight heparin
- Seems to be safe

Warfarin
- Embryopathy, especially if mother is exposed between 6 and 12 weeks
- Central nervous system malformations during any time of the gestation

Low dose aspirin
- Potential risk of birth defects and bleeding risk in the first trimester
- Safe in second and third trimester



Diagnosis of suspected deep vein thrombosis in pregnancy

molecular weight heparin is a useful alternative to unfractionated heparin and has been shown to be safe and effective in pregnancy.

Patients who develop thromboembolism during pregnancy could be treated initially with at least five days of intravenous heparin treatment, followed by a twice daily subcutaneous dose of unfractionated heparin. The dose is adjusted by maintaining the activated partial thromboplastin time (APTT) within the therapeutic levels. Heparin could temporarily be stopped immediately before delivery and then resumed in the postpartum period to minimise the risk of haemorrhage during labour. The duration of antithrombotic therapy in the postpartum period should be maintained for a minimum of three months, possibly up to six months.

Patients with prosthetic heart valves

The precise safety of warfarin during pregnancy continues to be debated, but it is probably appropriate to withhold warfarin between six and 12 weeks of gestation and for the latter half of the third trimester, because of the risk of causing embryopathy and postpartum haemorrhage respectively. Based on this, the recommended options for use of antithrombotic therapy in patients with a mechanical heart valve during pregnancy include:

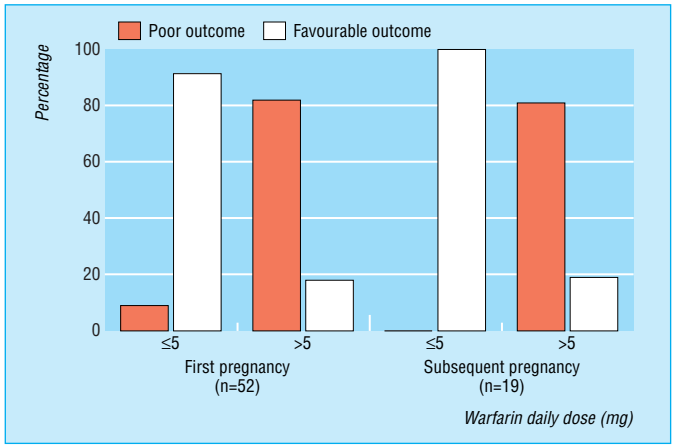
- Adjusted dose of unfractionated heparin twice daily to maintain the APTT within therapeutic range, or low molecular weight heparin throughout the pregnancy
- Warfarin throughout the pregnancy, except for the first trimester (either for the entire trimester or between six and 12 weeks) and for the latter half of the third trimester (when warfarin should be replaced by unfractionated or low molecular weight heparin).

The risks and benefits of this approach should be explained to patients, who should be allowed to make an informed choice. There are real concerns over the incidence of abortions and fetal malformations in patients treated with warfarin in the first trimester. Concerns over long term heparin treatment in pregnant women include heparin induced thrombocytopenia and osteoporosis.

Prepregnancy counselling is vital for patients who are receiving long term warfarin treatment, and a cardiologist and obstetrician should explain the risks to patients. Patients who are established on long term warfarin treatment and plan to become pregnant could then take twice daily heparin before getting pregnant. Alternatively, and assuming that the risk of warfarin to the fetus in the first six weeks of gestation is not worrisome, they could continue taking warfarin and have frequent checks to see if they are pregnant. If they are, they should immediately switch over to heparin. Again, close liaison between obstetrician, midwife, general practitioner, cardiologist, and neonatologist is vital.

Antiphospholipid syndrome

Antiphospholipid syndrome predisposes a pregnant woman to thromboembolism and pregnancy losses. Abortion in a previous pregnancy predisposes to further abortions or stillbirths in subsequent pregnancies. A combination of aspirin and heparin to prolong the APTT to within the therapeutic range (APTT ratio 2.0-3.0) thoughout the pregnancy would substantially decrease pregnancy losses and other complications. However, one recent trial suggested that low dose aspirin (75 mg) may suffice (see box), but small numbers merit some caution pending further data.



Distribution of warfarin dose and poor outcome according to order of pregnancy. The risk of pregnancy complication in patients treated with warfarin is higher when the mean daily dose exceeds 5 mg (P<0.001)

Antiphospholipid syndrome in pregnancy: a randomised controlled trial of aspirin *v* aspirin plus heparin

	Low dose aspirin (n=47)	Low dose aspirin plus heparin (n=51)
Live birth rate (%)	72	78
Mean (SD) birth weight (g):	3221 (781)	3127 (657)
Range	890-5300	718-4319
Gestation at delivery (%):		
<30 weeks	1	1
30-36 weeks	3	1
>36 weeks	30	38
Embryo loss (%)	9	3
Fetal loss (%)	4	8

Antithrombotic therapy in antiphospholipid syndrome

Scenario	Management
History of pregnancy loss	Aspirin plus heparin (APTT in therapeutic range)
History of thromboembolism but no pregnancy loss	Heparin alone (APTT in therapeutic range)
No history of adverse events	Heparin alone 5000 IU twice daily; close observation

Pre-eclampsia and intrauterine growth retardation

On the basis of small, retrospective studies, low dose aspirin (< 150 mg daily) was thought to be useful as prophylaxis in patients with a history of pre-eclampsia and intrauterine growth retardation in preventing similar adverse events during the current pregnancy. However, a large (nearly 10 000 women) randomised controlled trial (CLASP) of aspirin 60 mg compared with placebo, reported that, although aspirin was associated with a 12% reduction in the incidence of pre-eclampsia, this was not significant nor was there any substantial impact on intrauterine growth retardation, stillbirth, or neonatal death. Thus, routine use of low dose aspirin is not recommended. However, some experts recommend its use in patients who are liable to develop early onset (before 32 weeks) pre-eclampsia or in high risk groups for pre-eclampsia, such as women with type 1 diabetes, chronic hypertension, multiple pregnancies, or previous pre-eclampsia. However, the safety of higher doses of aspirin and aspirin ingestion during the first trimester remains uncertain.

Study	Aspirin	Placebo	Peto odds ratio (95% CI fixed)	Weight (%)	Peto odds ratio (95% CI fixed)
McParland, 1990	1/48	10/52		18.5	0.18 (0.05 to 0.61)
Morris, 1996	4/52	7/50		18.6	0.52 (0.15 to 1.82)
Bower, 1996	9/31	12/29		26.0	0.59 (0.20 to 1.68)
Zimmerman, 1997	4/13	2/13		9.0	2.30 (0.38 to 13.77)
Harrington, 2000	7/107	9/103		27.9	0.73 (0.27 to 2.03)
Total (95% CI)	25/251	40/247		100.0	0.55 (0.32 to 0.95)
Test for hetergeneity $\chi 2$=5.97, df=4, P=0.2					
Test for overall effect z=-2.16, P=0.03					

Effect of aspirin in preventing pre-eclampsia: meta-analysis of randomised trials showing numbers of cases of pre-eclampsia

Antithrombotic therapy in cancer

Venous thromboembolism is a frequent complication in patients with cancer, and it is a common clinical problem. It can even precede the diagnosis of cancer by months or years. Patients with cancer are nearly twice as likely to die from pulmonary embolism in hospital as those with benign disease, and about 60% of these deaths occur prematurely. Thromboembolism seems to be particularly predominant in patients with mucinous carcinoma of the pancreas, lung, or gastrointestinal tract. This may be because cancer can be associated with raised levels of procoagulants such as fibrinogen, von Willebrand factor, and tissue factor, as well as excess platelet activity. Raised levels of plasminogen activator inhibitor are often present, and this will impair fibrinolysis. Therapeutic interventions in patients with cancer, such as surgery, standard chemotherapy, or hormone based treatment (such as oestrogens for prostatic cancer), further increase the risk for thrombosis. One reason for this may be that certain types of chemotherapy impair the natural anticoagulant properties of the endothelium, thus promoting a procoagulant state. Unfortunately, no standardised protocols exist for the management of patients with cancer and the approaches vary.

Virchow's triad* in cancer

Abnormal blood flow
- Increased viscosity and turbulence
- Increased stasis from immobility

Abnormal blood constituents
- Increased platelet activation and aggregation
- Increased procoagulant factors
- Decreased anticoagulant and fibrinolytic factors

Abnormal blood vessel wall
- Damaged or dysfunctional endothelium
- Loss of anticoagulant nature
- Possibly angiogenesis

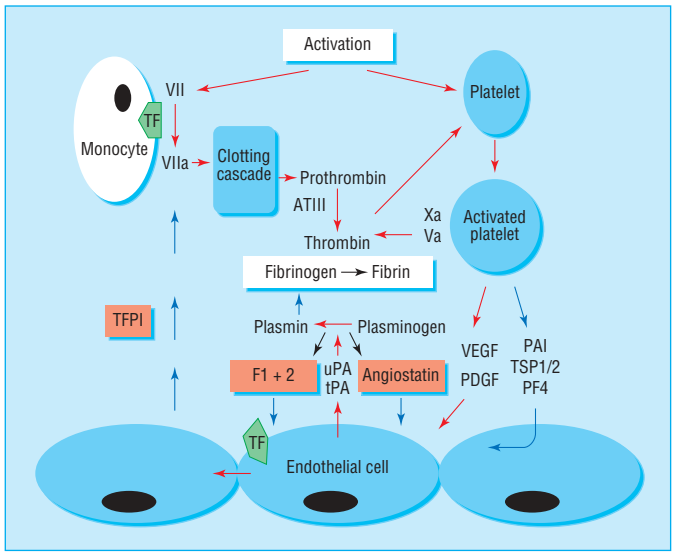
*For thrombogenesis (thrombus formation) there needs to be a triad of abnormalities (abnormal blood flow, abnormal blood constituents, and abnormal blood vessel wall)

Risk factors for thromboembolism in patients with cancer
- Prolonged immobility
- Chemotherapy
- Surgical procedures
- Indwelling vascular catheters

Primary prophylaxis

In patients with cancer who are confined to bed or having low risk surgical procedures a low dose of unfractionated heparin or low molecular weight heparin is administered subcutaneously, along with physical measures, as primary prophylaxis to reduce thromboembolic risk. Patients having major abdominal or pelvic surgery for cancer are recommended to receive adjusted dose heparin, low molecular weight heparin, or oral anticoagulants (therapeutic international normalised ratio (INR) 2.0-3.0) similar to those for major orthopaedic surgery.

A low dose warfarin regimen is recommended for patients receiving chemotherapy or those with indwelling venous catheters to decrease the incidence of thromboembolism. For example, one double blind randomised study of patients with metastatic breast cancer receiving chemotherapy showed that a very low dose (1 mg/day) of warfarin for six weeks followed by a dose to maintain the INR at 1.3-1.9 was effective. Low dose low molecular weight heparin (for example, daltaparin 2500 IU/day) is an alternative for patients with indwelling venous catheters.



Overview of coagulation, fibrinolysis, and angiogenesis in cancer. The activation of platelets leads to their swelling and the release of angiogenic factors. These affect the vascular endothelia of healing and tumour tissues. The blue arrows facilitate angiogenesis and the red arrows are inhibitory (ATIII=antithrombin III, F1 + 2=prothrombin fragments, PAI=plasminogen activator inhibitor, PDGF=platelet derived growth factor, PF4=platelet factor 4, TF=tissue factor, TFPI=tissue factor pathway inhibitor, TSP1/2=thrombospondin 1 and 2, tPA=tissue type plasminogen activator, uPA=urokinase type plasminogen activator, VEGF=vascular endothelial growth factor)

Treatment and secondary prevention

Patients with cancer who develop a thromboembolism should be treated in a similar manner to patients without cancer. An initial period of therapeutic unfractionated heparin or low molecular weight heparin which is overlapped and followed by warfarin for a minimum of three months is recommended. Anticoagulation should be continued in patients who have active disease or who receive chemotherapy while these risk factors last. The dose should maintain an INR of between 2.0 and 3.0.

Risk of haemorrhage

Patients with cancer who are receiving antithrombotic therapy are thought to be at higher risk of bleeding than patients without cancer. This assumption has been disputed, however, in light of the evidence from some studies in which the risk of major bleeding did not differ greatly between the two groups of patients. For practical purposes, the recommended therapeutic levels of anticoagulation remain the same (for example, if warfarin, then INR 2.0-3.0) as long as patients are educated about the risks and the anticoagulation levels are strictly monitored. The propensity for chemotherapy to be given in cycles or boluses, followed by periods free of chemotherapy, seems likely to frustrate attempts to maintain the INR within its target range.

Definite conclusions cannot be drawn about the safety of antithrombotic therapy in patients with primary or secondary brain malignancy. Some small studies report that it is probably safe to give these patients anticoagulants. However, definite decisions about anticoagulation in such patients have to be individualised and carefully considered. Anticoagulation should probably be avoided in patients with brain metastasis because of the chances of renal cell carcinoma or melanoma, as these tumours are highly vascular.

Recurrent venous thromboembolism

Patients with cancer are at a higher risk than non-cancer patients of recurrence of thromboembolism despite adequate anticoagulation. Again, no strict evidence based guidelines exist for the management of these patients. The recommended options include maintenance of a higher level of anticoagulation (INR 3.0 to 4.5), substitution with adjusted dose heparin or low molecular weight heparin (some evidence suggests heparin is probably better in this situation), and placement of inferior venacaval filters with or without anticoagulation.

The figure showing diagnosis of deep vein thrombosis is adapted from Chan W-S et al, *Thromb Res* 2002;107:85-91. The table showing the results of aspirin *v* aspirin plus heparin in treating antiphospholipid syndrome in pregnancy is adapted from Farquharson RG et al, *Obstet Gynecol* 2002;100:408-13. The table showing Virchow's triad in cancer is adapted from Lip GYH et al, *Lancet Oncol* 2002;3:27-34. The histogram showing distribution of warfarin dose and poor outcome according to order of pregnancy is adapted from Cotrufo M et al, *Obstet Gynecol* 2002;99:35-40. The meta-analysis showing the effect of aspirin in preventing pre-eclampsia is adapted from Coomarasamy A et al, *Obstet Gynecol* 2001;98:861-6. The figure showing the overview of coagulation, fibrinolysis, and angiogenesis in cancer is adapted from Nash G et al, *Lancet Oncol* 2001;2:608-13.

Concerns about antithrombotic therapy in cancer

- Recurrent venous thromboembolism
- Increased tendency for minor and major bleeds
- Inconsistency in therapeutic anticoagulant levels
- Procoagulant effects of chemotherapy (for example, endothelial cell dysfuntion)

Further reading

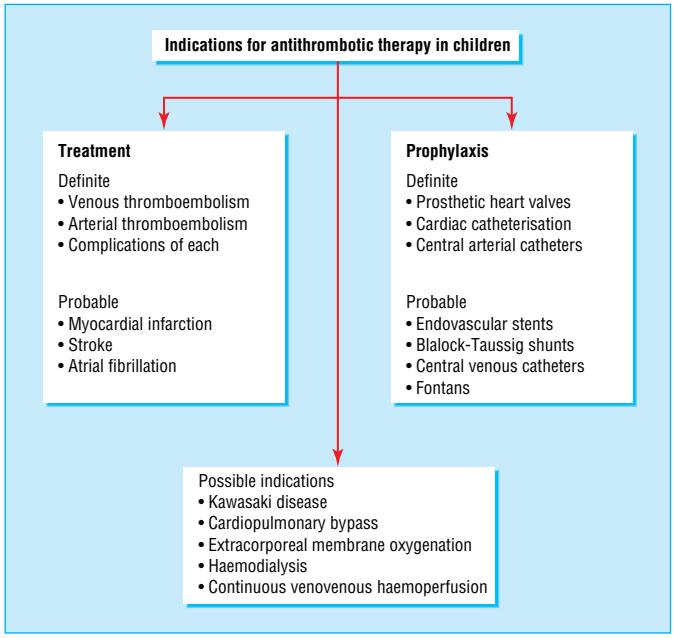
- Barbour LA. Current concepts of anticoagulation therapy in pregnancy. *Obstet Gynecol Clin North Am* 1997;24:499-521
- CLASP (Collaborative Low-dose Aspirin Study in Pregnancy) Collaborative Group. CLASP: a randomised trial of low-dose aspirin for the prevention and treatment of pre-eclampsia among 9364 pregnant women. *Lancet* 1994;343:619-29
- Cumming AM, Shiach CR. The investigation and management of inherited thrombophilia. *Clin Lab Haem* 1999;21:77-92
- Coomarasamy A, Papaioannou S, Gee H, Khan KS. Aspirin for the prevention of preeclampsia in women with abnormal uterine artery Doppler: a meta-analysis. *Obstet Gynecol* 2001;98:861-6
- Cotrufo M, De Feo M, De Santo LS, Romano G, Della Corte A, Renzilli A, et al. Risk of warfarin during pregnancy with mechanical valve prostheses. *Obstet Gynecol* 2002;99:35-40
- Ginsberg JS, Greer I, Hirsh J. Use of antithrombotic agents during pregnancy. *Chest* 2001:119;S122-31
- Farquarson RG, Quenby S, Greaves M. Antiphospholipid syndrome in pregnancy: a randomized, controlled trial of treatment. *Obstet Gynecol* 2002;100:408-13
- Letai A, Kuter DJ. Cancer, coagulation, and anticoagulation. *Oncologist* 1999;4:443-9
- Levine M, Hirsh J, Gent M, Arnold A, Warr D, Falanga A, et al. Double-blind randomised trial of a very-low-dose warfarin for prevention of thromboembolism in stage IV breast cancer. *Lancet* 1994;343:886-9
- Lip GYH, Chin BSP, Blann AD. Cancer and the prothrombotic state. *Lancet Oncol* 2002;3:27-34
- Prandoni P. Antithrombotic strategies in patients with cancer. *Thromb Haemost* 1997;78:141-4
- Sanson BJ, Lensing AW, Prins MH, Ginsberg JS, Barkagan ZS, Lavanne-Pardlonge E, et al. Safety of low-molecular-weight heparin in pregnancy: a systematic review. *Thromb Haemost* 1999;81:668-72
- Chan W-S, Ginsberg JS. Diagnosis of deep vein thrombosis and pulmonary embolism in pregnancy. *Thromb Res* 2002;107:85-91

15 Antithrombotic therapy in special circumstances. II–children, thrombophilia, and miscellaneous conditions

Bernd Jilma, Sridhar Kamath, Gregory Y H Lip

Treatments for children

Most of the recommendations on antithrombotic therapy in children are based on the extrapolation of results from randomised studies of adults or from small cross sectional, and mainly retrospective, clinical studies of children. Although antithrombotic therapy in children usually follows the same indications as in adults, the distribution of diseases requiring antithrombotic therapy differs in the paediatric population. For example, some predisposing factors for thromboembolism are encountered only in paediatric populations. Most of the indications for antithrombotic therapy in children arise because of an underlying medical disorder or an intervention for the management of the disorder. Management of antithrombotic therapy in children differs from that in adults because of ongoing changes in physiology that may alter the thrombotic process and potentially influence the response of the body to antithrombotic therapy.



Indications for antithrombotic therapy in children

Drug treatments

Antiplatelet treatment

Aspirin, dipyramidole, and indomethacin are probably the most used antiplatelet treatments among children. Low doses of aspirin (antiplatelet doses) usually have minimal side effects in children, but in general aspirin should not be prescribed to children aged <16 years unless there are compelling clinical indications. The particular concerns about Reye's syndrome usually seem to be related to higher doses of aspirin (>40 mg/kg).

Heparin

Heparin is probably the most commonly used antithrombotic drug in children. Varying concentrations of antithrombin in the body during different developmental stages mean that the therapeutic concentration of heparin in children has to be maintained by regular checks of the activated partial thromboplastin time (APTT) or anti-Xa concentrations. The recommended therapeutic level of APTT is the one which corresponds to a heparin concentration of 0.2-0.4 U/ml or an anti-Xa concentration of 0.3-0.7 U/ml.

In children, the advantages of low molecular weight heparin over unfractionated heparin are similar to those in adults. In addition, low molecular weight heparin may be preferred for children with difficult venous access because regular blood checks to monitor the therapeutic levels are not mandatory. The recommended therapeutic dose of a low molecular weight heparin is the one that reflects the plasma anti-Xa concentrations of 0.5-1.0 U/ml four to six hours after injection.

Oral anticoagulants

Certain problems are associated with the use of oral anticoagulants in children. Sensitivity to oral anticoagulants changes during different phases of life, especially during infancy, because of varying concentrations of vitamin K and vitamin K dependent proteins in the body. Neonates (during the first month

Adjusting low molecular weight heparin in children

Anti-Xa level (U/ml)	Hold next dose?	Dose change?	Repeat anti-Xa measurement
<0.35	No	Increase by 25%	4 hours after next dose
0.35-0.49	No	Increase by 10%	4 hours after next dose
0.5-1.0	No	No	Next day, then 1 week later, and monthly thereafter while receiving reviparin-Na treatment (4 hours after morning dose)
1.1-1.5	No	Decrease by 20%	Before next dose
1.6-2.0	3 hours	Decrease by 30%	Before next dose then 4 hours after next dose
>2.0	Until anti-Xa 0.5 U/ml	Decrease by 40%	Before next dose, then every 12 hours until anti-Xa level <0.5 U/ml

of life) are especially sensitive because of their relative deficiency of vitamin K, and therefore warfarin should be avoided in such patients if possible. However, formula fed infants are resistant to oral anticoagulants because of a high concentration of vitamin K in their diet. In general, young children need more oral anticoagulation for each kilogram of body weight than older children and adults. Poor venous access (for international normalised ratio (INR) checks) and non-compliance are added problems of anticoagulation in children.

Recommended therapeutic ranges and duration of anticoagulation for a variety of disorders in children are usually similar to those for adults.

Thrombolytic treatment

Thrombolytic treatment is used primarily for maintaining catheter patency and in the management of thromboembolism that threatens the viability of the affected organ. Thrombolytic drugs are used locally or systemically and their concentration can be monitored with plasma fibrinogen levels or total clotting time. Decreased plasma plasminogen levels in newborns may reduce the thrombolytic actions of the drugs. Thrombolytic drugs pose similar risks to children as to adults.

Venous thromboembolism

Venous thromboembolism in children usually occurs secondary to an underlying disorder, such as in the upper arm secondary to a central venous line being inserted. Such lines are usually placed for intensive care management and treatment of cancer. The patency of these lines is traditionally maintained through therapeutic local instillation of urokinase for blocked lines or prophylactic intermittent boluses of heparin (which have doubtful efficacy).

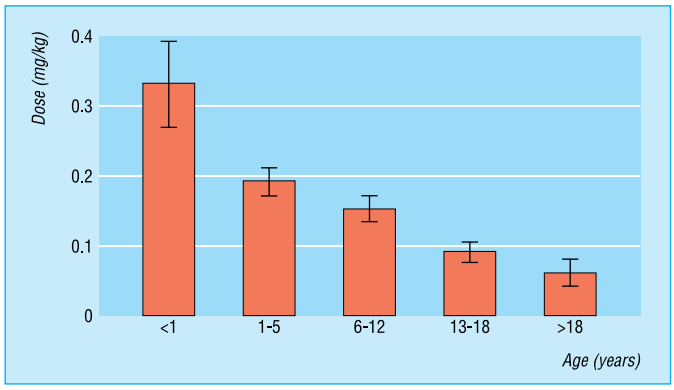
Established venous thromboembolism requires removal of the predisposing factor and anticoagulation similar to that in adults (standard heparin for five days followed by maintenance with oral anticoagulation for at least three months). Oral anticoagulation can be started on the same day as heparin. Low molecular weight heparin is a useful option for maintaining anti-Xa level of 0.5-1.0 U/ml. Patients with a first recurrence of venous thromboembolism or with an initial episode with continuing risk factors either could be closely monitored for any early signs of thromboembolism or should be given anticoagulant drugs prophylactically after the period of initial therapeutic anticoagulation for the episode. Patients with a second recurrence of venous thromboembolism or with a first recurrence with continuing risk factors should be given anticoagulants for life, as in adults.

Arterial thromboembolism

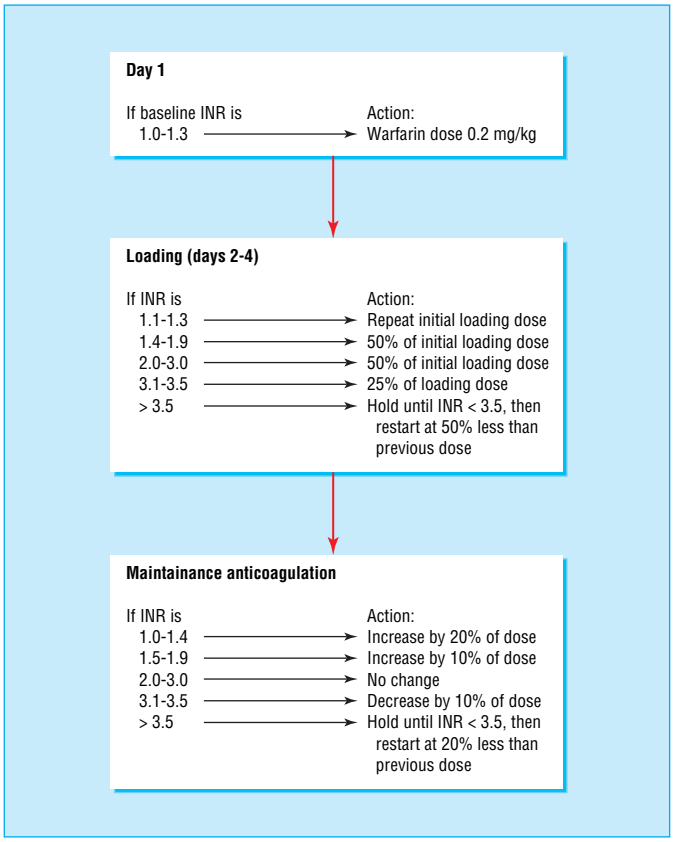
The usual predisposing factors include placement of central and peripheral arterial catheters for cardiac catheterisation and intensive care settings. A bolus of heparin (50-150 U/kg) at the time of arterial puncture and continuous low dose heparin infusion are common methods for cardiac and umbilical artery catheterisation, respectively.

Prosthetic heart valves

Oral anticoagulation is needed in children with mechanical heart valves. An INR of 2.5-3.5 is recommended as the target range. Patients who are predisposed to high risk of thromboembolism despite anticoagulation treatment and those with thromboembolism while taking warfarin could benefit from the addition of antiplatelet drugs, such as aspirin (6-20 mg/kg/day) or dipyridamole (2-5 mg/kg/day), to oral anticoagulation.



Effect of age on dose of warfarin needed to sustain an international normalised ratio (INR) of 2.0-3.0 in 262 children. Younger children required significantly more warfarin than older children (P<0.001)



Protocol for oral anticoagulation treatment to maintain an INR ratio of 2.0-3.0 for children

Commonly used drugs in children that affect INR values

Drug	Usual effect on INR
Amiodarone	Increase
Aspirin	Increase or no change
Amoxicillin	Slight increase
Cefaclor	Increase
Carbamazepine	Decrease
Phenytoin	Decrease
Phenobarbital	Decrease
Cloxacillin	Increase
Prednisone	Increase
Co-trimoxazole	Increase
Ranitidine	Increase

Other cardiac disorders
No universally accepted guidelines or randomised trials exist for the antithrombotic therapy in patients undergoing operations where there is risk of thromboembolism (such as Blalock-Taussig shunts, Fontan operations, and endovascular stents). A variety of antithrombotic regimens have been used after these operations, including intraoperative heparin only and intraoperative heparin followed by oral anticoagulation or aspirin.

Hereditary prothrombotic states
Deficiencies of protein C, protein S, or antithrombin III and factor V Leiden mutation can lead to thromboembolism especially in the presence of a secondary risk factor. Homozygous deficiency of these proteins could lead to fatal purpura fulminans in newborns, which is treated immediately by rapid replacement of these factors with fresh frozen plasma or protein concentrates. This is followed by careful initiation of lifelong oral anticoagulation to maintain the INR at higher levels of 3.0-4.5. Heterozygous patients could be given prophylactic antithrombotic therapy during exposure to secondary risk factors or be followed up with close observation.

Antithrombotic therapy in thrombophilia and miscellaneous conditions

A detailed discussion of management of thrombophilic disorders is beyond the scope of this article. The guidelines on the management of these disorders are based on small and non-controlled series of patients because of the paucity of randomised trials (as reviewed by the Haemostasis and Thrombosis Task Force in 2001).

Inherited thrombophilic disorders are genetically determined, and most of the affected patients are heterozygotes. Homozygotes are extremely rare. Antithrombin III, protein C, and protein S are produced in the liver and act by inactivating coagulation factors. Deficiency of these proteins could lead to uncontrolled activation of the coagulation cascade and therefore thromboembolism.

Activated protein C resistance is the commonest inherited thrombophilic disorder and accounts for 20-50% of cases. Antithrombin III deficiency is the rarest of the mentioned inherited thrombophilic disorders but carries the highest thrombogenic risk. High plasma concentration of homocysteine is linked to genetic enzyme deficiencies and low plasma concentrations of folate and vitamin B-6, and an investigation of vitamin B-12 metabolism is warranted.

Though thrombophilic disorders predispose patients to thromboembolism, the routine use of anticoagulation for primary prophylaxis entails greater risks than benefit (except probably in homozygotes). Therefore primary prophylaxis is warranted only in the presence of a second risk factor, and for as long as the risk factor lasts. Common predisposing factors that require prophylaxis include surgery, immobilisation, pregnancy and the puerperium, and oral contraception.

Special caution is needed when giving anticoagulation to patients with protein C deficiency. Because protein C is a vitamin K dependent factor, the administration of warfarin could lead to sudden decrease in protein C before any noticeable decrease in coagulation factors. This could cause enhanced thrombosis and diffuse skin necrosis. This adverse response can be avoided by gradual initiation of oral anticoagulation with low doses of warfarin, preferably overlapped by adequate heparinisation. In cases of severe deficiency, replacement of protein C is indicated before starting warfarin.

Common thrombophilic disorders

Inherited
- Antithrombin III deficiency
- Protein C deficiency
- Protein S deficiency
- Activated protein C resistance (factor V Leiden mutation)
- Inherited hyperhomocysteinaemia
- Raised factor VIII levels
- Prothrombin gene G20210 A variant

Acquired
- Antiphospholipid syndrome
- Acquired hyperhomocysteinaemia

Guidelines for antithrombotic therapy in inherited thrombophilia*

Indication	Treatment
Primary prophylaxis	
Any surgery	Unfractionated heparin subcutaneously 5000 IU three times daily
Malignancy or orthopaedic surgery	Unfractionated heparin subcutaneously 5000 IU three times daily, possibly with replacement of deficient factors
Pregnancy	Unfractionated heparin subcutaneously 5000 IU three times daily
Pregnancy in antithrombin III deficiency	Therapeutic dose of unfractionated heparin to prolong APTT or dose adjusted warfarin (INR 2.0-3.0), except during first trimester and latter part of third trimester, when unfractionated heparin is used
Puerperium (for 4-6 weeks)	Unfractionated heparin subcutaneously 5000 IU three times daily or dose adjusted warfarin (INR 2.0-3.0)
Secondary prophylaxis	
First episode of thrombosis	Dose adjusted warfarin for 6 months
First episode of life threatening thrombosis, multiple deficiencies, continuing predisposing factor	Lifelong oral anticoagulation
Recurrent thrombosis	Lifelong oral anticoagulation
Treatment of established thrombosis	
Treatment of acute thrombosis	Unfractionated heparin to prolong APTT followed by oral anticoagulation treatment, possibly with replacement of the deficient factors

*For full guidelines see Haemostasis and Thrombosis Task Force, *Br J Haematol* 2001;114:512-28
Low molecular weight heparins are increasingly used as alternatives to unfractionated heparin

Little evidence exists to support the use of antithrombotic agents in hyperhomocysteinaemia. Although replacement of folic acid and vitamin B-6 has been shown to reduce plasma homocysteine levels, no study has found reduction in thromboembolic events with this intervention.

Antiphospholipid syndrome

The long term prognosis for this syndrome is influenced by the risk of recurrent thrombosis. As with other thrombophilic disorders, primary prophylaxis is not indicated in the absence of other risk factors. A patient with one episode of thrombosis is at considerable risk of further thrombosis and should be given lifelong anticoagulation with warfarin as secondary prophylaxis. The target INR should be 2.0-3.0 (although some authorities advocate a higher INR level (≥3.0)). Patients with this syndrome may be relatively resistant to warfarin and so will need high doses. However, some authorities believe that the antiphospholipid antibodies interfere with the generation of the INR and lead to spurious results. Consequently, other routes to monitoring anticoagulation may be needed.

Low molecular weight heparin is used increasingly in patients with various thrombophilia and seems to be safe and reliable.

Kawasaki disease

Aspirin continues to be used in Kawasaki disease despite a lack of unequivocal evidence from randomised trials of its benefit in reducing coronary artery aneurysm or thrombosis. Aspirin is used in anti-inflammatory doses (50-100 mg/kg/day) during the acute stage of the disease, followed by antiplatelet doses (1-5 mg/kg/day) for seven weeks or longer.

The figure showing effect of age on dose of warfarin in 262 children is adapted from Streif W et al, *Curr Opin Pediatr* 1999;11:56-64. The diagram of protocol for oral anticoagulation for children and the tables showing adjustment of low molecular weight heparin in children and commonly used drugs in children are adapted from Monagle P et al, *Chest* 2001;119:S344-70. The guidelines for antithrombotic therapy in inherited thrombophilia are adapted from the Haemostasis and Thrombosis Task Force, *Br J Haemotol* 2001;114:512-28. The box of recommendations from the College of Amercian Pathologists consensus conference on diagnostic issues in thrombophilia is adapted from Olson JD, *Arch Pathol Lab Med* 2002;126:1277-80.

Recommendations from the College of American Pathologists consensus conference XXXVI: diagnostic issues in thrombophilia

- Patients, and especially asymptomatic family members, should provide informed consent before thrombophilia testing is performed
- Individuals testing positive for a thrombophilia need counselling on:
  - *Risks of thrombosis* to themselves and their family members
  - *Importance of early recognition* of the signs and symptoms of venous thromboembolism that would require immediate medical attention
  - *Risks and benefits* of antithrombotic prophylaxis in situations in which their risk of thrombosis is increased, such as surgery or pregnancy
- Laboratory testing for other inherited and acquired thrombophilias should be considered even after the identification of a known thrombophilia because more than one thrombophilia could coexist, compounding the risk for thrombosis in many cases
- When available, World Health Organization (WHO) standards, or standards that can be linked to the WHO standard, should be used to calibrate funtional and antigenic assays
- Effect of age and sex should always be taken into consideration when interpreting the results of antigenic and functional assays
- Before concluding that a patient has an inherited thrombophilia, diagnostic assays for function or antigen should be repeated after excluding acquired aetiologies of the defect

Further reading

- Cumming AM, Shiach CR. The investigation and management of inherited thrombophilia. *Clin Lab Haem* 1999;21:77-92
- Monagle P, Michelson AD, Bovill E, Andrew M. Antithrombotic therapy in children. *Chest* 2001;119:S344-70
- Streif W, Mitchell LG, Andrew M. Antithrombotic therapy in children. *Curr Opin Pediatr* 1999;11:56-64
- Haemostasis and Thrombosis Task Force, British Committee for Standards in Haematology. Guideline: investigation and management of heritable thrombophilia. *Br J Haematol* 2001;114:512-28

16 Anticoagulation in hospitals and general practice

Andrew D Blann, David A Fitzmaurice, Gregory Y H Lip

Service requirements for warfarin management include phlebotomy or finger pricking, accurate measurement of the international normalised ratio (INR) by a coagulometer (with associated standards and quality control), interpretation of the result, and advice on the warfarin dose. Clinical management of the complications of treatment (predominantly overdose) are also required. Furthermore, almost any drug can interact with oral anticoagulants, and many (such as steroids and antibiotics) often increase the anticoagulant effect.

When introducing a new drug, if the duration of treatment is short (such as an antibiotic for less than five days), then adjustment of dose is often not essential. If, however, the treatment is to last more than five days, then the INR should be checked after starting treatment with the new drug and the warfarin dose adjusted on the basis of the results.

Warfarin tablets used routinely in the United Kingdom

Starting treatment in hospital inpatients

Once the indications for anticoagulation have been confirmed (for example, for suspected deep vein thrombosis do venography or D-dimer measurement), the initial dose of oral anticoagulant depends on a patient's coagulation status, age, clinical situation, and degree of heart failure (if present). In older patients, those with impaired liver function, and those with congestive heart failure oral anticoagulation should be started cautiously and the resulting INR checked often (every three to five days). The dose of warfarin needed to maintain an INR at 2.0-3.0, for example, falls with age and is greater in patients of Indo-Asian or African origin than in Europeans. Where possible, take routine blood samples for prothrombin time and activated partial thromboplastin time (APTT), platelet count, and liver function tests before starting treatment. Oral anticoagulation with warfarin should be started on day one, preferably in conjunction with heparin because the initial period of treatment with warfarin may be associated with a procoagulant state caused by a rapid reduction in protein C concentration (itself a vitamin K dependent protein). Heparin should not be stopped until the INR has been in the therapeutic range for two consecutive days. Patients at a high risk of thrombosis and those with a large atrial thrombus may need longer treatment with heparin.

Similarly, a specific anticoagulant treatment chart that contains the treatment protocol, the results of coagulation tests (INR and APTT ratios), and the prescribed doses based on the results should be the basis of treatment and is a useful way of assessing and monitoring patients' anticoagulation in the follow up period. Daily INR measurement for at least four days is recommended in patients needing rapid anticoagulation (for example, in those with high risk of thrombosis). Adjustment of the oral anticoagulant loading dose may be necessary if baseline coagulation results are abnormal. Some patients may be particularly sensitive to warfarin, such as older people and those with liver disease, congestive cardiac failure, or who are recieving drug treatment (such as antibiotics) likely to increase the effects of oral anticoagulants.

Once the therapeutic INR range is achieved it should be monitored weekly until control is stable. The British Society for Haematology's guidelines suggest that thereafter blood testing can be extended to fortnightly checks, then checks every four

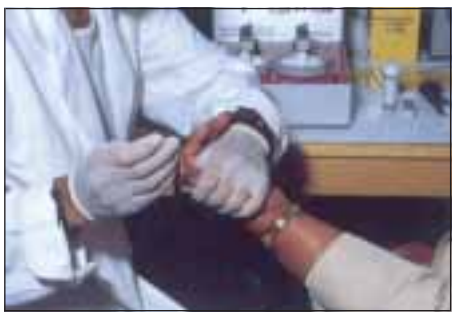
Drug interactions with warfarin*

Enhanced anticoagulant effect—Alcohol, allopurinol, anabolic steroids, analgesics (for example, paracetamol), antiarrhythmics (for example, amiodarone), antidepressants (for example, selective serotonin reuptake inhibitors), antidiabetics, antimalarials, antiplatelets, anxiolytics, disulfiram, influenza vaccine, leukotriene antagonists, levothyroxine, lipid regulating agents, testosterone, uricosurics

Reduced anticoagulant effect—Oral contraceptives, raloxifene, retinoids, rowachol, vitamin K (possibly present in enteral feeds)

Variable effect—Antibiotics (but, generally, more likely to enhance), colestyramine, antiepileptics, antifungals, barbiturates, cytotoxics (for example, effect enhanced by ifosfamide but often reduced by azathioprine), hormone antagonists, ulcer healing drugs

*This list is not exhaustive or definitive but provides perspective: the effect of each particular agent should be observed on each particular patient. Considerable variation exists in different drugs within a single class (for example, antibiotics). Refer to the *BNF* for guidance



Anticoagulation monitoring by fingerprick. Note coagulometer in the background

weeks, eight weeks, and 12 weeks (maximum). By this time, the checks are most likely to be in the setting of an experienced hospital outpatient clinic.

At the time of discharge from hospital, follow up arrangements for each patient should include sufficient tablets to allow adequate cover until the general practitioner can provide a prescription (two to three weeks' worth) and an appointment for further INR measurements, generally in an outpatient clinic. This period should not exceed seven days and should be detailed in the patient case notes and the yellow Department of Health anticoagulant booklets. Information in the yellow booklet should indicate the target INR range for each patient and other pertinent information, such as the presence of diabetes and indication for anticoagulation.

Starting treatment in outpatients

Without the benefit of the management procedures described above, starting anticoagulant treatment in outpatients can be difficult, especially if patients are referred without their notes or adequate information (such as other drugs prescribed or reason for anticoagulation). Nevertheless, local conditions and guidelines will generally recommend a starting dose, and patients will need to be recalled weekly for INR management until they are deemed to be stable. In many cases the introduction of computer assisted dosing (an algorithm software) is of immense benefit.

Requirement for daily dose of warfarin to maintain an INR between 2.0 and 3.0 and 3.0 and 4.5

Age (years)	No of patients	Daily dose of warfarin (mg)*
INR to be in range 2.0 to 3.0		
40-49	36	7.3 (6.21 to 8.39)
50-59	76	5.5 (5.0 to 6.0)
60-69	209	4.3 (4.05 to 4.55)
70-79	233	3.9 (3.68 to 4.12)
≥80	107	3.3 (3.01 to 3.59)
INR to be in range 3.0 to 4.5		
40-49	9	6.5 (5.23 to 7.77)
50-59	20	6.0 (5.2 to 6.8)
60-69	45	5.9 (5.13 to 6.67)
70-79	24	4.8 (4.15 to 5.45)
≥80	2	4.2 (2.65 to 5.75)

*Data are presented as mean (95% CI)
Relationship between age and daily dose in INR range 2.0-3.0 is correlation coefficient $r = -0.45$, $P < 0.001$, and for INR range 3.0-4.5 is correlation coefficient $r = -0.23$, $P = 0.022$
Data from Blann AD, et al, *Br J Haematol* 1999;107:207-9

Complications and reversal of oral anticoagulation

Bleeding complications while patients are receiving oral anticoagulants increase substantially when INR levels exceed 5.0, and therapeutic decisions depend on the presence of minor or major bleeding. However, in those cases with evidence of severe bleeding or haemodynamic compromise, hospitalisation, intensive monitoring, and resuscitation with intravenous fluids may be needed. Sometimes the bleeding point can be treated (for example, endoscopic treatment of bleeding peptic ulcer). Fresh frozen plasma is recommended when quick reversal of over-anticoagulation is needed. If plasma is unavailable then vitamin K, given by slow intravenous injection at doses of 0.5-1.0 mg or orally at doses of 1-10 mg, may reduce the INR within six to eight hours without the risk of over-correction. However, the effects of vitamin K can last for a week and may delay the restarting of warfarin treatment, although retesting (thus restarting with warfarin) after 48-72 hours is common.

Yellow Department of Health anticoagulant booklet. Columns are provided for the date of each visit, INR result, recommended daily dose, and signature

Treatment of excessive antithrombotic therapy effects

Class	Drug	Antidote
Oral anticoagulants	Warfarin	Oral or intravenous vitamin K; clotting factors or fresh frozen plasma, or both; recombinant factor VII
Intravenous or subdermal anticoagulants	Heparin	Protamine; clotting factors or fresh frozen plasma, or both
Thrombolytics	Streptokinase, tissue plasminogen activator (examples)	Transexamic acid

Maintenance in hospital practice

The traditional model of care for patients taking oral anticoagulants requires them to attend a hospital outpatient clinic so that the INR can be estimated. Capillary or venous blood samples are used, with the result being available either immediately or at a later stage. However, the INR derived from capillary (finger prick) blood is likely to be different from that obtained from plasma from a peripheral blood sample, and this should be considered. If possible, it is preferable to use consistently either finger prick or venous blood. Rarely, phlebotomists will visit housebound patients and return a venous sample to the laboratory for INR management. Where INR results are available with the patient present, dosing recommendations are made and the patient is given a date for the next appointment. When there is a delay in the INR estimation, patients receive dosing and recall advice through the

post or by telephone. Although this service has been traditionally led by a physician (usually a consultant haematologist) or pathologist, more recently biomedical scientists, nurse specialists, and pharmacists have been taking responsibility for anticoagulant clinics. This model has been widely used in the United Kingdom but has come under more strain because of increasing numbers of patients referred for warfarin treatment, particularly for stroke prophylaxis in atrial fibrillation. However, in terms of INR control, adverse events, or patient satisfaction, long term oral anticoagulant care has traditionally required patients to attend a hospital anticoagulant clinic repeatedly because of the need for laboratory testing, specialist interpretation of the result, and adjustment of warfarin dose.

Anticoagulation in general practice

Concerns over general practice involvement in anticoagulation monitoring have been expressed—namely, lack of resources (machines and reagents to generate the INR) and lack of expertise (experience and training), although these can be overcome. Despite various moves to decentralisation, no large scale development in a primary care setting has occurred. Understandably, general practitioners are anxious that decentralisation of anticoagulation care represents an additional, unwanted, and possibly dangerous burden. Local circumstances vary enormously so the process of decentralisation will need to be modified according to local needs and resources.

The establishment of a local development group consisting of general practitioners and hospital clinicians responsible for the anticoagulant clinic is one way of promoting decentralisation and identifying problem areas. There is increasing evidence that general practitioners or healthcare professionals such as biomedical scientists, pharmacists, and practice nurses, with or without computer assisted dosing, are able to achieve high standards of anticoagulation care with "near patient" testing. As the principle of near patient testing is well developed in glucose monitoring by or for diabetic patients, it seems logical that it can be transferred to oral anticoagulation, provided that adequate levels of accuracy and safety are achievable.

In one of the more widespread models general practitioners take a blood sample and dosing decisions are made by a hospital department, with patients receiving dosing information through the post or by telephone. This model retains the expertise and quality assurance of the laboratory process while decentralising at minimal cost to primary care. Patients can attend their (usually more convenient) general practitioner's surgery and a venous blood sample is sent to the central laboratory. INR is determined and information on dose and the next appointment is sent to the patient. There are no clinically significant changes in the INR when analysis is delayed for up to three days, and the quality control with near patient sampling is at least equal to that in a hospital based setting. This process requires access to phlebotomy in general practice, and the cost of testing and dosing remains in the central laboratory.

General practices with limited access to hospital clinics are more likely to undertake the second level of care and give dosing advice. General practitioners who do not have access to computer assisted dosing seem to have similar success to hospital clinics in achieving optimum INR control.

The third level of care uses near patient testing for INR estimation and computer assisted dosing for recommendation of dose and recall. Anticoagulant clinics are managed by practice nurses with support from the general practitioner and hospital laboratory. Liaison with the hospital laboratory is

Factors affecting delivery of anticoagulation therapy

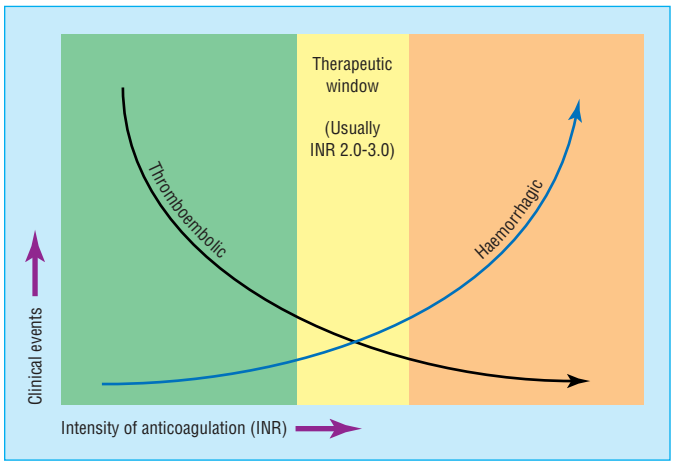
Hospital anticoagulation clinics
- Usually busy and congested
- Congestion is an increasing problem caused by the ageing population with more indications for warfarin (especially atrial fibrillation)
- Inconvenient

Domiciliary anticoagulation service
- Depends on resources
- Limited availability
- Useful for those who are immobile or housebound

"Near patient" testing
- Requires considerable resources
- Dependent on primary care, facilities, and training

Potential levels of involvement in general practice for managing anticoagulation treatment

- Phlebotomy in the practice (by practice or hospital staff), blood sent (post or van) to the hospital laboratory with the result returned to the practice (telephone, fax, post, or email), dosing decisions being made in the practice, then communicated to the patient
- Phlebotomy in the practice, blood sent (post or van) to the hospital laboratory, dosing with INR estimation performed in the hospital and patient managed directly (telephone or post)
- Phlebotomy, INR estimation (plus dosing) and management all performed in the practice with hospital equipment and by hospital staff
- Phlebotomy, INR estimation plus dosing and management made by the practice (that is, full near patient testing); minimal input from the hospital



The therapeutic "window" is a balance between the best reduction in thromboembolic events and increased risk of bleeding with higher intensities of anticoagulation. Adapted from Hylek EM, et al. *New Engl J Med* 1993;120:897-902

paramount to the success of such a clinic as it needs to provide training and guidance on near patient testing technique, quality assurance, and health and safety issues. In a study by Fitzmaurice, et al (2000), INR therapeutic range analyses as point prevalence, proportion of tests in range, number of serious adverse events, and proportion of time in range all compared well with the hospital control patients. However, the proportion of time spent in the INR range showed substantial improvement for patients in the intervention group.

Computer assisted dosing aids interpretation of results, although it can be over-ridden if the suggestion made is not clinically indicated. For an effective and reliable service it is essential to ensure formal training and quality assurance procedures for near patient testing at the initial stages of the clinic development. This model of care gives an immediately available result, and, with close liaison with a hospital laboratory, it offers patients a complete model of care that would be a useful alternative to traditional care.

Another primary care model that has had limited evaluation is that of anticoagulant clinics that are managed entirely by scientists and pharmacists. These specialist healthcare professionals make use of their expertise in coagulation and pharmacology respectively. Secondary care anticoagulant clinics run by scientists and pharmacists have existed in the United Kingdom since 1979, and in terms of INR control they perform as well as clinics run by pathologists. Patients also prefer general practice management and welcome reduced waiting times and travelling costs. Improved patient understanding may also occur, which can help compliance. Further clinics managed by scientists or pharmacists, or both, are currently being evaluated.

Patient self monitoring and dosing

Diabetic patients have long been able to use portable monitoring machines to check their own blood glucose concentrations and administer insulin accordingly. As equivalent machines for checking INR are now available, increased patient demand is likely to rise. The machine will appeal especially to those receiving long term anticoagulation whose lifestyle is not suited to the inconvenience of attending outpatient clinics. As with diabetic patients, well trained and motivated patients can probably attain a level of control of their own warfarin dose similar to that of the hospital. As yet, there are no comparison data on the safety and reliability of such an approach, so great caution is needed in offering (or even recommending) this option, which will be applicable to a well defined subset of patients. However, most pilot data suggest that patient self management is as safe as primary care management for a selected population, and further study is needed to show if this model of care is suitable for a larger population.

Conclusion

The quality of anticoagulant care has improved in recent years with the development of clinical guidelines (for example, by the haemostasis and thrombosis task force of the British Society for Haematology), adoption of the INR system, quality control assurance, computerised decision support systems, and clinical audit. This allows a gradual movement of dosing from hospital to general practice. New models of delivering care (such as near patient testing) are now being developed to meet the increasing demand from an ageing population, such as from the growing number of patients with atrial fibrillation, whose risk of stroke is markedly reduced by anticoagulant therapy.

Contraindications to warfarin use and management

The patient
- Comorbidity—including comorbid medical conditions, falls, frailty, exposure to trauma
- Impaired cognitive function
- Possibly housebound
- Poor compliance

The doctor
- Poor appreciation of drug interactions
- Inefficient organisation of INR monitoring

The system
- General practice *v* hospital facilities—for example remote location and poor communication and support
- Inadequate resources and facilities available

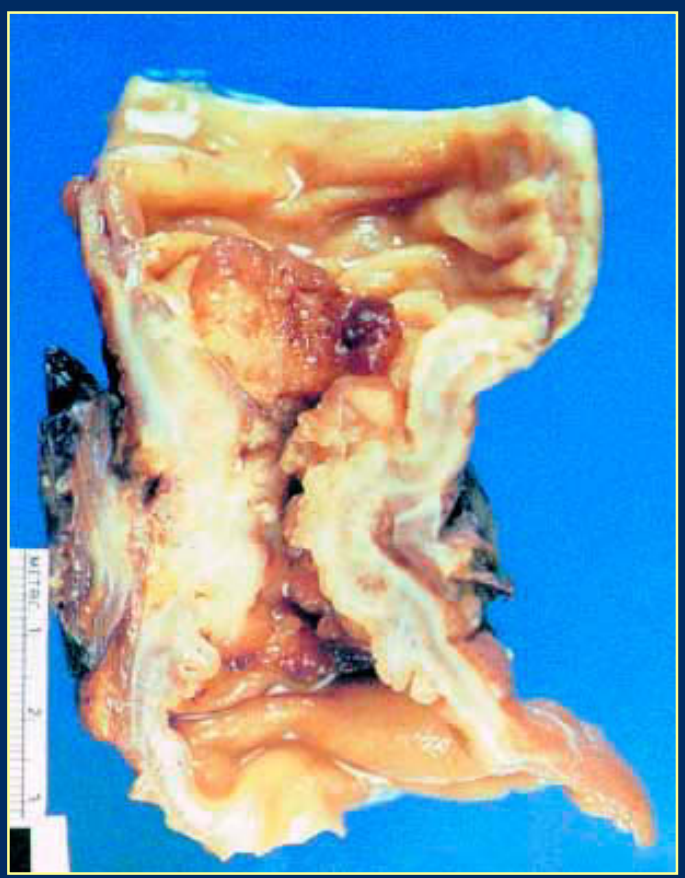
Further reading

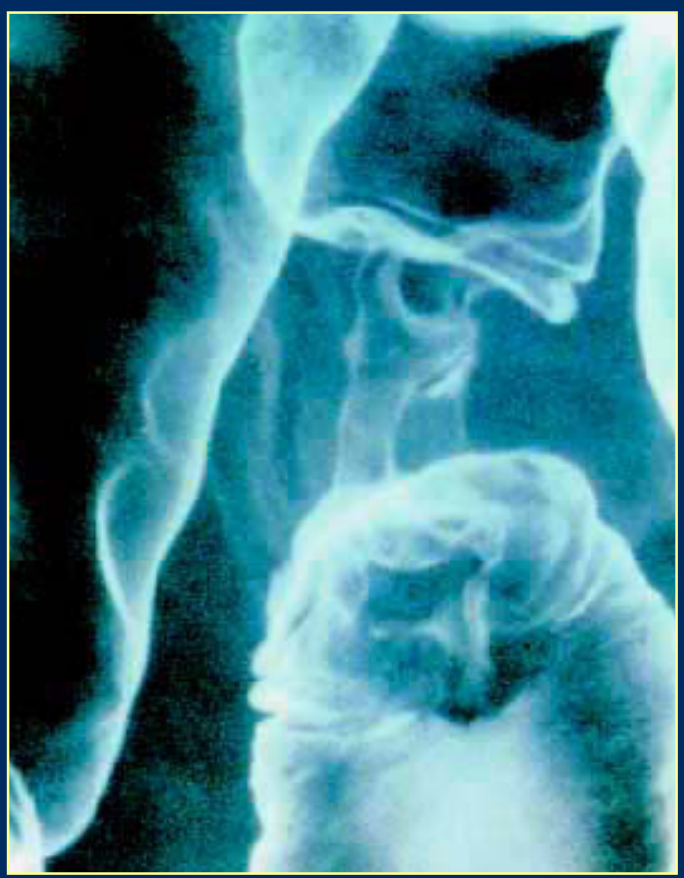
- Baglin T, Luddington R. Reliability of delayed INR determination: implications for decentralised anticoagulant care with off-site blood sampling. *Br J Haematol* 1997;96:431-4
- Blann AD, Hewitt J, Siddique F, Bareford D. Racial background is a determinant of average warfarin dose required to maintain the INR between 2.0 and 3.0. *Br J Haematol* 1999;107:207-9
- Fitzmaurice DA, Hobbs FDR, Delaney BC, Wilson S, McManus R. Review of computerized decision support systems for oral anticoagulation management. *Br J Haematol* 1998;102:907-9
- Fitzmaurice DA, Murray ET, Gee KM, Allan TF, Hobbs FD. A randomised controlled trial of patient self management of oral anticoagulation treatment compared with primary care management. *J Clin Pathol* 2002;55:845-9
- Fitzmaurice DA, Hobbs FD, Murray ET, Holder RL, Allan TF, Rose PE. Oral anticoagulation management in primary care with the use of computerized decision support and near-patient testing: a randomized, controlled trial. *Arch Intern Med* 2000;160:2343-8
- Haemostasis and Thrombosis Task Force of the British Society for Haematology. Guidelines on anticoagulation: third edition. *Br J Haematol* 1998;101:374-87
- MacGregor SH, Hamley JG, Dunbar JA, Dodd TRP, Cromarty JA. Evaluation of a primary care anticoagulation clinic managed by a pharmacist. *BMJ* 1996;312:560
- Pell JP, McIver B, Stuart P, Malone DNS, Alcock J. Comparison of anticoagulant control among patients attending general practice and a hospital anticoagulant clinic. *Br J Gen Pract* 1993;43:152-4
- Radley AS, Hall J, Farrow M, Carey PJ, Evaluation of anticoagulant control in a pharmacist operated anticoagulant clinic. *J Clin Pathol* 1995;48:545-7

Index

Page numbers in **bold** type refer to figures; those in *italic* refer to tables or boxed material.

ABC OF COLORECTAL CANCER





Edited by
David J Kerr, Annie M Young and FD Richard Hobbs

BMJ Books

ABC OF COLORECTAL CANCER

Edited by

D J KERR
Professor, Institute for Cancer Studies, University of Birmingham

ANNIE M YOUNG
Research Fellow, School of Health Sciences, University of Birmingham

and

F D RICHARD HOBBS
Professor, Department of Primary Care and General Practice, University of Birmingham

BMJ Books

ABC OF COLORECTAL CANCER

Edited by

D J KERR
Professor, Institute for Cancer Studies, University of Birmingham

ANNIE M YOUNG
Research Fellow, School of Health Sciences, University of Birmingham

and

F D RICHARD HOBBS
Professor, Department of Primary Care and General Practice, University of Birmingham

BMJ Books

BMJ Books is an imprint of the BMJ Publishing Group

First published in 2001
by BMJ Books, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0-7279-1526-6

Cover design by Marritt Associates, Harrow, Middlesex
Composition by Scribe Design, Gillingham, Kent
Printed and bound in Spain by GraphyCems

Contents

This Page Intentionally Left Blank

Contributors

Peter Boyle
Director of Division of Epidemiology and Biostatistics, European Institute, Milan, and Honorary Professor of Cancer Epidemiology, University of Birmingham

G A Chung-Faye
Clinical Research Fellow, Institute for Cancer Studies, University of Birmingham

T R P Cole
Consultant in Clinical Cancer Genetics, Birmingham Women's Hospital, Birmingham

Robert G Hardy
Wellcome Trust Clinical Research Fellow, Departments of Medicine and Surgery, University Hospital, Birmingham

F D Richard Hobbs
Professor, Department of Primary Care and General Practice, University of Birmingham

Janusz A Jankowski
Reader and Consultant Gastroenterologist, Department of Medicine, University Hospital, Birmingham and Imperial Cancer Research Fund, London

D J Kerr
Professor, Institute for Cancer Studies, University of Birmingham

Michael J S Langman
Professor of Medicine, University of Birmingham

Colin McArdle
Professor, University Department of Surgery, Royal Infirmary, Edinburgh

Stephen J Meltzer
Professor of Medicine and Director, Functional Genomics Laboratory, University of Maryland Greenbarn Cancer Centre, Baltimore, USA

Rachel S J Midgley
Clinical Research Fellow, CRC Institute for Cancer Studies, University of Birmingham

Daniel Rea
Senior Lecturer in Medical Oncology, Institute for Cancer Studies, University of Birmingham

John H Scholefield
Professor of Surgery, Division of Gastrointestinal Surgery, University Hospital, Nottingham

H V Sleightholme
Regional Cancer Coordinator, Department of Clinical Genetics, Birmingham Women's Hospital, Birmingham

Annie M Young
Research Fellow, School of Health Sciences, University of Birmingham

This Page Intentionally Left Blank

Preface

The inspiration for this book stemmed from the widely shared optimism of colorectal cancer specialists that after many decades of often painfully slow progress (despite much action), we are at the brink of a new era with several positive developments in the prevention, diagnosis and treatment of colorectal cancer, bringing hope to the hundreds of thousands of people who develop the disease.

It is crucial that this evidence-based sanguinity spreads to the entire multiprofessional colorectal cancer team, in particular to general practitioners who are by and large the first, the intermediate and the last point of contact for our patients. They have the complex task of firstly identifying suspected colorectal cancer and then working in partnership with the patient, carers and the specialists at all stages along the patient pathway. This book is written for them – the primary care physician, the nurses, the junior doctors, the dieticians, the radiographers and countless other healthcare professionals, all caring for colorectal cancer patients.

It isn't just that the book walks us through contemporary knowledge in the prevention, diagnosis, prognosis and modalities of treatment for colorectal cancer (and many other things besides) but it also acknowledges and debates the numerous uncertainties around the disease in a balanced manner, in addition to peering into future approaches towards screening, molecular biology, genetics and therapies.

The book, in short, presents a concise story of the full spectrum of colorectal cancer in a kind of chronological order. All of us who care for colorectal cancer patients, should make it our duty to be acquainted with the detail in order to provide optimal patient care.

Annie Young

This Page Intentionally Left Blank

1 Epidemiology

Peter Boyle, Michael J S Langman

In countries with a westernised lifestyle about half of all deaths are caused by circulatory disease and a quarter by cancer. Cancer is an important problem in both public health and political terms worldwide, irrespective of a country's development. The most recent estimates of the global cancer burden suggest that there were 8.1 million new cases, excluding non-melanoma skin cancer, worldwide in 1990. About 10 million new cases are now diagnosed each year.

Colorectal cancer is the fourth commonest form of cancer occurring worldwide, with an estimated 783 000 new cases diagnosed in 1990, the most recent year for which international estimates are available. It affects men and women almost equally, with about 401 000 new cases in men annually and 381 000 in women. The number of new cases of colorectal cancer worldwide has been increasing rapidly since 1975 (when it was 500 000).

Worldwide, colorectal cancer represents 9.4% of all incident cancer in men and 10.1% in women. Colorectal cancer, however, is not equally common throughout the world. If the westernised countries (North America; those in northern, southern, and western Europe; Australasia; and New Zealand) are combined, colorectal cancer represents 12.6% of all incident cancer in westernised countries in men and 14.1% in women. Elsewhere colorectal cancer represents 7.7% and 7.9% of all incident cases in men and women respectively.

Large differences exist in survival, according to the stage of disease. It is estimated that 394 000 deaths from colorectal cancer still occur worldwide annually, and colorectal cancer is the second commonest cause of death from any cancer in men in the European Union. Substantial differences in cancer survival seem to exist between Great Britain, Europe as a whole, and the United States. This variation in survival is not easily explained but could be related to stage of disease at presentation or treatment delivery, or both of these.

The numbers of new cases of colorectal cancer worldwide has increased rapidly since 1975

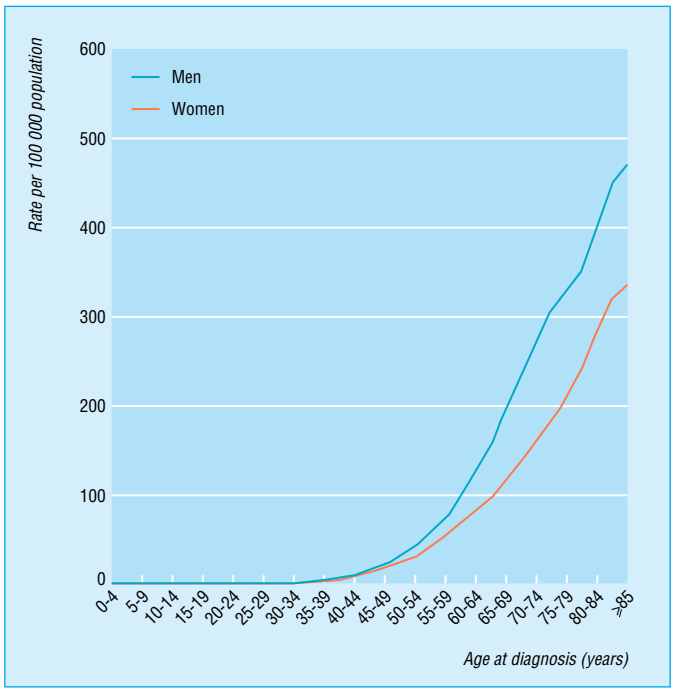


Figure 1.1 Estimated incidence of colorectal cancer in United Kingdom, by age and sex, 1995

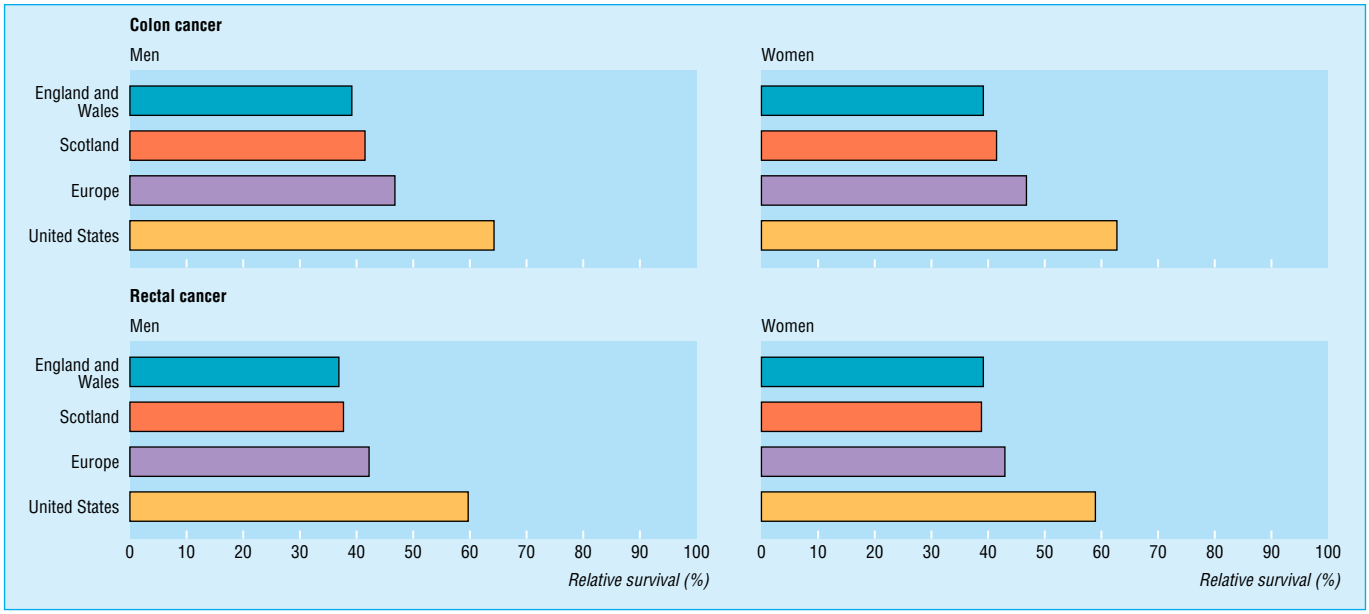


Figure 1.2 International comparison of five year relative survival for colon and rectal cancer in adults aged 15-99 at diagnosis (based on Coleman et al, *Cancer survival trends in England and Wales, 1971-1975*; Berrino et al, *Survival of cancer patients in Europe: the EUROCARE-2 study*; and Surveillance Epidemiology and End Results (SEER) programme, National Cancer Institute, 1998)

Survival and deprivation

The relation between poverty and ill health has been researched for more than 100 years. In Scotland, since the 1851 census, all cause occupational mortality has been routinely reported, and since 1911, inequalities in health, as shown by mortality, have been examined in decennial reports classified by social class (based on occupation) and by occupational group alone.

No single, generally agreed definition of deprivation exists. Deprivation is a concept that overlaps but is not synonymous with poverty. Absolute poverty can be defined as the absence of the minimum resources for physical survival, whereas relative poverty relates to the standards of living in a particular society. Deprivation includes material, social, and multiple deprivation. In Scotland the Carstairs and Morris index of deprivation was derived from 1981 census data with the postcode sector as the basic geographical unit (covering a population of about 5000). This index describes a deprivation category on a scale of 1 (least deprived) to 7 (most deprived) for each household address in Scotland.

The incidence of colorectal cancer is higher in men than women among each of the seven deprivation categories in Scotland, although incidence varies little with deprivation category. Survival, however, clearly improves with decreasing deprivation. At each milestone, there is a notable gradient in survival, with the most affluent doing best and the least affluent doing worst. The reasons that such variations exist are unclear and highlight an important priority for research.

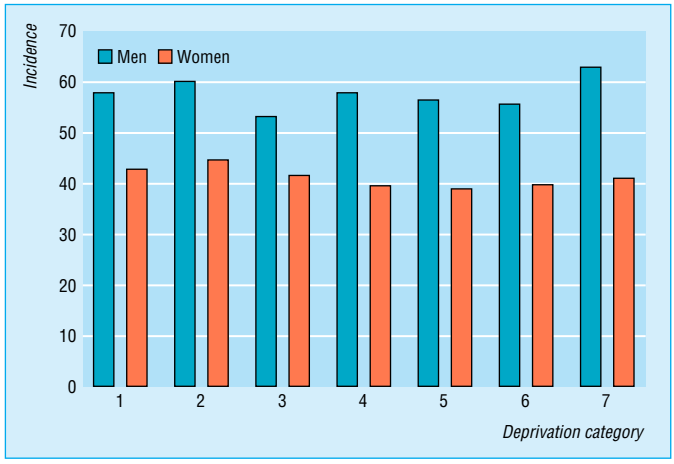


Figure 1.3 Incidence according to deprivation category in Scotland, 1998 (1=least deprived, 7=most deprived)

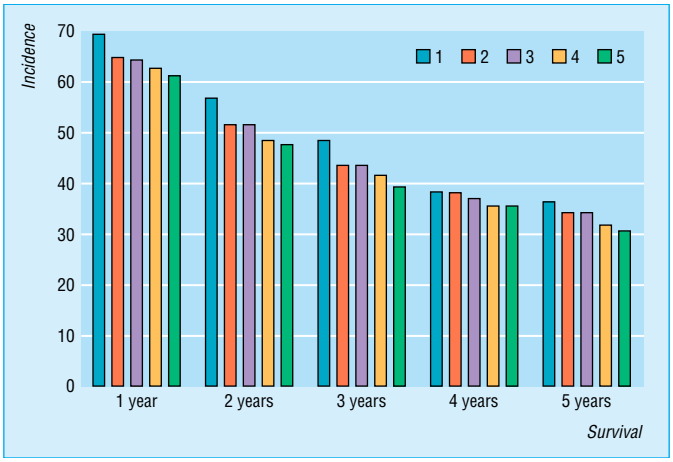


Figure 1.4 Survival according to deprivation category in Scotland, 1998 (1=least deprived, 5=most deprived)

Descriptive epidemiology

Different populations worldwide experience different levels of colorectal cancer, and these levels change with time. Populations living in one community whose lifestyles differ from those of others in the same community also experience different levels of colorectal cancer. Groups of migrants quickly lose the risk associated with their original home community and acquire the patterns of the new community, often starting within one generation of arrival.

Ethnic and racial differences in colorectal cancer, as well as studies on migrants, suggest that environmental factors play a major part in the aetiology of the disease. In Israel male Jews born in Europe or the United States are at higher risk of colon cancer than those born in Africa or Asia. Risk in the offspring of Japanese populations who have migrated to the United States has changed—incidence now approaches or surpasses that in white people in the same population and is three or four times higher than among the Japanese in Japan.

For reasons such as these, colorectal cancer is widely believed to be an environmental disease, with "environmental" defined broadly to include a wide range of ill defined cultural, social, and lifestyle practices. As much as 70-80% of colorectal cancers may owe their appearance to such factors; this clearly identifies colorectal cancer as one of the major neoplasms in which causes may be rapidly identified, and a large portion of the disease is theoretically avoidable.

The move from theoretically avoidable causes to implementation of preventive strategies depends on the identification of risk factors, exposures that have been associated with an increased (or decreased) risk of colorectal cancer, and the smaller subset of risk determinants, whose alteration would lead directly to a reduction in risk. From analytical epidemiology some clear ideas have now emerged about measures for reducing the risk of colorectal cancer.

Table 1.1 Highest incidence of colorectal cancer in men worldwide around 1990

Registry	Age standardised incidence per 100 000
US (Hawaii: Japanese), 1988-92	53.48
New Zealand (non-Maori), 1988-92	51.30
Japan (Hiroshima), 1986-90	50.99
France (Haut-Rhin), 1988-92	49.90
Italy (Trieste), 1989-92	49.37
France (Bas-Rhin), 1988-92	49.24
Canada (Yukon), 1983-92	48.98
US (Detroit: black), 1988-92	48.32
Czech Republic, 1988-92	48.23
US (Los Angeles: black), 1988-92	47.89
Canada (Nova Scotia), 1988-92	47.84
Canada (Newfoundland), 1988-92	47.29
Australia (New South Wales), 1988-92	46.92
US (San Francisco: black), 1988-92	46.82
Israel (Jews born in America or Europe), 1988-92	46.79

Data taken from Parkin et al, eds (*Cancer incidence in five continents*. Vol 7. IARC Scientific Publications, 1997:120)

Dietary and nutritional practices

Evidence from epidemiological studies seems to show consistently that intake of dietary fat and meat is positively related to risk of colorectal cancer. This evidence is obtained from ecological studies, animal experiments, and case-control and cohort studies.

In 1990 Willett et al published the results from the US nurses health study involving follow up of 88 751 women aged 34-59 years who were without cancer or inflammatory bowel disease at recruitment. After adjustment for total energy intake, consumption of animal fat was found to be associated with increased risk of colon cancer. The trend in risk was highly significant ($P = 0.01$), with the relative risk in the highest compared with the lowest quintile being 1.89 (95% confidence interval 1.13 to 3.15). No association was found with vegetable fat. The relative risk in women who ate beef, pork, or lamb as a main dish every day was 2.49 (1.24 to 5.03) compared with women reporting consumption less than once a month. The authors suggested that their data supported the hypothesis that a high intake of animal fat increases the risk of colon cancer, and they supported existing recommendations to substitute fish and chicken for meats high in fat.

Figure 1.5 Intake of dietary fat and meat may increase risk of colorectal cancer

Intake of vegetables, fruit, and fibre

Dietary fibre has been proposed as accounting for the differences in the rates of colorectal cancer between Africa and westernised countries—on the basis that increased intake of dietary fibre may increase faecal bulk and reduce transit time. Various other factors, related to risk of colorectal cancer, are now thought to contribute to explaining these differences.

Fibre has many components, each of which has specific physiological functions. The components are most commonly grouped into insoluble, non-degradable constituents (mainly found in cereal fibre) and soluble, degradable constituents, such as pectin and plant gums (mainly found in fruits and vegetables). Epidemiological studies have reported differences in the effect of these components. Many studies, however, found no protective effect of fibre in cereals but have consistently found a protective effect of fibre in vegetables and perhaps fruits. This might reflect an association with other components of fruits and vegetables, with fibre intake acting merely as an indicator of consumption.

Figure 1.6 Fruits are a good source of fibre and may protect against cancer

Physical activity, body mass index, and energy intake

Evidence from epidemiological studies is strong that men with high occupational or recreational physical activity seem to have a decreased risk of colon cancer. Such evidence comes from follow up studies of cohorts who are physically active or who have physically demanding jobs, as well as from case-control studies that have assessed physical activity by, for example, measurement of resting heart rate or questionnaire. The association remains even after potential confounding factors, such as diet and body mass index, are controlled for.

The available data, however, show no consistent association between obesity and risk of colorectal cancer (analysis and interpretation of this factor is difficult in retrospective studies, where weight loss may be a sign of the disease), although evidence now suggests an association between obesity and adenomas. This increased risk associated with energy intake does not seem to be the result merely of overeating; it may reflect differences in metabolic efficiency. If the possibility that the association with energy intake is a methodological artefact is excluded (as such a consistent finding is unlikely to emerge from such a variety of study designs in diverse population groups), it would imply that individuals who use energy more efficiently may be at a lower risk of colorectal cancer.

Box 1.1 Physical activity and colorectal cancer

- Giovannucci et al examined the role of physical activity, body mass index, and the pattern of adipose distribution in the risk of colorectal adenomas
- In the nurses health study, 13 057 female nurses, aged 40-65 years in 1986, had an endoscopy during 1986-92. During this period, adenoma of the distal colorectum was newly diagnosed in 439 nurses
- After age, prior endoscopy, parental history of colorectal cancer, smoking, aspirin use, and dietary intake were controlled for, physical activity was associated inversely with the risk of large adenomas (≥1 cm) in the distal colon (relative risk 0.57 (95% confidence interval 0.30 to 1.08)) when high and low quintiles of average weekly energy expenditure from leisure activities were compared
- Much of this benefit came from activities of moderate intensity, such as brisk walking

Hormone replacement therapy

Increasing evidence supports an (originally unexpected) association between hormone replacement therapy and a reduced risk of colorectal cancer.

Of 19 published studies of hormonal replacement therapy and risk of colorectal cancer, 10 support an inverse association and a further five show a significant reduction in risk. The risk seems lowest among long term users. Although some contradictions still exist in the available literature, hormone replacement therapy seems likely to reduce the risk of colorectal cancer in women. The risk seems to halve with 5-10 years' use. The role of unopposed versus combination hormone replacement therapy needs further research.

Whether this association is causal or is associated with some selection factor that directs women to using hormone replacement therapy is, however, not known. This question is important; if the link is indeed causal, women who are at high risk of colorectal cancer could be offered the therapy to lower their risk.

Control of colorectal cancer

Prospects for preventing death from colorectal cancer are now more promising than even 10 years ago. To achieve this goal public health decisions have to be taken, and part of this decision process involves deciding at which point enough epidemiological evidence is available to change focus comfortably from information generation to health actions.

To turn research findings into public health strategies for controlling the incidence of and mortality from colorectal cancer requires a profound change of mentality in the epidemiological community. It is easy to say that more studies are needed, but they would be unlikely to alter existing conclusions. Moreover, the implementation of strategies to control cancer must be considered separately from research into the control of cancer.

One consequence of epidemiological research into the contribution of lifestyle factors to cancer risk has been to blame the individual who develops cancer. Smoking, alcohol, dietary imprudence, and exposure to sunlight tend to assign responsibility to the individual. The individual is often not principally responsible for decisions about factors that influence his or her risk of cancer, and society—including government and industry—could do more to discourage lifestyles associated with cancer risk. Government legislation, including taxation policy and other actions, could have profound effects on smoking habits, for example.

The goal of all cancer research and treatment is to prevent people dying from the disease. Knowledge has been accruing rapidly about actions and interventions that could lead to a reduction in death from colorectal cancer by reducing the risk of developing the disease, identifying the disease at a stage when it is more curable, or improving the outcome of treatment.

Box 1.2 How individuals can reduce their risk of colorectal cancer

- Increase intake of vegetables and fruits (eat five servings of fruits and vegetables each day); replace snacks such as chocolate, biscuits, and crisps with an apple, orange, or other fruit or vegetable
- Reduce intake of calories (animal fats in particular); often replace beef, lamb, and pork with fish and poultry
- Increase physical activity—by activities of moderate intensity, such as brisk walking
- Participate in population screening programmes; when these are not in place, strongly consider having a colonoscopy with polyp removal once between ages 50 and 59
- Consult a doctor as soon as possible if a noticeable and unexplained change in bowel habits occurs, blood is present in the stool, colicky pain occurs in the abdomen, or a sensation of incomplete evacuation after defecation recurs

Further reading

- Boyle P. Progress in preventing death from colorectal cancer [editorial]. *Br J Cancer* 1995;72:528-30.
- Berrino F, Capocaccia R, Estève J, Gatta G, Hakulinen T, Micheli A, et al, eds. *Survival of cancer patients in Europe: the EUROCARE-2 study*. Lyons: International Agency for Research on Cancer, 1999. (Scientific publication No 151.)
- Coleman MP, Babb P, Damiecki P, Grosclaude P, Honjo S, Jones J, et al. Cancer survival trends in England and Wales, 1971-1995: deprivation and NHS region. London: Stationery Office, 1999.
- Giovannucci E, Colditz GA, Stampfer MJ, Willett WC. Physical activity, obesity and risk of colorectal cancer in women (United States). *Cancer Causes Control* 1996;7:253-63.
- McLaren G, Bain M. *Deprivation and health in Scotland: insights from NHS data*. Edinburgh: ISD Scotland, 1998.
- MacLennan SC, MacLennan AH, Ryan P. Colorectal cancer and oestrogen replacement therapy: a meta-analysis of epidemiological studies. *Med J Aust* 1991;162:491-3.
- Parkin DM, Pisani P, Ferlay J. Estimates of the worldwide incidence of 25 major cancers in 1990. *Int J Cancer* 1999;80:827-41.
- Shephard RJ. Exercise in the prevention and treatment of cancer—an update. *Sports Med* 1993;15:258-80.
- Willett WC. The search for the causes of breast and colon cancer. *Nature* 1989;338:389-94.
- Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Speizer FE. Relation of meat, fat, and fiber intake to the risk of colon cancer in a prospective study among women. *N Engl J Med* 1990;323:1664-72.

The two graphs showing incidence of colorectal cancer in the United Kingdom and an international comparison of five year relative survival for colon and rectal cancer are adapted with permission from the Cancer Research Campaign *(CRC CancerStats: Large Bowel—UK*; factsheet, November 1999) The graphs of incidence and survival according to deprivation category are adapted from McLaren G et al (*Deprivation and health in Scotland*. ISD Scotland Publications, 1998). The photograph of meat is published with permission from Tim Hall/CEPHAS.

2 Molecular basis for risk factors

Robert G Hardy, Stephen J Meltzer, Janusz A Jankowski

Evidence for the molecular basis of colorectal cancer comes from genetic analysis of tissues either from patients with a family history of the disease or from patients with sporadic adenomatous colorectal polyps or extensive ulcerative colitis. The traditional view is that background rates of genetic mutation, combined with several rounds of clonal expansion, are necessary for a tumour to develop. It has recently been argued, however, that inherent genetic instability not only is necessary but may also be sufficient for cancer to develop.

Sporadic colorectal adenomas

More than 70% of colorectal cancers develop from sporadic adenomatous polyps, and postmortem studies have shown the incidence of adenomas to be 30-40% in Western populations. Polyps are asymptomatic in most cases and are often multiple. Flat adenomas, which are more difficult to detect at endoscopy, account for about 10% of all polyps and may have a higher rate of malignant change or may predispose to a more aggressive cancer phenotype.

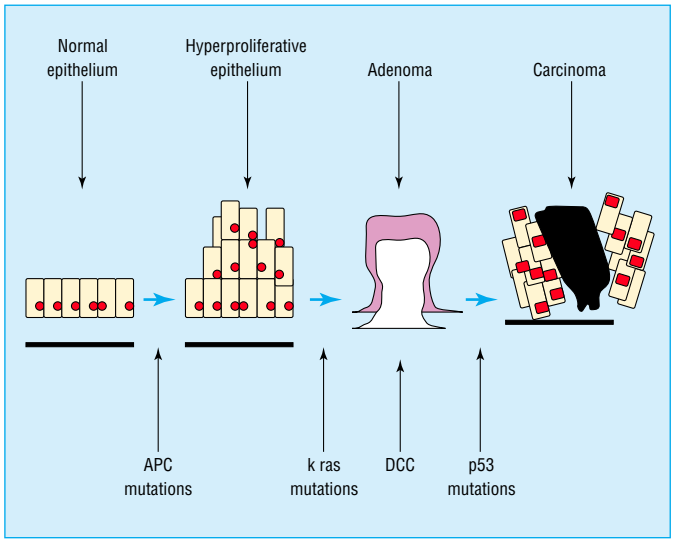


Figure 2.1 Proposed adenoma to carcinoma sequence in colorectal cancer. Adenomatous polyposis coli (APC) gene mutations and hypermethylation occur early, followed by k ras mutations. Deleted in colon cancer (DCC) and p53 gene mutations occur later in the sequence, although the exact order may vary

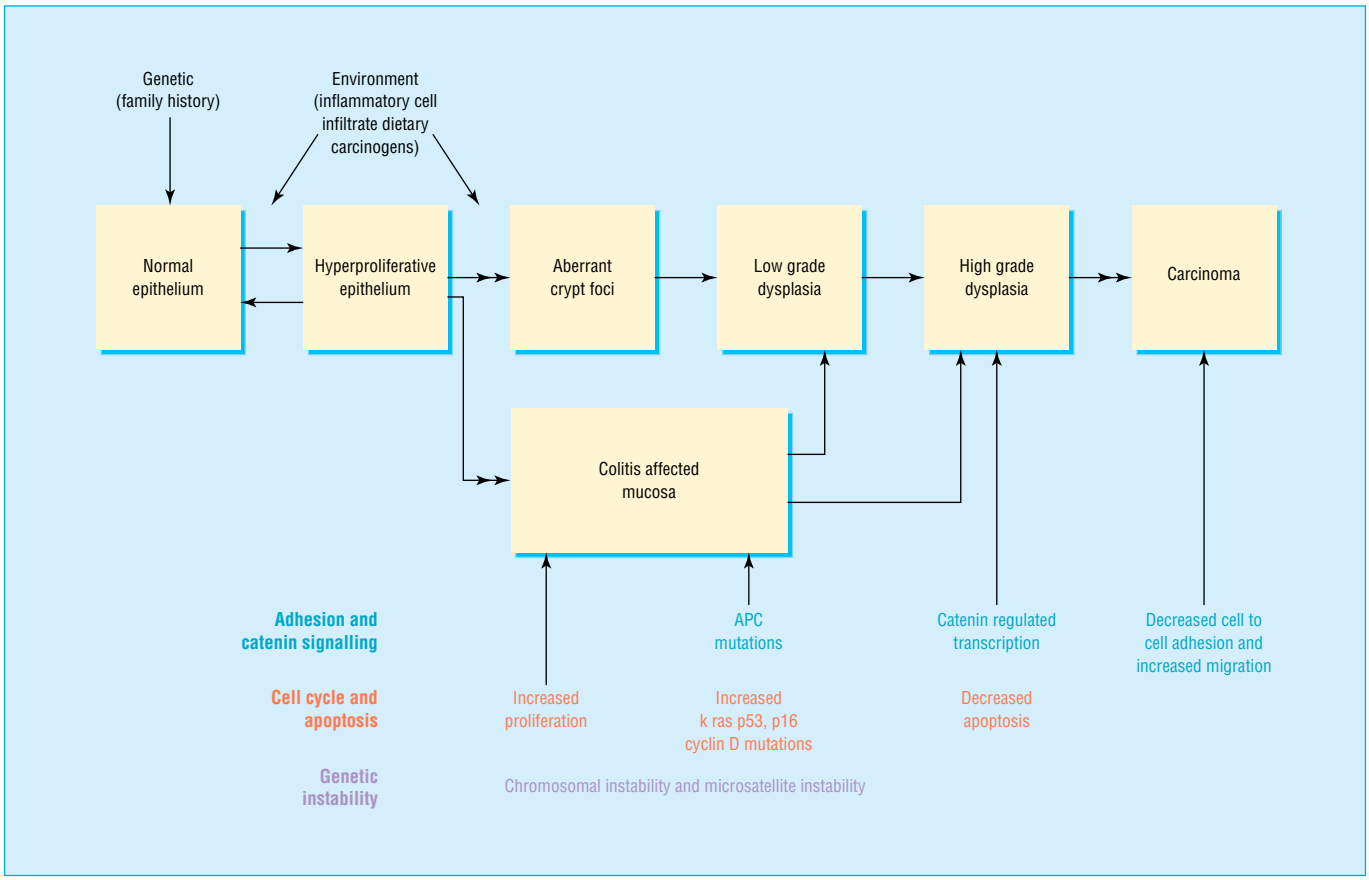


Figure 2.2 Key molecular events in colorectal premalignancy: comparisons between the adenoma carcinoma sequence and ulcerative colitis associated neoplasia

Family history

Recognised familial syndromes account for about 5% of colorectal cancers. The commonest hereditary syndromes are familial adenomatous polyposis and heredity non-polyposis colon cancer. Patients with these syndromes usually have a family history of colorectal cancer presenting at an early age. Attenuated familial adenomatous polyposis, juvenile polyposis syndrome, and Peutz-Jeghers syndrome are rarer, mendelian causes of colorectal cancer. In familial adenomatous polyposis (a mendelian dominant disorder with almost complete penetrance) there is a germline mutation in the tumour suppressor gene for adenomatous polyposis coli (APC) on chromosome 5.

Heredity non-polyposis colon cancer also shows dominant inheritance, and cancers develop mainly in the proximal colon. Patients with heredity non-polyposis colon cancer show germline mutations in DNA mismatch repair enzymes (which normally remove misincorporated single or multiple nucleotide bases as a result of random errors during recombination or replications). Mutations are particularly demonstrable in DNA with multiple microsatellites ("microsatellite instability").

In addition to the well recognised syndromes described above, clusters of colorectal cancer occur in families much more often than would be expected by chance. Postulated reasons for this increased risk include "mild" APC and mismatch repair gene mutations, as well as polymorphisms of genes involved in nutrient or carcinogen metabolism.

Box 2.1 Factors determining risk of malignant transformation within colonic adenomatous polyps

High risk
Large size (especially > 1.5 cm)
Sessile or flat
Severe dysplasia
Presence of squamous metaplasia
Villous architecture
Polyposis syndrome (multiple polyps)

Low risk
Small size (especially < 1.0 cm)
Pedunculated
Mild dysplasia
No metaplastic areas
Tubular architecture
Single polyp

The immediate family members of a patient with colorectal cancer will have a twofold to threefold increased risk of the disease

Table 2.1 Clinical and molecular correlates of familial adenomatous polyposis coli; attenuated familial adenomatous polyposis coli/hereditary flat adenoma syndrome; hereditary non-polyposis colon cancer/Lynch forms of hereditary colorectal cancer; and ulcerative colitis associated neoplasia

	FAP	AFAP/HFAS	HNPCC/Lynch	UCAN
Mean age at diagnosis of colorectal cancer	32-39	45-55	42-49	40-70
Distribution of cancer	Random	Mainly right colon	Mainly right colon	Mainly left colon
No of polyps	>100	1-100	1 (ie tumour)	
Sex ratio (male:female)	1:1	1:1	1.5:1	1:1
Endoscopic view of polyp	Pedunculated	Mainly flat	Pedunculated (45%); flat (55%)	None
Lag time (years) from early adenoma to occurrence of cancer	10-20	10	5	?<8
Proportion (%) of colonic cancer	1	0.5	1-5	<0.5
Superficial physical stigmata	80% have retinal pigmentation	None	Only in Muir-Torre syndrome	None
Distribution of polyps	Distal colon or universal	Mainly proximal to splenic flexure with rectal sparing	Mainly proximal to splenic flexure	None
Carcinoma histology	More exophytic growth	Non-exophytic but very variable	Inflammation increased mucin	Mucosal ulceration and inflammation
Other associated tumours	Duodenal adenoma cerebral and thyroid tumours, medulloblastoma and desmoids	Duodenal adenoma	Endometrial ovarian, gastric cancer, glioblastoma, many other cancers	
Gene (chromosome) mutation	APC (5q 21) distal to 5	APC (5q 21) proximal to 5	MHS2 (2p), MLH1 (3p21), PMS1 (2q31), PMS2 (7p22)	Multiple mutations, 17p (p53), 5q (APC), 9p (p16)

FAP = familial adenomatous polyposis coli; AFAP = attenuated familial adenomatous polyposis coli; HFAS = hereditary flat adenoma syndrome; HNPPC = hereditary non-polyposis colon cancer; UCAN = ulcerative colitis associated neoplasia.

Risk from ulcerative colitis

Several studies have indicated that patients with ulcerative colitis have a 2-8.2 relative risk of colorectal cancer compared with the normal population, accounting for about 2% of colorectal cancers. One of the factors influencing an individual's risk is duration of colitis—the cumulative incidence of colorectal cancer is 5% at 15 years and 8-13% at 25 years. The extent of disease is also important: patients with involvement of right and transverse colon are more likely to develop colorectal cancer (the relative risk in these patients is 15 compared with the normal population). Coexisting primary sclerosing cholangitis independently increases the relative risk of ulcerative colitis associated neoplasia (UCAN) by 3-15%. In addition, high grade dysplasia in random rectosigmoid biopsies is associated with an unsuspected cancer at colectomy in 33% of patients.

Box 2.2 Factors affecting risk of colorectal cancer in patients with ulcerative colitis

High risk
Long duration of disease (especially >10 years)
Extensive disease
Dysplasia
Presence of primary sclerosing cholangitis
Family history of colorectal cancer
Coexisting adenomatous polyp

Low risk
Short duration of disease (especially <10 years)
Proctitis only
No dysplasia
No primary sclerosing cholangitis
No family history of colorectal cancer
No coexisting adenomatous polyp

Molecular basis of adenoma carcinoma sequence and UCAN

Cancers arising in colitis versus those in adenomas
Important clinical and biological differences exist between the adenoma carcinoma sequence and ulcerative colitis associated neoplasia. Firstly, cancer in ulcerative colitis probably evolves from microscopic dysplasia with or without a mass lesion rather than from adenomas. Secondly, the time interval from the presence of adenoma to progression to carcinoma probably exceeds the interval separating ulcerative colitis associated dysplasia from ulcerative colitis associated neoplasia. Thirdly, patients with a family history of colorectal cancer (but not ulcerative colitis associated neoplasia) and who also have ulcerative colitis are at further increased risk, suggesting additive factors.

Box 2.3 Chromosomal and microsatellite instability

- Molecular alterations in colorectal cancer can be grouped into two broad categories: chromosomal instability (subdivided into aneuploidy and chromosomal alterations) and microsatellite instability
- As a consequence of these two phenomena, other specific genetic events occur at increased frequency
- These include inactivation of tumour suppressor genes by deletion or mutation, activation of proto-oncogenes by mutation, and dysregulated expression of diverse molecules, such as the cell to cell adhesion molecule E cadherin and mucin related sialosyl-Tn antigen

Chromosomal instability
Aneuploidy indicates gross losses or gains in chromosomal DNA and is often seen in many human primary tumours and premalignant conditions. It has been shown that aneuploid "fields" tend to populate the epithelium of patients with ulcerative colitis even in histologically benign colitis. These changes may occur initially in some cases by loss of one allele at a chromosomal locus (loss of heterozygosity) and may imply the presence of a tumour suppressor gene at that site. Loss of both alleles at a given locus (homozygous deletion) is an even stronger indicator of the existence of a tumour suppressor gene. Loss of heterozygosity occurs clonally in both the adenoma carcinoma sequence and ulcerative colitis associated neoplasia. Many of these loci are already associated with one or more known candidate tumour suppressor genes. These include 3p21 (β catenin gene), 5q21 (APC gene), 9p (p16 and p15 genes), 13q (retinoblastoma gene), 17p (p53 gene), 17q (BRCA1 gene), 18q (DCC and SMAD4 genes), and less frequently 16q (E cadherin gene).

The p53 gene locus is the commonest site demonstrating loss of heterozygosity. p53 is a DNA binding protein transcriptional activator and can arrest the cell cycle in response to DNA damage—hence its title "guardian of the genome." The effect of normal (wild-type) p53 is antagonised by mutation or by action of the antiapoptotic gene Bcl-2, which is significantly less frequently overexpressed in ulcerative colitis associated neoplasia than in the adenoma carcinoma sequence. Most mutations in p53 cause the protein to become hyperstable and lead to its accumulation in the nucleus.

A second tumour suppressor gene necessary for development of sporadic colorectal cancer is APC, which is

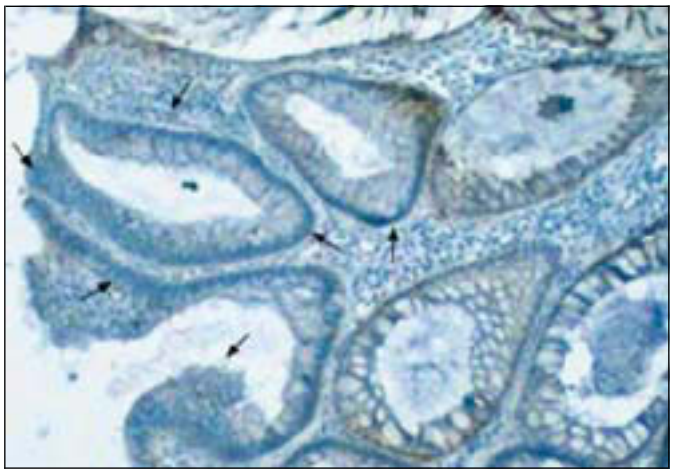


Figure 2.3 Downregulation of E cadherin (arrowed) within colonic adenoma. Normal membranous E cadherin staining (brown) is seen in non-dysplastic crypts in the right of the picture

inactivated in >80% of early colorectal cancers. Consequently this gene has been termed the "gatekeeper" for adenoma development as adenoma formation requires perturbation of the APC gene's function or that of related proteins such as catenins. An important function of the APC gene is to prevent the accumulation of molecules associated with cancer, such as catenins. Accumulation of catenins can lead to the transcription of the oncogene c-myc, giving a proliferative advantage to the cell. APC mutations occur later and are somewhat less common in ulcerative colitis associated neoplasia (4-27%) than in sporadic colorectal adenomas and carcinomas. Catenins also bind E cadherin, which functions as a tumour suppressor gene in the gastrointestinal tract. It is currently thought that mutated catenins may not bind to APC and thus accumulate.

Microsatellite instability

A further important category of alteration studied in the adenoma carcinoma sequence and ulcerative colitis associated neoplasia is microsatellite instability. This comprises length alterations of oligonucleotide repeat sequences that occur somatically in human tumours. This mechanism is also responsible for the germline defects found in heredity non-polyposis colon cancer. The incidence of microsatellite instability has been noted to be about 15% for adenomas and 25% for colorectal cancers overall. Microsatellite instability also occurs in patients with ulcerative colitis and is fairly common in premalignant (dysplastic) and malignant lesions (21% and 19% respectively). Indeed it has also been reported in "histologically normal" ulcerative colitis mucosa. It can therefore be considered to be an early event in the adenoma carcinoma sequence and in ulcerative colitis associated neoplasia.

E cadherin mutations are not commonly associated with the adenoma carcinoma sequence, but loss of heterozygosity and, rarely, missense mutations have been reported in 5% of ulcerative colitis associated neoplasia

Prognosis

The prognosis of colorectal cancer is determined by both pathological and molecular characteristics of the tumour.

Pathology

Pathology has an essential role in the staging of colorectal cancer. There has been a gradual move from using Dukes's classification to using the TNM classification system as this is thought to lead to a more accurate, independent description of the primary tumour and its spread. More advanced disease naturally leads to reduced disease-free interval and survival. Independent factors affecting survival include incomplete resection margins, grade of tumour, and number of lymph nodes involved (particularly apical node metastasis—main node draining a lymphatic segment).

Molecular biology

Reports on correlations between tumour genotype and prognosis are currently incomplete. However, analysis of survival data from patients with sporadic colorectal cancer and from those with colorectal cancer associated with familial adenomatous polyposis and hereditary non-polyposis colon cancer has not shown any reproducible significant differences between these groups. In premalignancy, however, the onset of p53 mutations in histologically normal mucosa in ulcerative colitis suggests that detection of such mutations may be a useful strategy in determining mucosal areas with a high risk of dysplastic transformation.

Table 2.2 Staging and survival of colorectal cancers

TNM classification	Modified Dukes's classification	Survival (%)
Stage 0—Carcinoma in situ		
Stage I—No nodal involvement, no metastases; tumour invades submucosa (T1, N0, M0); tumour invades muscularis propria (T2, N0, M0)	A	90-100
Stage II—No nodal involvement, no metastases; tumour invades into subserosa (T3, N0, M0); tumour invades other organs (T4, N0, M0)	B	75-85
Stage III—Regional lymph nodes involved (any T, N1, M0)	C	30-40
Stage IV—Distant metastases	D	<5

Further reading

- Oates GD, Finan PJ, Marks CG, Bartram CI, Reznek RH, Shepherd NA, et al. *Handbook for the clinico-pathological assessment and staging of colorectal cancer*. London: UK Co-ordinating Committee on Cancer Research, 1997.
- Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities inhuman cancers. *Nature* 1998;396:643-9.
- Tomlinson I, Bodmer W. Selection, the mutation rate and cancer: ensuring that the tail does not wag the dog. *Nature Med* 1999;5:11-2.
- Powell SM, Zilz N, Beazer-Barclay Y, Bryan TM, Hamilton SR, Thibodeau SN, et al. APC mutations occur early during colorectal tumorigenesis. *Nature* 1992;359:235-7.
- Jankowski J, Bedford F, Boulton RA, Cruickshank N, Hall C, Elder J, et al. Alterations in classical cadherins in the progression of ulcerative and Crohn's colitis. *Lab Invest* 1998;78:1155-67.
- Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990;61:759-67.
- Liu B, Parsons R, Papadopoulos N, Nicolaides NC, Lynch HT, Watson P, et al. Analysis of mismatch repair genes in hereditary non-polyposis colorectal cancer patients. *Nature Med* 1996;2:169-74.

This work was funded by the Cancer Research Campaign and the Medical Research Department of Veteran Affairs. Fiona K Bedford, Robert Allan, Michael Langman, William Doe, and Dion Morton provided useful comments.

3 The role of clinical genetics in management

T R P Cole, H V Sleightholme

Before 1990 the role of inherited factors in the aetiology of adult cancer was relatively poorly understood and aroused little interest among doctors and the public alike—although familial adenomatous polyposis (the autosomal dominant colon cancer syndrome) was an exception in this respect. In the past decade, however, interest has increased markedly. In the West Midlands, for example, familial cancer referrals constituted $<1\%$ (<20 cases) of all clinical genetic referrals in 1991, whereas now they represent over 30% of cases (>1000).

Despite the estimate that 5-10% of colorectal cancer has an inherited basis, only a small percentage of referred families have mutations in one of the currently identified genes. Furthermore, mutation studies are usually possible only if DNA is available from an affected patient, so molecular investigation will facilitate the management of only a small minority of cases. The remaining referrals must be managed with clinically derived strategies. This article discusses the clinical features and management of dominant colon cancer syndromes and provides referral guidelines and screening protocols for more common familial clusters.

Genetic counselling for families with a history of cancer requires a full and accurate family history. When possible, histological confirmation of the reported tumours should be obtained. It should then be possible to recognise the specific cancer syndromes. It is important to emphasise to families that however extensive the family history of cancer (unless present on both sides), a person will always have a greater than 50% chance of not developing that particular tumour. This may surprise but greatly reassure many families.

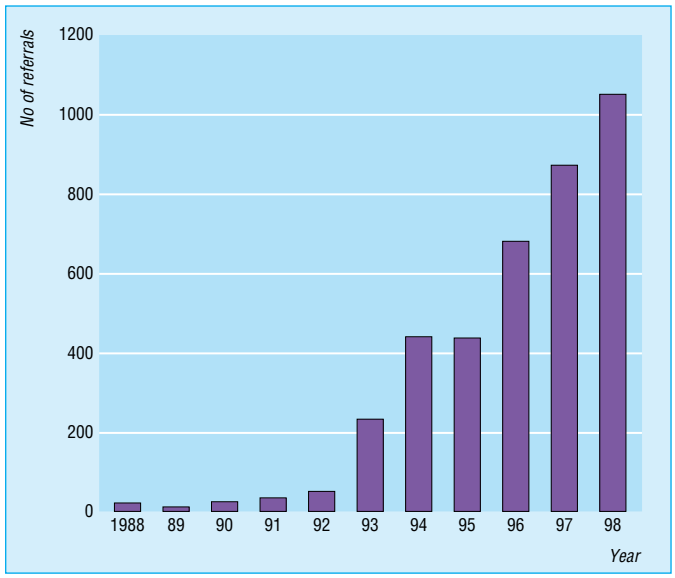


Figure 3.1 Number of referrals of patients with cancer (except familial adenomatous polyposis) to West Midlands regional clinical genetics service, 1988-98

Familial adenomatous polyposis

Familial adenomatous polyposis, previously called polyposis coli (or Gardner's syndrome if extra colonic manifestations were present), is the best recognised of the colorectal cancer syndromes but accounts for less than 1% of all colorectal cancers and has an incidence of 1 in 10 000. It is characterised by the presence of 100 or more tubovillous adenomas in the colon, with intervening microadenoma on histological examination. The mean age of diagnosis of polyps is during teenage years, and almost all of gene carriers have polyps by the age of 40. If these polyps are left untreated, malignant transformation is inevitable, with a mean age of colorectal cancer occurring during the patients' mid-30s, often with synchronous tumours.

This condition is an autosomal dominant disorder, with the offspring of affected individuals at 50% risk of being gene carriers. The diagnosis of familial adenomatous polyposis should always result in a careful and full evaluation of the family history. Wherever possible, parents should have at least one colonoscopy, irrespective of age. In most cases without a family history, parental examination will be negative and the proband will probably be one of 30% of cases that represent new mutations. The siblings of all probands, however, should be offered annual flexible sigmoidoscopy up to the age of 40 or until proved to be non-gene carriers.

The cloning of the causative gene (APC) on chromosome 5 in 1991 dramatically changed the management of familial adenomatous polyposis. If DNA is available from an affected

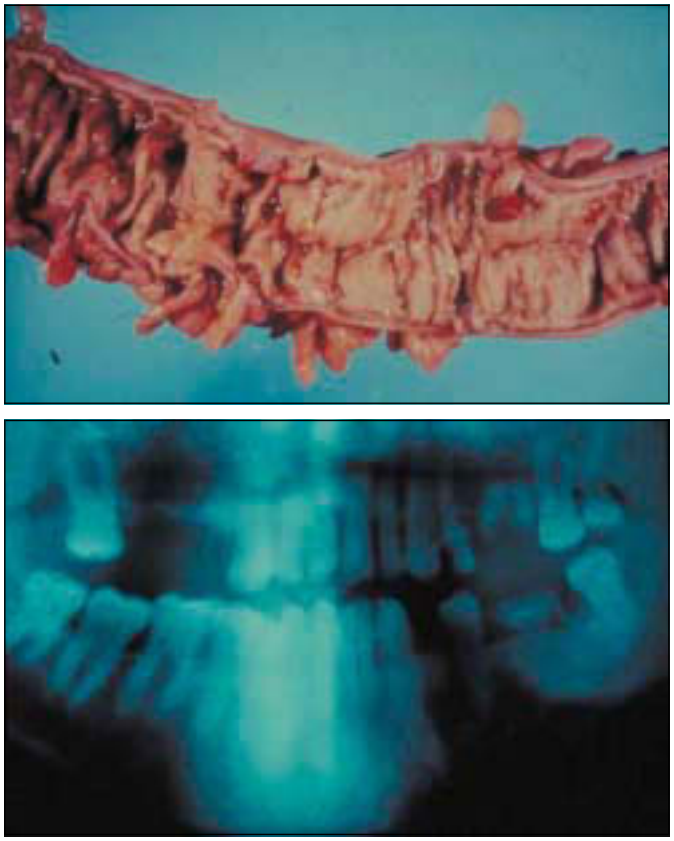


Figure 3.2 Features of familial adenomatous polyposis: colon with multiple polyps (top) and jaw cysts (bottom)

individual, mutation detection is possible in about 70% of families. In these families first degree relatives should be offered predictive testing with appropriate genetic counselling. In families with no identified mutation, linkage studies to identify the "high risk" chromosome 5 are possible in many cases. Non-gene carriers should be reassured and surveillance stopped. Gene carriers should be offered annual flexible sigmoidoscopy from the age of 12. Once several polyps have been identified, the timing and type of surgery available should be discussed (a sensitive issue in teenagers and young adults). The two most common options are ileal-rectal anastomosis with annual surveillance of the remaining rectal tissue; and ileal-anal anastomosis with reconstruction of a rectal pouch using terminal small bowel.

Molecular testing is usually offered to "at risk" children at age 10-14 before starting annual sigmoidoscopy. However, parental pressure for earlier testing (before the child can give consent) is not uncommon, and ongoing studies may help to clarify when to proceed with testing.

Cloning APC explained several clinical features and aided studies of genotypes and phenotypes. For example, congenital hypertrophy of the retinal pigment epithelium, an attenuated phenotype (that is, fewer than 100 polyps or late onset), and non-malignant but debilitating and potentially lethal desmoid disease each show an association with mutations in specific exon regions. The cloning also confirmed clinical findings that familial adenomatous polyposis and Gardner's syndrome were different manifestations of the same disease spectrum that could coexist within the same family.

With greater clinical awareness, regular surveillance, and the advent of molecular investigation, almost all colorectal cancer deaths in inherited cases of familial adenomatous polyposis can be avoided. Increased survival has revealed later complications, in particular periampullary or duodenal adenocarcinoma (present in about 12% of postcolectomy cases). Also important are aggressive desmoid disease and other rare malignant disease.

Chemoprophylactic approaches to reduce polyp growth (for example, aspirin and non-digestible starch) are the subject of multicentre trials.

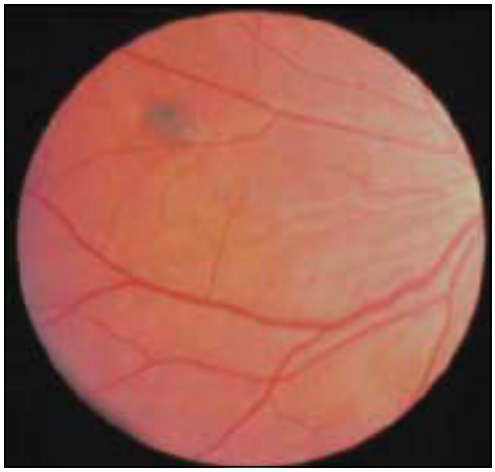


Figure 3.3 Congenital hypertrophy of the retinal pigmentary epithelium is a feature of familial adenomatous polyposis

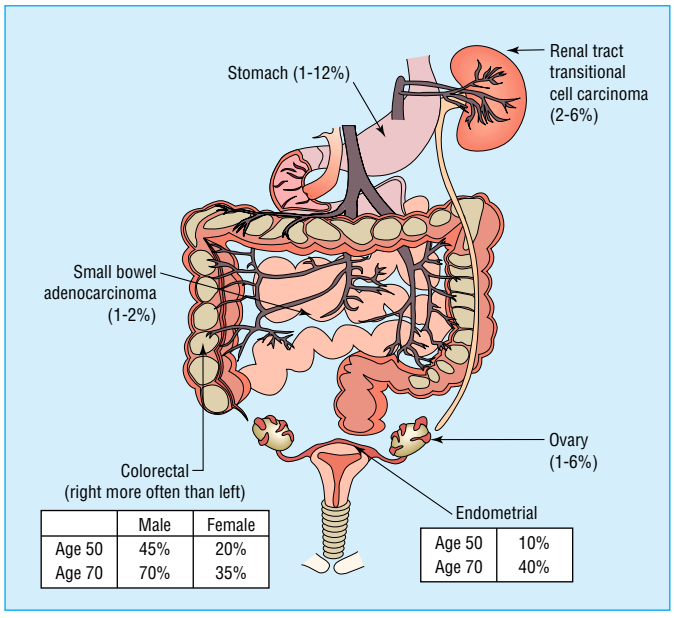
Box 3.1 Early and late extracolonic tumours in familial adenomatous polyposis

- Hepatoblastoma (early)
- Adrenal adenoma (early or late)
- Desmoid disease (early or late)
- Papillary thyroid cancer—females only (late)
- Periampullary carcinoma (late)

Hereditary non-polyposis colon cancer

Hereditary non-polyposis colon cancer (also known as Lynch syndrome) became more widely recognised about 30 years ago in families manifesting mainly colorectal cancer segregating in an autosomal dominant fashion. Many families also exhibit extracolonic tumours, particularly gynaecological, small bowel, or urinary tract carcinomas, and these became known as Lynch type 2 to distinguish them from site specific colorectal cancers, designated Lynch type 1. The subsequent name change to hereditary non-polyposis colon cancer is potentially misleading as many gene carriers will develop a small number of tubovillous adenomas, but not more than 100, as seen in familial adenomatous polyposis. The proportion of colorectal cancers due to hereditary non-polyposis colon cancer is controversial, and estimates range from 1% to 20%; most observers, however, suggest about 2%.

The diagnosis of hereditary non-polyposis colon cancer is made on the family history as the appearance of the bowel, unlike in familial adenomatous polyposis, is not diagnostic. To improve the recognition of hereditary non-polyposis colon cancer, diagnostic criteria were devised in Amsterdam in 1991 and were subsequently modified to include non-colonic tumours.



	Male	Female
Age 50	45%	20%
Age 70	70%	35%

Age 50	10%
Age 70	40%

Figure 3.4 Site of tumours and frequency in hereditary non-polyposis colon cancer (upper figures in ranges may be overestimates owing to ascertainment bias)

Box 3.2 Modified Amsterdam criteria

- Three or more cases of colorectal cancer in a minimum of two generations
- One affected individual must be a first degree relative of the other two (or more) cases
- One case must be diagnosed before age 50
- Colorectal cancer can be replaced by endometrial or small bowel adenocarcinoma
- Familial adenomatous polyposis must be excluded

In 1994 the first of the genes for hereditary non-polyposis colon cancer (hMSH2 on chromosome 2) was cloned, and since then four further genes have been identified; all are mismatch repair genes. If both copies of the genes are mutated, as postulated in Knudson's "two hit" hypothesis, that cell and all its daughter cells are missing a vital mechanism for repair of DNA in somatic tissue. Failure to repair mutations in tumour suppressor genes will in some instances result in initiation of the adenoma carcinoma sequence. Molecular studies showed that about 30% of colorectal cancers with early onset (under age 35) are due to the mismatch repair genes, and the typical age of onset and the spectrum of tumours in families with hereditary non-polyposis colon cancer were revised.

The limited available evidence suggests that screening for colorectal cancer in hereditary non-polyposis colon cancer is beneficial. In 1999 Vasen et al published figures showing clinical benefit and cost effectiveness of screening in hereditary non-polyposis colon cancer after a Finnish study reporting reduced morbidity and mortality in a non-randomised observational study of colonoscopy versus no screening.

The optimal surveillance frequency has not yet been defined in families with hereditary non-polyposis colon cancer. The method of choice, however, is colonoscopy rather than flexible sigmoidoscopy as 80% of cancers are proximal to the rectum (compared with only 57% in sporadic colorectal cancer). The screening interval should be at most three years and probably every 18-24 months in gene carriers. Failure to reach the caecum should be followed by barium enema examination, although surveillance using radiological techniques should probably be used sparingly owing to the theoretical mutagenic consequences in patients with DNA repair defects. Patients should understand that the strategy of colonoscopy is the removal of polyps and prevention of tumours or early diagnosis, but that complete prevention is impossible.

No large series of patients fulfilling the Amsterdam criteria has a mutation detection rate >70%. The figure is much lower for families that do not meet the criteria described here. Case selection using tumour DNA instability or immunohistochemical studies can improve mutation detection rates

- **Screening of other organ systems has not yet been proved beneficial**
- **It is prudent to screen for gynaecological tumours in mutation positive families, irrespective of family history, as 40% of female gene carriers develop endometrial carcinomas**
- **If tumours have been identified in the gynaecological or urinary tract in the family, surveillance should be offered (see the West Midlands guidelines)**

Box 3.3 West Midlands site specific screening strategies in hereditary non-polyposis colon cancer

Colorectal (all cases)—colonoscopy every two years at age 25-65
*Endometrial**—annual pipelle biopsy (suction curettage) and ultrasound at age 30-65
Ovarian†—annual transvaginal ultrasound and serum Ca125 concentration at age 30-65
Transitional cell carcinoma in the urinary tract—annual haematuria test at age 25-40; annual urine cytology at age 40-65 (with or without cystoscopy every one to two years); annual renal ultrasound at age 40-65

*Families with history of endometrial cancer and mutation positive families.
†Families with history of ovarian cancer.

Familial clusters with no recognisable single gene disorder

Families whose cancers do not meet the diagnostic criteria of familial adenomatous polyposis, hereditary non-polyposis colon cancer, or rarer colorectal cancer syndromes (such as syndromes related to the PTEN gene, Turcot's syndrome, Peutz-Jeghers syndrome, or juvenile polyposis) make up the largest and most difficult group of patients requesting management. There is rarely any indication of the aetiological basis of the cluster of colorectal cancer, and many instances will be coincidental occurrences. Other tumours clusters may be due to common environmental exposures, the effect of multiple low penetrance genes, or an interaction of both these elements. The risk of colorectal cancer may be assessed with empirical risk figures. These figures are estimates, however, and do not take into account factors such as the number of unaffected relatives. Further enquiry is usually justified if features such as multiple relatives with the same tumour or early onset of tumours are present in a family.

Concerns about not having precise risk figures may be misguided as many patients have difficulty interpreting risk figures and are often requesting only general guidance on risk and a discussion of management options. Many different screening protocols have been suggested in the past, however, and the lack of consistency and long term audit in these families is a problem.

To manage familial cancer in the West Midlands (population 5.2 million), a protocol has been developed that builds on the Calman-Hine model for cancer services and maximise the role

Table 3.1 Lifetime risk of colorectal cancer in first degree relatives of patient with colorectal cancer (from Houlston et al, 1990)

Population risk	1 in 50
One first degree relative affected (any age)	1 in 17
One first degree and one second degree relative affected	1 in 12
One first degree relative affected (age <45)	1 in 10
Two first degree relatives affected	1 in 6
Autosomal dominant pedigree	1 in 2

Box 3.4 Four pointers to recognition of familial cancer clusters

- High frequency of the same tumour in the family
- Early age of onset of tumours
- Multiple primary tumours
- Recognised associations—for example, colorectal and endometrial adenocarcinomas

of primary care. The protocol provides clear inclusion and screening guidance for cancer units and centres. This has promoted a consistency of management in and between families and is now allowing data collection for audit.

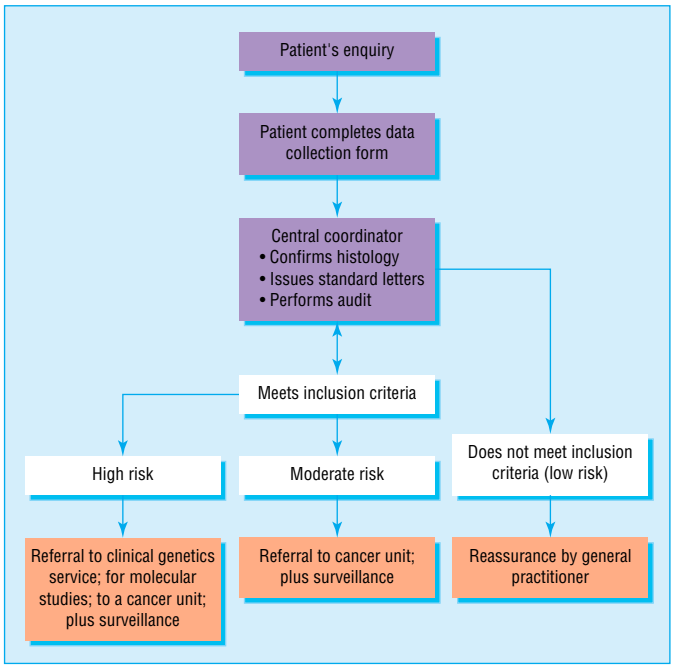


Figure 3.5 Algorithm for West Midlands family cancer strategy

It may be wise for general practitioners to use a reactive approach to patient enquiries until evidence exists to support a proactive approach. In the West Midlands, patients requesting advice are asked to complete a data collection sheet at home. This form and the inclusion criteria are available at www.bham.ac.uk/ich/clingen.htm. Completion of the form in the patient's own time, at home, facilitates discussion with relatives to clarify the relevant information and has saved time in primary care if a referral is required.

After histological confirmation in suspected familial cases, the data are evaluated centrally to identify high risk families requiring specialist investigation or referral to a cancer unit.

In a pilot study (population 200 000) the protocol reduced referrals from primary care by 50%, with a greater reduction in screening owing to a fall in low risk referrals to cancer units. This was associated with an increased referral rate for high risk referrals to clinical genetics departments. Central coordination prevents unnecessary investigations for different branches of any one family and facilitates audit.

Reports from general practitioners and patients suggest that individuals at no or minimal increased risk of cancer avoid unnecessary outpatient appointments and screening and for the most part are reassured by standardised protocols with explanations on the data collection forms. Such systems need to be studied further but seem to be preferable to continuing the current exponential rise of ad hoc responses from individual clinicians without long term audit.

Table 3.2 West Midlands inclusion and screening criteria for a family history of colorectal cancer

Inclusion criteria	Screening method	Age range (years) for surveillance
One first degree relative aged >40	Reassure, plus general advice on symptoms	Not applicable
One first degree relative aged <40	Colonoscopy every 5 years; appointment at local screening unit	25-65, or 5 years before tumour if later
Two first degree relatives average age >70	Reassure, plus general advice on symptoms	Not applicable
Two first degree relatives average age 60-70	Single colonoscopy; appointment at local screening unit	About 55
Two first degree relatives average age 50-60	Colonoscopy every 5 years; appointment at local screening unit	35-65
Two first degree relatives average age <50	Colonoscopy every 3-5 years; referral to genetics unit	30-65
Three close relatives but not meeting Amsterdam criteria	Colonoscopy every 3-5 years; referral to genetics unit	Begin age 30-40; stop at 65
Three close relatives meeting Amsterdam criteria	Colonoscopy every 2 years; referral to genetics unit	25-65
Familial adenomatous polyposis	Annual sigmoidoscopy; referral to genetics unit	12-40

Further reading

- Foulkes W. A tale of four syndromes: familial adenomatous polyposis, Gardner syndrome, attenuated APC and Turcot syndrome. *Q J Med* 1995;88:853-63.
- Vasen HF, van Ballegooijen, Buskens E, Kleibeuker JK, Taal BG, Griffioen G, et al. A cost-effectiveness analysis of colorectal screening of hereditary nonpolyposis colorectal carcinoma gene carriers. *Cancer* 1998;82:1632-7.
- Burke W, Petersen G, Lynch P, Bokin J, Daly M, Garber J, et al. Recommendations for follow-up care of individuals with an inherited predisposition to cancer. 1: Hereditary nonpolyposis colon cancer. *JAMA* 1997;277:915-9.
- Houlston RS, Murday V, Harocopos C, Williams CB, Slack J. Screening and genetic counselling for relatives of patients with colorectal cancer in a family cancer clinic. *BMJ* 1990;301:366-8.
- Hodgson SV, Maher ER, eds. *A practical guide to human cancer genetics*. Cambridge: Cambridge University Press, 1999.

Professor Eamonn Maher gave helpful comments and support.

4 Screening

John H Scholefield

Colorectal cancer is the third commonest malignancy in the United Kingdom, after lung and breast cancer, and kills about 20 000 people a year. It is equally prevalent in men and women, usually occurring in later life (at age 60-70 years). The incidence of the disease has generally increased over recent decades in both developed and developing countries. Despite this trend, mortality in both sexes has slowly declined. This decrease in mortality may reflect a trend towards earlier diagnosis—perhaps as a result of increased public awareness of the disease.

Surgery remains the mainstay of treatment for colorectal cancer, but early diagnosis makes it more likely that the tumour can be completely resected and thereby improves the chance of cure

Why screen?

Most colorectal cancers result from malignant change in polyps (adenomas) that have developed in the lining of the bowel 10-15 years earlier. The best available evidence suggests that only 10% of 1cm adenomas become malignant after 10 years. The incidence of adenomatous polyps in the colon increases with age, and although adenomatous polyps can be identified in about 20% of the population, most of these are small and unlikely to undergo malignant change. The vast majority (90%) of adenomas can be removed at colonoscopy, obviating the need for surgery. Other types of polyps occurring in the colon—such as metaplastic (or hyperplastic) polyps—are usually small and are much less likely than adenomas to become malignant.

Colorectal cancer is therefore a common condition, with a known premalignant lesion (adenoma). As it takes a relatively long time for malignant transformation from adenoma to carcinoma, and outcomes are markedly improved by early detection of adenomas and early cancers, the potential exists to reduce disease mortality through screening asymptomatic individuals for adenomas and early cancers.

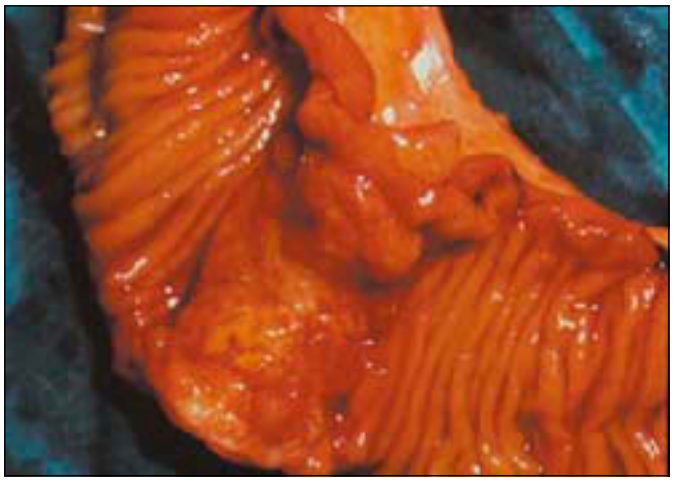


Figure 4.1 Colon cancer

Which screening test for population screening?

Education about bowel cancer is poor. A survey in 1991 showed that only 30% of the British population were aware that cancer of the bowel could occur. Such ignorance only adds to the difficulties of early detection for this form of cancer.

For a screening test to be applicable to large populations it has to be inexpensive, reliable, and acceptable. Many different screening tests for detecting early colorectal cancer have been tried. The simplest and least expensive is a questionnaire about symptoms, but this has proved predictably insensitive and becomes reliable only when the tumour is relatively advanced. Digital rectal examination and rigid sigmoidoscopy both suffer from the limitation that they detect only rectal or rectosigmoid cancers and are unpleasant and invasive.

Flexible sigmoidoscopy

Flexible sigmoidoscopy can detect 80% of colorectal cancers as it examines the whole of the left colon and rectum. A strategy of providing single flexible sigmoidoscopy for adults aged 55-65 years—with the aim of detecting adenomas—may be cost effective. A multicentre trial of this strategy for population screening is currently under evaluation.

Although flexible sigmoidoscopy is more expensive than rigid sigmoidoscopy, it is generally more acceptable to patients (it is less uncomfortable) and has much higher yield than the

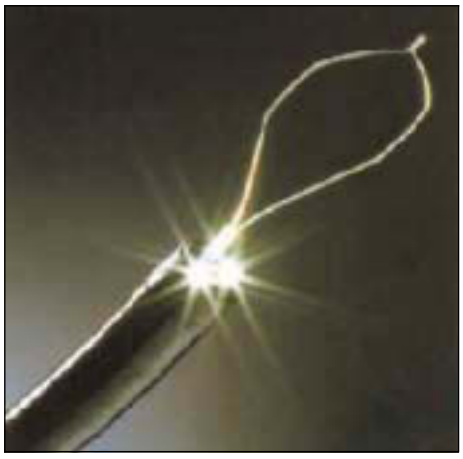


Figure 4.2 Flexible sigmoidoscope: used for endoluminal visualisation and therapeutic removal of adenomas

rigid instrument. Many nurses are now trained to perform flexible sigmoidoscopy, making potential screening programmes using this technique more cost effective. In a population screening programme, uptake of the offer of the screening test is crucial. Uptake is likely to be around 45%, and, of these, 6% will subsequently need full colonoscopy. The effect that this will have on the incidence of and mortality from colorectal cancer is uncertain until the completion of the multicentre trial in 2003.

Colonoscopy

Colonoscopy is the gold standard technique for examination of the colon and rectum, but its expense, the need for full bowel preparation and sedation, and the small risk of perforation of the colon make it unacceptable for population screening. Colonoscopy is, however, the investigation of choice for screening high risk patients (those at risk of hereditary non-polyposis colon cancer or with longstanding ulcerative colitis).

Barium enema

Barium enema, like colonoscopy, examines the whole colon and rectum, and, although it is cheaper and has a lower complication rate than colonoscopy, it is invasive and requires full bowel preparation. Whereas colonoscopy may be therapeutic (polypectomy), barium enema does not allow removal or biopsy of lesions seen. There are no population screening studies using barium enema.

Faecal occult blood tests

Faecal occult blood tests are the most extensively studied screening tests for colorectal cancer. These tests detect haematin from partially digested blood in the stool. Their overall sensitivity for colorectal neoplasia is only 50-60%, though their specificity is high. In screening studies of faecal occult blood tests, individuals are invited to take two samples from each of three consecutive stools. Compliance is around 50-60%, but with population education this might be improved. Individuals with more than four out of six positive tests (about 2% of participants) need colonoscopy.

Several large randomised studies have shown that screening with faecal occult blood testing is feasible, and two studies have shown that such screening reduces the mortality from colorectal cancer. In a study in Nottingham, for every 100 individuals with a positive test result, 12 had cancer and 23 had adenomatous polyps. The cancers detected at screening tended to be at an earlier stage than those presenting symptomatically (Dukes's A classification: 26% screen detected *v* 11% in controls). The disadvantage of screening with faecal occult bloods is its relatively low sensitivity—a third to a half of cancers will be missed on each round of screening. The Nottingham data suggest that screening every two years detects only 72% of cancers. This could be improved by testing annually and using more sensitive immunologically based faecal occult blood tests.

Who should be screened?

Although about 20% of the population will develop adenomatous polyps, only 5% of these will develop colorectal cancer. This equates to a 1 in 20 lifetime risk for colorectal cancer. The cancer occurs most often in the age group 65-75 years, but for adenomas the peak incidence is in a slightly earlier age group (55-65 years). Thus population screening for colorectal cancer should target both these age groups.

In addition, some people inherit a much higher susceptibility to colorectal cancer. Some inherit a well

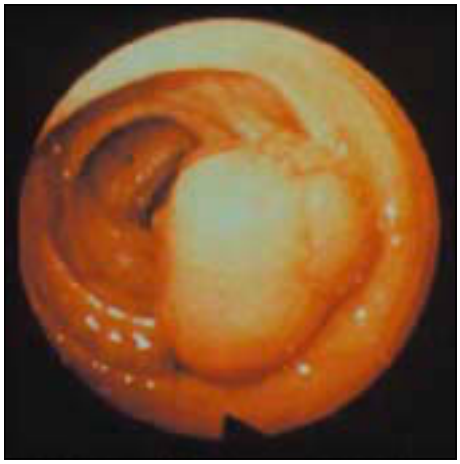


Figure 4.3 Colonoscopic view of colonic adenoma (about 1.5 cm diameter)

Figure 4.4 Double contrast barium enema showing carcinoma of sigmoid colon

Box 4.1 CT colography

- CT (computed tomographic) colography—virtual colonoscopy—is a new radiological technique that may have a role in population screening
- Although it requires full bowel preparation, highly expensive computed tomography scanners, and computing facilities, it is minimally invasive, and views of the whole colon can be obtained in five minutes
- Preliminary data suggest that this technique is as sensitive as colonoscopy or barium enema for detecting large polyps and cancers
- As yet, no trials of CT colography in population screening have been published
- CT colography has the potential to be cost effective and to reduce the need for colonoscopy in population screening

recognised single gene disorder, such as familial adenomatous polyposis or hereditary non-polyposis colon cancer, whereas most inherit an undetermined genetic abnormality. These people tend to develop colorectal cancer before the age of 50, and therefore screening in this high risk population needs to be tailored to each individual's risk pattern. They may also be at risk for cancers at other sites, and screening for ovarian, breast, and endometrial cancers may be appropriate in some of these cases. The advice of clinical geneticists in these cases can be invaluable.

Cost effectiveness of screening

If screening for colorectal cancer is to be acceptable to healthcare providers it must be shown to be cost effective. Estimates of the cost of screening for colorectal cancer range from £1000 to £3000 per life year saved, depending on the screening technique used. The cost of using faecal occult blood testing would be the lowest—similar to estimates for breast cancer screening.

Cost estimates are associated with several unknown factors. The factors that cause greatest concern to those considering funding screening programmes are the cost of cancers missed and the potential damage caused to asymptomatic individuals by invasive procedures such as colonoscopy.

Potential harm from screening

Although it has been suggested that considerable anxiety and psychological morbidity may be caused by inviting populations to participate in screening for colorectal cancer, little evidence exists to substantiate this. Indeed in the Nottingham trial no longstanding psychological morbidity from the screening programme was found. Similarly, no evidence exists that screening for colorectal cancer leads to false reassurance from negative tests.

Complications from colonoscopy (perforation and haemorrhage), however, can occur. The incidence of these complications is around 1 in 2000 procedures, and complications usually occur in therapeutic colonoscopy (endoscopic polypectomy) rather than in diagnostic procedures. Mortality from such events is rare.

Conclusions

- Screening for colorectal cancer using faecal occult blood tests is feasible; increasingly compelling evidence shows that such programmes can save lives at a cost similar to that of the existing breast cancer screening programme
- Once-only flexible sigmoidoscopy presents a promising alternative to faecal occult blood screening, but conclusive data will not be available for about five years
- For a screening programme to operate in the United Kingdom, considerable investment in colonoscopy facilities and expertise would be needed
- Several countries, including the United States, have screening programmes that use faecal occult blood tests or once-only flexible sigmoidoscopy, or both of these procedures. The United Kingdom has undertaken a pilot study in three areas to determine the feasibility of delivering a practicable, population based screening programme

Box 4.2 Inherited risk of colorectal cancer

High risk
- Familial adenomatous polyposis
- Hereditary non-polyposis colon cancer

Medium risk
- One first degree relative with colorectal cancer presenting at <45 years
- Two or more first degree relatives with colorectal cancer

Low risk
- Only one first degree relative with colorectal cancer presenting at >55 years
- No family history of colorectal cancer

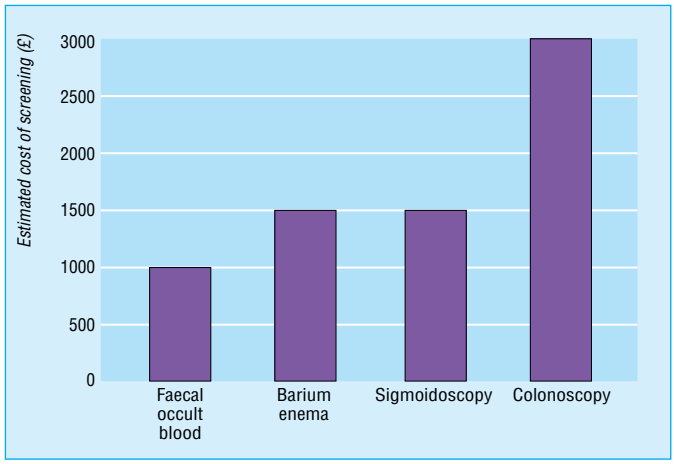


Figure 4.5 Estimates of costs for different methods of screening for colorectal cancer. Costs are based on biennial testing (faecal occult blood), testing at intervals of 5 years (barium enema and colonoscopy), or once-only testing (sigmoidoscopy)

Further reading

- Young GP, Rozen P, Levin B, eds. *Prevention and early detection of colorectal cancer*. London: Saunders, 1996.
- Kune G, ed. *Causes and control of colorectal cancer—a model for cancer prevention*. Boston, MA: Kluer Academic Publishers, 1996.
- Hardcastle JD, Chamberlain JO, Robinson MHE, Moss SM, Amar SA, Balfour TW, et al. Randomised controlled trial of faecal-occult-blood screening in colorectal cancer. *Lancet* 1996;348:1472-7.
- Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, et al. Reducing mortality from colorectal cancer by screening for faecal occult blood. *N Engl J Med* 1993;328:1365-71.
- Winawer SJ, Schottenfield D, Felhinger BJ. Colorectal cancer screening. *J Natl Cancer Inst* 1991;83:243-53.

The picture of the flexible sigmoidoscope is published with permission from Endoscopy Support Services.

5 The role of primary care

F D Richard Hobbs

Every general practitioner in the United Kingdom will on average see one new case of colorectal cancer each year. For most primary care doctors the most important contributions they make to the care of patients with colorectal cancer relate to early diagnosis of the condition (including the point of referral) and to palliation of symptoms in those with established disease. Further roles in the future primary care service are screening for colorectal cancer (possibly using faecal occult blood testing) and a greater involvement in monitoring patients after curative procedures.

As colorectal cancer is the sixth most common cause of mortality in the United Kingdom, a general practitioner will on average care for a patient dying from colorectal cancer every 18 months

Early diagnosis and referral guidelines

Early diagnosis of colorectal cancer is essential in view of the stage related prognosis. Three potential levels of delay occur in the diagnosis of the disease: delay by the patient in presenting to the general practitioner; delay in referral by the general practitioner to a specialist; and delay by the hospital in either establishing the diagnosis or starting treatment. Detrimental differences between England and Wales and the rest of western Europe in survival rates for colorectal cancer arise primarily in the first six months after diagnosis, suggesting that these differences relate to late presentations or delays in treatment.

Box 5.1 Guidelines for urgent referral of patients with suspected colorectal cancer based on symptoms presented*

These combinations of symptoms and signs, when occurring for the first time, should be used to identify patients for urgent referral (that is, within two weeks). Patients need not have all symptoms

All ages
- Definite, palpable, right sided, abdominal mass
- Definite, palpable, rectal (not pelvic) mass
- Rectal bleeding with change in bowel habit to more frequent defecation or looser stools (or both) persistent over six weeks
- Iron deficiency anaemia (haemoglobin concentration <110 g/l in men or <100 g/l in postmenopausal women) without obvious cause

Age over 60 years (maximum age threshold could be 55 or 50)
- Rectal bleeding persistently without anal symptoms (soreness, discomfort, itching, lumps, prolapse, pain)
- Change of bowel habit to more frequent defecation or looser stools (or both), without rectal bleeding, and persistent for six weeks

*Adapted from the NHS Executive's *Referral Guidelines for Suspected Cancer* (London: Department of Health, 2000)

Patients presenting with symptoms

Most patients developing colorectal cancer will eventually present with symptoms. Primary symptoms include rectal bleeding persistently without anal symptoms and change in bowel habit—most commonly, increased frequency or looser stools (or both)—persistently over six weeks. Secondary effects include severe iron deficiency anaemia and clear signs of intestinal obstruction. Clinical examination may show a definite right sided abdominal mass or definite rectal mass.

Unfortunately, many large bowel symptoms are common and non-specific and often present late. Recently published guidelines, however, make specific recommendations about which patients should be urgently referred—within two weeks—for further investigation in the NHS. The guidelines also indicate which symptoms are highly unlikely to be caused by colorectal cancer.

The risk of colorectal cancer in young people is low (99% occurs in people aged over 40 years and 85% in those aged over 60). In patients aged under 45, therefore, initial management will depend on whether they have a family history of colorectal cancer—namely, a first degree relative (brother, sister, parent, or child) with colorectal cancer presenting below the age of 55, or two or more affected second degree relatives. Patients aged under 45 presenting with alarm symptoms and a family history of the disease should also be urgently referred for further investigation.

In patients suspected of having colorectal cancer, referral should be indicated as urgent (with an appointment expected within two weeks); the referral letter should include any relevant family history and details about symptoms and risk factors. An increasing number of general practitioners will have direct access to investigations, often via a rapid access rectal bleeding clinic. The usual investigations needed will be flexible colonoscopy or barium enema studies.

Figure 5.1 The NHS Executive's *Referral Guidelines for Suspected Cancer*

Box 5.2 Symptoms associated with low risk of malignancy*

Patients with the following symptoms but with no abdominal or rectal mass are at very low risk of colorectal cancer
- Rectal bleeding with anal symptoms (soreness, discomfort, itching, lumps, prolapse, pain)
- Change in bowel habit to less frequent defecation and harder stools
- Abdominal pain without clear evidence of intestinal obstruction

*Adapted from *Referral Guidelines for Suspected Cancer*

In the absence of a family history of the disease, younger patients with a negative physical examination, including a digital rectal examination, can be initially treated symptomatically. If symptoms persist, however, patients should be considered for further investigation.

Patients with genetic predisposition

All patients registering with a practice for the first time should provide details of their medical history. Patients with a history of familial adenomatous polyposis should be referred for DNA testing after the age of 15. Familial adenomatous polyposis accounts for about 1% of cases of colorectal cancer, with the defect gene identified on chromosome 5. Patients with a positive result should enter a programme of surveillance with flexible sigmoidoscopy.

The second common genetic predisposition to colorectal cancer is hereditary non-polyposis colon cancer. This condition should be suspected in patients describing three or more cases of colorectal cancer (or andenocarcinoma of the uterus) within their family. Such patients should be referred for endoscopic screening at the age of 25. Genetic testing for this condition is currently not feasible.

In patients with a first degree relative with colorectal cancer aged under 45 or with two first degree relatives with the disease, the lifetime risk of the cancer rises to over 1 in 10. Such patients should be referred for lower endoscopy screening once they are 10 years younger than the age at which the disease was diagnosed in the youngest affected relative. An earlier article in this series gives more detail on the genetics of colorectal cancer.

Population screening in primary care

The United Kingdom currently has no national screening programme for colorectal cancer. Several studies in the United States and Europe have shown that screening with faecal occult blood testing will reduce the overall mortality of colorectal cancer by about 15%. Such testing is a fairly simple procedure: only two small samples from different sites of a stool need to be collected on each of three consecutive days. In the United States, the specimens are then normally hydrated, whereas research in the United Kingdom and Denmark advocates using dry samples. The latter technique results in a lower sensitivity, but higher specificity—desirable test performance characteristics for an asymptomatic population screening procedure.

Faecal occult blood testing is therefore a cheap and easy method of screening, with reasonable levels of acceptability to the population. The main disadvantages of this test are the low sensitivity—with about 40% of cancers missed by a single screen, leading to the need for frequent faecal occult blood tests—and the fact that bleeding tends to occur late in the development of the disease. Furthermore there are no direct studies to guide on the most cost effective method of establishing a national screening programme using faecal occult blood testing. However, evidence from the cervical screening programme suggests that general practice led "call/recall" programmes would have the greatest impact.

A large Medical Research Council trial is currently evaluating once-only flexible sigmoidoscopy as a method of screening patients aged 50-60 years. The results of this trial will not be available for several years.

The American Cancer Society recommends an annual digital rectal examination for people aged over 40, an annual faecal occult blood test for people aged over 50, and flexible sigmoidoscopy every three to five years for people aged over 50. More detail on screening for colorectal cancer appears in an earlier article in this series.

Most (85-90%) colorectal cancers arise in people with no known risk factors, so opportunistic asymptomatic screening is of little value in colorectal cancer

Box 5.3 Patients with iron deficiency

- Patients aged 45 and over presenting with iron deficiency anaemia should be investigated to determine the cause of anaemia
- This will normally require both upper and lower bowel endoscopy
- In patients aged under 45, the cause of the anaemia should also be established, although the likelihood of this being colorectal cancer is low

Figure 5.2 Haemoccult (SmithKline Beecham) has been the faecal occult blood test most often used in studies of the feasibility of screening for colorectal cancer

Table 5.1 Results from European population colorectal cancer screening trials using faecal occult blood testing kits (Haemoccult)

	Funen, Denmark (1985-95)	Nottingham, UK (1985-91)
Uptake (% of population screened)	67 (>92 in later rounds)	57 (range in general practices 29-74)
% of positive tests (range in rounds)	1-1.8 (n=215-261)	1.9-2.1 (n=837-924)
No of cases of colorectal cancer*	37/215, 25/261	83/837, 22/924
No of cases of adenomas (>10 mm)*	68/215, 56/261	311/837, 304/924
% predictive value for neoplasia	38-58	44-47
% predictive value for cancer	25-37	10-12 (17 for late responders)
% of patients with Dukes's A classification†:		
Intervention group	22	20
Control group	11	11
% of patients with Dukes's C and D classification†:		
Intervention group	39	46
Control group	47	52

*Funen: rounds 1 and 5; Nottingham: first screen and rescreen.
†$P < 0.01$ for intervention versus control, both in Funen trial and in Nottingham trial.

Managing patients with established disease

After confirmation of diagnosis, the role of the primary care doctor revolves around advice, support, possibly monitoring for recurrence, and palliative care. Some general practices are involved with home based chemotherapy, usually coordinated by specialist outreach nurses.

In the United Kingdom primary care does not currently have a formal role in monitoring for disease recurrence after curative treatments. Data on this option are limited (see a later article in this series) but suggest that such surveillance could be safely conducted in primary care. Ideally, this monitoring should be accompanied by adequate infrastructure and training in primary care, with good liaison between the practice and secondary (or tertiary) care.

Limited evidence from other types of shared care indicate that certain factors are likely to improve outcomes: structured and planned discharge policies; the use of shared (preferably patient held) cards that document patient information (disease progress and drug treatments, as a minimum); locally agreed guidelines specifying the appropriate follow up and delineating responsibilities; and access to rapid referral clinics. As with follow up in all chronic diseases, the more communication between doctors and with the patients (and their families), the better the quality of care.

Where appropriate, the doctor should also counsel patients on any possible familial risk and the need for genetic counselling of relatives. The primary care doctor may also advise patients with diagnosed colorectal cancer about practical considerations, including access to social security benefits. In the United Kingdom eligibility for attendance allowance may be immediately available in the exceptional circumstance of cancer with a short terminal prognosis of less than six months.

For some patients, especially those with rectal tumours, the diagnosis of cancer is also accompanied by the necessity for either colostomy or ileostomy. Such patients will often require further specialised support, and liaison between the primary care team and specialist stoma nurses is important.

As the disease progresses, management will shift towards palliative care. Ideally, this would be delivered jointly by the primary care team and specialist palliative care services, such as those based at a hospice or provided by Macmillan nurses. Few data exist to guide on the most effective models for palliative care in colorectal cancer. However, non-randomised studies have shown high satisfaction among patients when they are kept fully involved in understanding the progression of their disease and their treatment options, when shared care cards are used, and when home care teams are provided.

The main priorities in palliative care in colorectal cancer include the management of pain, jaundice, ascites, constipation, and nausea. The importance of attempting to correct these symptoms cannot be overstated: as much distress may be caused by constipation or nausea as by pain. Full explanations of signs such as jaundice are likely to be reassuring. Moreover, the advent of specialist home care teams (with access to specialist equipment—such as bed aids to preserve pressure areas or syringe drivers for pain control) and skilled counsellors for patients and their families, enables virtually all patients who wish it to remain at home.

Such an option is further enhanced by relief admission—when necessary for the patient or the family—to specialist palliative care wards or, more likely, to a hospice. In the United Kingdom only a minority of patients with colorectal cancer currently die from their disease in hospital or in a hospice.

Figure 5.3 Macmillan nurses have an important role in community palliative care, liaising with both professionals and patients

Further reading

- Carter S, Winslet M. Delay in the presentation of colorectal cancer: a review of causation. *Int J Colorectal Dis* 1998;13(1):27-31.
- Crossland A, Jones R. Rectal bleeding: prevalence and consultation behaviour. *BMJ* 1995;311:486-8.
- Curless R, French J, Williams G, James O. Comparison of gastro-intestinal symptoms in colorectal carcinoma patients and community control to the respect of age. *Gut* 1994;35:1267-70.
- Fijten GH, Starmans R, Muris JW, Schouten HJ, Blijham GH, Knottnerus J. Predictive value of signs and symptoms for colorectal cancer in patients with rectal bleeding in general practice. *Fam Pract* 1995;12:279-86.
- Hardcastle JD, Chamberlain JO, Robinson MHE, Moss SM, Amar SS, Balfour TW, et al. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *Lancet* 1996;348:1472-7.
- Hobbs FDR, Cherry RC, Fielding JWL, Pike L, Holder R. Acceptability of opportunistic screening for occult gastrointestinal blood loss. *BMJ* 1992;304:483-6.
- Kronborg O, Olsen J, Jorgensen O, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult blood test. *Lancet* 1996;348:1467-71.
- Spurgeon P, Barwell F, Kerr D. Waiting times for cancer patients in England after general practitioners' referrals: retrospective national survey. *BMJ* 2000;320;838-9.
- St John DJ, McDermott FT, Hopper JL, Debney EA, Johnson WR, Hughes ES. Cancer risk in relatives of patients with common colorectal cancer. *Ann Intern Med* 1993;118:785-90.

The photograph of the Macmillan nurses is published with pemission from Macmillan Cancer Relief.

6 Primary treatment—does the surgeon matter?

Colin McArdle

The dominant factor contributing to the relatively poor prognosis for colorectal cancer is the advanced stage of the disease at the time of initial presentation: up to a third of patients have locally advanced or metastatic disease, which precludes surgical cure. Even in the patients who undergo apparently curative resection, almost half die within five years.

In the west of Scotland, for example, about a third of 1842 patients presenting with colorectal cancer to seven hospitals between 1991 and 1994 presented as emergencies. Potentially curative resection was achieved in about 70% of patients presenting electively; the curative resection rate was lower in those presenting as emergencies. Five per cent of patients admitted for elective surgery and 13% of those admitted as emergencies died. Almost 60% of elective patients survived two years, compared with 44% of patients admitted as emergencies. These results are typical of population based studies in the United Kingdom.

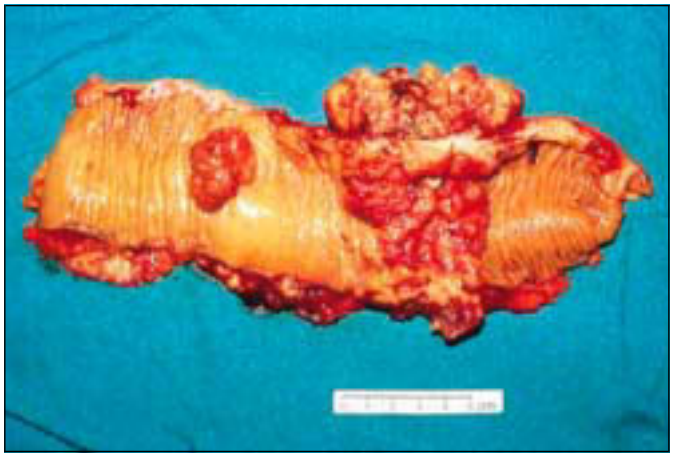


Figure 6.1 Colorectal adenoma and tumour—does a patient's survival depend on which surgeon operates?

Variation among surgeons

Most surgeons acknowledge that the incidence of postoperative complications varies widely among individual surgeons. It is now almost 20 years since Fielding and his colleagues in the large bowel cancer project drew attention to differences in anastomotic leak and local recurrence rates after resection for large bowel cancer.

In the original Glasgow Royal Infirmary study, which was conducted in the 1980s, similar differences in postoperative morbidity and mortality were noted. Furthermore, after apparently curative resection, survival at 10 years varied threefold among surgeons.

One might argue that these are historical data and therefore bear little relevance to the current situation. In the current west of Scotland study, however, although overall 33% of patients presented as emergencies, the proportion varied among hospitals from 24% to 41% and among surgeons from 10% to 50%.

Similarly, the proportion of patients undergoing curative resection varied among surgeons from 45% to 82%; postoperative mortality, in patients presenting electively, also varied, from 0% to 17%. Several out of the 16 surgeons studied performed less well than their colleagues.

Several factors apart from the individual surgeon's skill might influence these measurements of immediate and long

Table 6.1 Presentation, type of surgery, and postoperative mortality, by hospital and surgeon (n=1842), west of Scotland study. Values are percentages

	All (mean)	Hospital (range)	Surgeon (range)
Emergency admission	33	24-41	10-50
Dukes's classification A or B	49	43-56	29-68
Curative resection	68	63-75	45-82
Palliative resection	25	15-29	11-48
Postoperative mortality:			
Elective	5	0-7	0-17
Emergency	13	9-24	4-38

Table 6.2 Variation in outcome, by surgeon, after curative resection (n=338)

	Overall rate (%)	Range among surgeons (%)
Anastomotic leak	9	0-25
Local recurrence	11	0-21
Postoperative mortality	6	0-20
Survival (10 years)	41	20-63

Data are from the original Glasgow Royal Infirmary study (McArdle et al, *BMJ* 1991;302:1501-5)

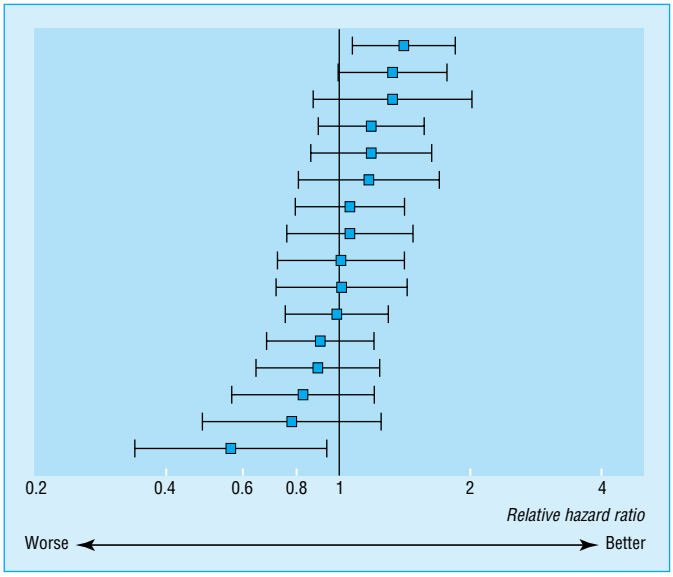


Figure 6.2 95% confidence intervals for relative risk of outcome, for all resections, by surgeon (n=16), west of Scotland study

term outcome: case mix; surgical philosophy; assessment of cure; quality of pathological reporting; other prognostic factors; small numbers (see box). Despite these factors it seems likely that the differences in the immediate postoperative morbidity and mortality observed among surgeons in the above studies are genuine. There have now been several analyses of immediate outcome after colorectal cancer surgery, and in each study, the results have been broadly comparable.

Effect of volume of surgery

Two explanations are possible for the differences in outcome among surgeons—namely, the number of patients treated by individual surgeons and whether these surgeons are specialists.

Although good evidence exists for other types of surgery that volume of work is important, in colorectal cancer convincing evidence that volume affects outcome is lacking. In the Lothian and Borders study, 5 of 20 consultants were responsible for 50% of the rectal cancer procedures. These five surgeons had a significantly lower anastomotic leak rate, but this may reflect specialisation rather than volume of work. In the German multicentre study, a group of surgeons with low work volume and performing only a few rectal cancer procedures had local recurrence rates well within the range of results obtained by individual surgeons with high work loads. Furthermore, in a recent analysis of outcome in 927 patients treated in the Manchester area, after correction for non-prognostic variables no relation between volume and outcome was noted.

Role of specialisation

The question of specialisation is more complex. Clearly rectal cancer surgery represents a greater technical challenge than colonic surgery. It therefore seems reasonable to expect—but it is remarkably difficult to show (largely because of the small numbers of patients treated by individual surgeons)—that specialist surgeons achieve better outcome. Analysis of outcome in almost 1400 patients with rectal cancer randomised in the Swedish preoperative radiotherapy studies, suggested that local recurrence and death rates were significantly lower in those patients operated on by surgeons with more than 10 years' experience as a specialist.

Perhaps the best information, however, comes from the Canadian study in which 683 patients with rectal cancer were treated by 52 different surgeons, five of whom were trained in colorectal surgery. These five surgeons performed 109 (16%) of the procedures. Independent of the type of training received by the surgeons, 323 procedures (47%) were performed by surgeons who each did fewer than 21 resections over the study period. Multivariate analysis showed that the risk of local recurrence was increased in patients treated both by surgeons not trained in colorectal surgery and by surgeons performing fewer than 21 resections. Similarly, disease specific survival was lower in the patients treated by these two groups of surgeons. These results suggest that both specialisation and volume may be important independent factors determining outcome.

Surgeons are currently under intense scrutiny, partly because readily available measures of outcome exist and partly because outcome seems to differ substantially among surgeons. The issues, however, are complex. Small numbers, annual accounting, and failure to take into account case mix, surgical intent, quality of staging, and prognostic factors may lead to inappropriate conclusions.

Box 6.1 Influences, apart from surgeon's skill, on immediate and long term outcome of colorectal surgery

Case mix
Non-specialist surgeons tend to have a high proportion of elderly patients, often with concomitant disease, who present as emergencies with advanced lesions; specialist surgeons may have fewer emergencies, with most patients being younger, fitter, and with less advanced disease

Surgical philosophy
Faced with the same problem, an aggressive surgeon might undertake radical surgery, thereby risking technical complications, in an attempt to improve quality and duration of life, whereas a conservative surgeon might opt for limited surgery, thereby minimising the risk of postoperative complications (but in doing so, he or she may compromise long term survival)

Assessment of cure
The decision on whether a resection is curative or palliative is often based on the surgeon's subjective impression at the time of laparotomy. In patients in whom the adequacy of resection was borderline an optimistic surgeon might believe a cure had been achieved, whereas a more pessimistic surgeon might believe that only palliation had been achieved

Quality of pathological reporting
Limited sampling might suggest that the lymph nodes and the lateral resection margins were clear of tumour, whereas more rigorous sampling might show the presence of more extensive disease. The resultant pathological stage migration might therefore alter expectation of outcome and lead to in inappropriate interpretation of the results

Other prognostic factors
Other factors—for example, socioeconomic deprivation—should be taken into consideration

Small numbers
Most surgeons at times have a cluster of patients who do less well than expected. This will vary from year to year. Any conclusion based on a small sample is likely to be misleading as it pertains to the individual surgeon

Table 6.3 Local recurrence and disease specific survival (n=683), according to specialisation and volume of work. Values are percentages

Training in colorectal surgery	Surgeons performing <21 resections (323 procedures)	Surgeons performing ≥21 resections (360 procedures)
No (n = 574):		
Local recurrence	44.6	27.8
Survival	39.2	49
Yes (n = 109):		
Local recurrence	21.1	10.4
Survival	54.5	67.3

Data are from the Canadian study (Porter et al, *Ann Surg* 1998;227:157-67)

Even if confounding variables are taken into account, some surgeons seem to be less competent than others, with poorer outcomes

Nevertheless, the results of the studies discussed here suggest that some surgeons are less competent than their colleagues and that these factors may compromise survival. Considerable effort and resources are currently being poured into large multicentre studies of adjuvant chemotherapy and radiotherapy in an effort to provide a marginal improvement in the survival of patients with colorectal cancer. If, by specialisation, the overall results of surgery could be improved—and evidence suggests that this is so—the impact on survival might be greater than that of any of the adjuvant therapies currently under study.

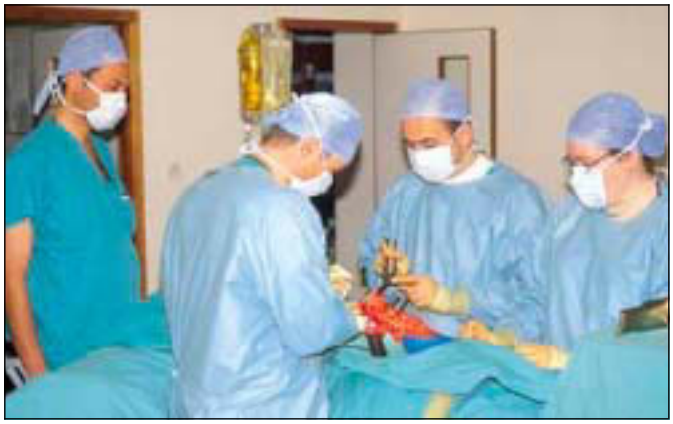


Figure 6.3 Outcome seems to differ substantially among surgeons performing colorectal surgery–specialisation rather than volume of work might be a way of improving overall outcome

Further reading

- Holm T, Johansson H, Cedermark B, Ekelund G, Rutqvist L-E. Influence of hospital and surgeon related factors and outcome after treatment of rectal cancer with or without pre-operative radiotherapy. *Br J Surg* 1997;87:657-63.
- McArdle CS, Hole D. Impact of variability among surgeons on post-operative morbidity and mortality and ultimate survival. *BMJ* 1991;302:1501-5.
- Parry JM, Collins S, Mathers J, Scott NA, Woodman CBJ. Influence of volume of work on the outcome of treatment for patients with colorectal cancer. *Br J Surg* 1998;86:475-81.
- Porter GA, Soskolne CL, Yakimets WW, Newman SC. Surgeon-related factors and outcome in rectal cancer. *Ann Surg* 1998;227:157-67.

7 Adjuvant therapy

Rachel S J Midgley, D J Kerr

Despite substantial improvements in surgical technique and postoperative care, colorectal cancer continues to kill 95 000 people in Europe alone each year.

Of the annual 150 000 newly diagnosed cases, about 80% have no macroscopic evidence of residual tumour after resection. More than half of patients, however, develop recurrence and die of their disease. This is a result of occult viable tumour cells that have metastasised before surgery and which are undetectable by current radiological techniques (the limit of detection of standard computed tomography is about $1cm^3$, equivalent to 10^9 cells).

Adjuvant treatment (chemotherapy and radiotherapy) has developed as an auxiliary weapon to surgery and is aimed at eradicating these micrometastatic cancer cells before they become established and refractory to intervention. As the presence of the primary tumour can exert an inhibitory influence on micrometastases, theoretically the removal of the tumour might stimulate growth of any residual cells, increasing the proliferating fraction and rendering them more susceptible to the cytotoxic effects of the widely used cytotoxic agent, fluorouracil.

It is reasonable to predict therefore that the earlier chemotherapy is started after surgery, the greater the potential benefit, although this has not yet been formally addressed in adjuvant trials. Implicit in this belief is a necessity for a multidisciplinary effort between surgeon, oncologist, and the community care team to provide seamless, streamlined cancer care for the individual patient.

Adjuvant—helpful, assisting, auxiliary (from Latin *ad* to, and *juvare* to help)

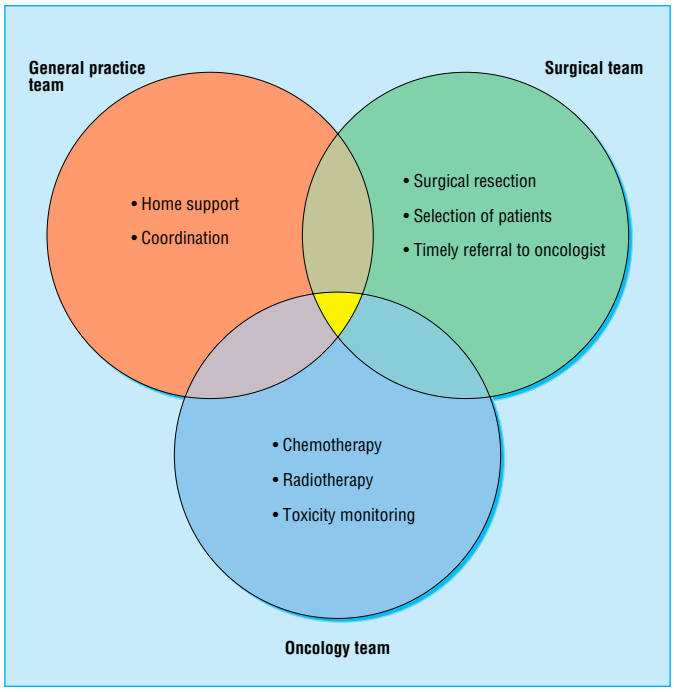


Figure 7.1 Optimising adjuvant therapy requires careful coordination between general practice, surgical, and oncology teams

Pharmacology of fluorouracil

Fluorouracil has remained the cornerstone chemotherapy for colorectal cancer for over 40 years. It is a prodrug that is converted intracellularly to various metabolites that bind to the enzyme thymidylate synthase, inhibiting synthesis of thymidine, DNA, and RNA. Increasing understanding of the molecular pharmacology of fluorouracil has led to the development of strategies to increase its efficacy.

The first strategy to be tested was coadministration with the immunostimulatory, antihelminthic drug levamisole, but despite promising early results, recent trials have not convincingly shown significant improvements in outcome compared with fluorouracil alone. In addition, no persuasive mechanism for the assumed synergism between fluorouracil and levamisole has been found.

In contrast, addition of folinic acid increases and prolongs the inhibition of the target enzyme (thymidylate synthase) and seems to confer improved clinical outcome compared with fluorouracil alone in advanced disease and when used in adjuvant therapy.

The side effects of chemotherapy based on fluorouracil vary according to the regimen (most commonly given as bolus intravenously daily for 5 days every 4 weeks or bolus weekly). They include nausea, vomiting, an increased susceptibility to infection, oral mucositis, diarrhoea, desquamation of the palms and soles, and, rarely, cardiac and neurological toxic effects.

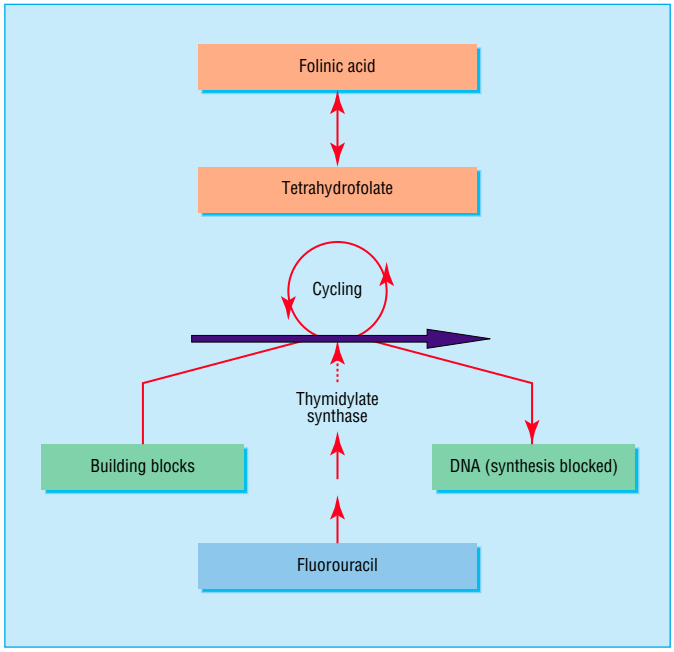


Figure 7.2 Intracellular metabolism and mechanism of action of fluorouracil and modulation by folinic acid

Established benefits of fluorouracil based adjuvant chemotherapy

Early adjuvant trials were retrospective and underpowered and failed to show any therapeutic benefit with respect to recurrence rate or survival. In 1990, however, the results of the intergroup trial were published. In this study 318 patients with stage B colorectal malignancy were randomised for surgical treatment alone or surgery followed by fluorouracil plus levamisole. In addition, 929 patients with stage C malignancy received surgery alone, surgery plus levamisole, or surgery plus fluorouracil and levamisole. For these patients there was a 33% reduction in the odds of death and a 41% decrease in recurrence among those treated with fluorouracil plus levamisole compared with surgery alone or surgery plus levamisole.

In contrast with levamisole, combining folinic acid with fluorouracil is pharmacologically rational, and documented benefit in advanced disease led to the logical extension of this combination into adjuvant therapy. Three large randomised adjuvant phase III trials produced confirmatory evidence of improved, disease-free survival at three years and improved overall survival in patients treated with fluorouracil plus folinic acid, with a 25-30% decrease in the odds of dying from colon cancer (or an absolute improvement in survival of 5-6% compared with controls).

Recently a meta-analysis of updated individual data from all unconfounded randomised studies of adjuvant chemotherapy (including the above three trials) has been undertaken (Colorectal Cancer Collaborative Group, unpublished). Overall, there was a 6-7% absolute improvement in survival with chemotherapy compared with surgery alone (SD 2.3, P = 0.01). The analysis advised that on current evidence the combination of fluorouracil plus folinic acid should be accepted as "standard" adjuvant chemotherapy for patients with Dukes's type C colon cancer.

Table 7.1 Results of three international randomised controlled trials of adjuvant chemotherapy (fluorouracil plus folinic acid *v* control) for patients with colon cancer. Values are percentage survival and P values

Trial	Disease-free survival	Overall survival
Overview of French, Italian, and Canadian trials (n = 1493)*	71 *v* 62 (< 0.0001)	83 *v* 78 (0.03)
Intergroup study (n = 309)†	74 *v* 58 (0.004)	74 *v* 63 (0.02)
NSABP C-03 trial (n = 1080)‡	73 *v* 64 (0.0004)	84 *v* 77 (0.003)

NSABP C-03 = national surgical adjuvant breast and bowel project—colon (protocol No 3).
*Fluorouracil plus high dose folinic acid *v* observation alone, 3 year follow up.
†Fluorouracil plus low dose folinic acid *v* observation alone, 5 year follow up.
‡Fluorouracil plus high dose folinic acid *v* methylCCNU, oncovin, or fluorouracil.

Box 7.1 Controversies still surrounding adjuvant therapy

- For how long should adjuvant therapy be continued, and what is the optimal dose of fluorouracil plus folinic acid?
- What is the role of adjuvant chemotherapy in lower risk groups?
- Is adjuvant therapy useful in rectal cancer?
- What is the role of new agents (eg irinotecan and oxaliplatin)?

Controversies in adjuvant therapy

Despite convincing evidence that adjuvant chemotherapy improves disease-free survival and overall survival in Dukes's type C colon cancer (an estimated six deaths prevented for 100 patients treated), several controversies surrounding the application of this form of treatment still exist.

Length of treatment and optimal dose of fluorouracil plus folinic acid

Lengthy adjuvant treatment has adverse effects on patients' quality of life as well as financial implications. A recent North American study, however, has shown that six months' treatment is as effective as 12 months'.

Determining the optimal dose is important: high dose folinic acid is 10 times as expensive as low dose. This issue has been addressed in the "certain" arm of the United Kingdom Co-ordinating Committee on Cancer Research's QUASAR ("quick and simple and reliable") trial (patients with Dukes's type C colon cancer). The trial uses the principle of randomising according to certain or uncertain indication: if, for a particular subgroup of patients the worth in receiving some form of adjuvant chemotherapy is definitely established from published randomised controlled trials (for example, patients with Dukes's type C colon cancer) then these patients are randomised to the certain indication arm (with a choice of different drugs and regimens); if, however, no definitive evidence exists of worth in a particular subgroup (for example, in patients with Dukes's type B colon cancer or with rectal

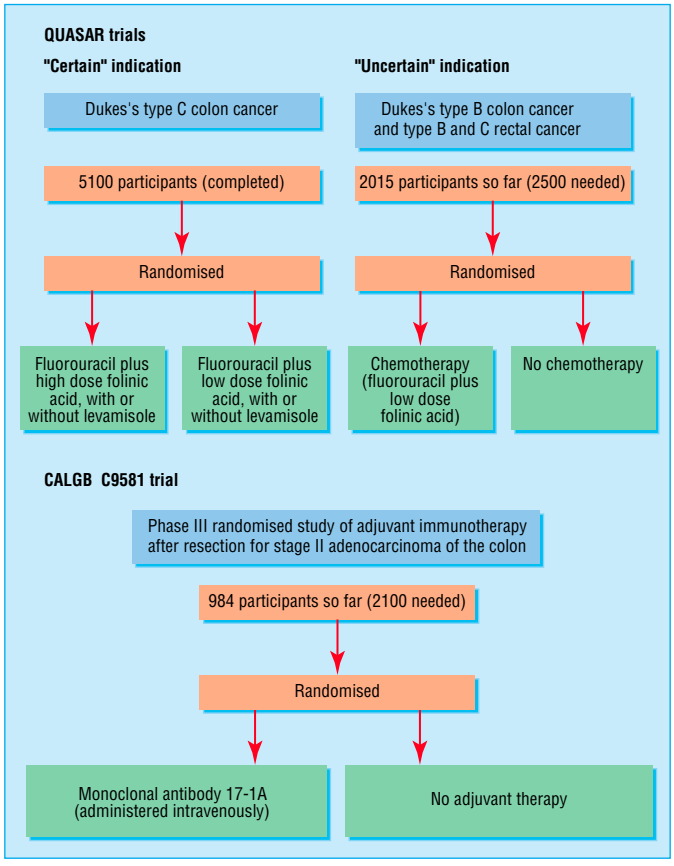


Figure 7.3 Ongoing adjuvant trials in colorectal cancer

cancer) then the patients are randomised into the uncertain indication arm (chemotherapy *v* no chemotherapy). The results from QUASAR's certain arm show that neither high dose folinic acid nor levamisole contribute to improved survvial.

Role of adjuvant chemotherapy in lower risk groups

Inadequate data exist on the effect of chemotherapy in stage B colon cancer. The proportional reduction in annual risk is probably similar for stage B and stage C patients. If the proportional reductions in mortality are similar, the absolute benefits in terms of five year survival would be somewhat smaller for stage B patients than for stage C patients because of lower risk of recurrence (perhaps two to three lives saved per 100 patients treated).

Patients with stage B cancers who have prognostic indicators that suggest a high risk of recurrence (for example, perforation, vascular invasion, poor differentiation) might benefit proportionately more than patients with stage B cancer without high risk indicators and these variables might define a subgroup of patients who might merit adjuvant chemotherapy. Little evidence exists, however, on the prognostic predictability of these various features.

Use of adjuvant therapy in rectal cancer

Insufficient evidence exists to support the routine use of systemic chemotherapy in either Dukes's type B or type C rectal cancer. Anatomical constraints make the rectum less accessible to the surgeon, so it is much more difficult to achieve wide excision of the tumour, and about 50% of recurrences are in the pelvis itself rather than at distant sites. This means that locally directed radiotherapy is a useful adjuvant weapon, and this has been assessed for rectal cancer both before and after surgery.

In the largest trial of preoperative radiotherapy (the Swedish rectal cancer trial), radiotherapy produced a 61% decrease in local recurrence and an improvement in overall survival (58% *v* 48%) compared with surgery alone. Radiotherapy after surgery seems to be less effective, even at higher doses, possibly because of rapid repopulation of tumour cells after surgery or relative hypoxia around the healing wound.

Only one trial, the Uppsala trial in Sweden, has directly compared radiotherapy before and after surgery. Despite a higher dose after surgery, a significant reduction occurred in local recurrence rates among patients treated before surgery (12% *v* 21%, $P < 0.02$).

Animal studies have suggested that fluorouracil may prime the tumour cells and increase the cytotoxic effect of subsequent radiotherapy. Some clinical data support the role of chemoradiotherapy combinations in rectal cancer, but further clinical evidence of benefit needs to be provided before this treatment could be considered for routine use. The uncertain arm of the QUASAR trial will help to resolve this issue.

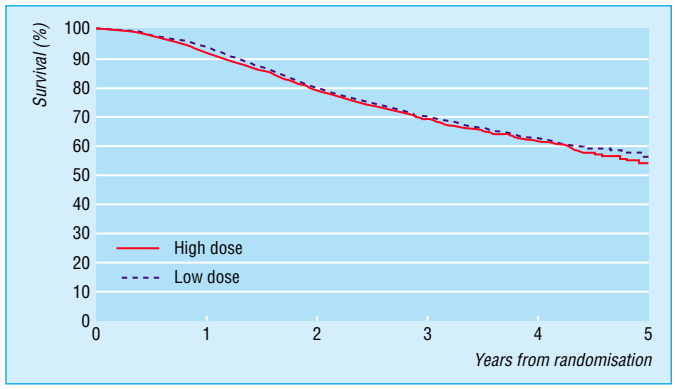


Figure 7.4 Five year survival for 4927 patients with colorectal cancer randomised to high dose or low dose folinic acid with fluorouracil

The uncertain arm of the QUASAR trial is aiming to establish whether chemotherapy is justified in Dukes's type B colon cancer and to define which factors might help to predict chemotherapeutic benefit

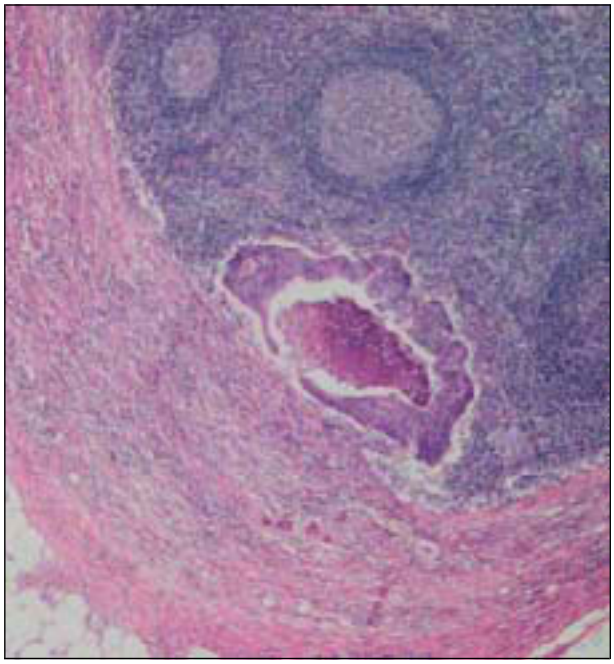


Figure 7.5 Microscopic metastasis in a draining lymph node from a rectal cancer

Table 7.2 Benefits of adjuvant therapy

Site and stage	Chemotherapy	Radiotherapy
Dukes's type C colon cancer	Worth established; fluorouracil plus folinic acid optimal to date; six lives per 100 treated saved; new drugs entering trial	No benefit
Dukes's type B colon cancer	Worth not established; particular subgroups of patients may benefit (on basis of histological and other prognostic factors)	No benefit
Dukes's type B or C rectal cancer	Worth not established	Worth established; preoperative may be superior to postoperative in terms of both efficacy and reducing toxicity

Role of portal venous infusional therapy

Fluorouracil is an S phase specific drug, and yet its active metabolites have a half life of about 10 minutes, which limits its target, when given as a bolus, to the small fraction of cells in the S phase at the time of administration. Infusional therapy can therefore affect a greater proportion of cells. In addition, the most common site for micrometastases after resection of a colorectal tumour is the liver. In contrast with macroscopically identifiable metastases of advanced disease, which derive their blood supply from the hepatic artery, these micrometastases are thought to be supplied by the portal vein. Therefore delivering chemotherapy via the portal vein should provide high concentrations of the drug at the most vulnerable site and lead to substantial first pass metabolism, which should attenuate any systemic toxicity. The established regimen for portal fluorouracil in adjuvant therapy is a course of 5-7 days starting immediately after surgery. A meta-analysis of 10 randomised trials showed a 4.7% improvement in absolute survival with portal venous infusion therapy compared with surgery alone; however, the confidence intervals were wide and the statistical benefit is not robust. Indeed the AXIS trial, the largest single trial of portal venous infusion to date, randomising 4000 patients after surgery either to the infusion therapy or to observation alone at five years, suggests no significant differences in overall survival.

The United Kingdom Co-ordinating Committee on Cancer Research's AXIS study, in which 4000 patients have been randomised to portal vein infusion versus surgery alone, will contribute substantially to the debate on this type of therapy

Future role of adjuvant therapy

The use of adjuvant therapy in colorectal cancer over the past 40 years has centred on fluorouracil, alone and in combination, and on the fine tuning of regimen and route of administration. Current trials are considering new drugs (eg irinotecan and oxaliplatin) and their sequencing, as well as innovative techniques, such as immunotherapy and gene therapy. These techniques will be considered in detail later in the series. Gene therapy and immunotherapy are likely to function optimally, however, when cellular load is low, blood supply is good, and small clusters of cells are surrounded by effectors of the immune system; these therapies may therefore be most suitable as adjuvant therapy rather than for use in advanced disease.

All cytotoxic agents are rigorously tested and applied in advanced disease before being used in adjuvant therapy. New agents that are now entering adjuvant trials will be fully described in the next article in this series.

Further reading

- Gray R. Adjuvant treatment for colorectal cancer. In: *Guidance on commissioning cancer services: improving outcomes in colorectal cancer*. London: Department of Health, 1997.
- Midgley RS, Kerr DJ. Adjuvant treatment of colorectal cancer. *Cancer Treat Rev* 1997;23:135-52.
- Midgley RS, Kerr DJ. Colorectal cancer. *Lancet* 1999;353:391-9.

The authors acknowledge the support of the Cancer Research Campaign and the Medical Research Council.

The survival graph is adpated from the *Lancet* (2000;355:1588-96); the illustration of microscopic metastasis was supplied by Dr D C Rowlands.

8 Treatment of advanced disease

Annie M Young, Daniel Rea

Advanced colorectal cancer can be defined as colorectal cancer that at presentation or recurrence is either metastatic or so locally advanced that surgical resection is unlikely to be carried out with curative intent. Despite most patients undergoing potentially curative surgery and the availability of adjuvant chemotherapy, about 50% of patients presenting with colorectal adenocarcinoma die from subsequent metastatic disease. The five year survival rate for advanced colorectal cancer is lower than 5%.

In the past few years several therapeutic advances—underpinned by multiprofessional, site specialised team working—have finally changed the view that advanced colorectal cancer is an untreatable disease. Although cytotoxic chemotherapy is not suitable for all patients, widespread use in appropriate situations can improve survival and quality of life

Clinical presentation

Local recurrence of a tumour is more common in rectal than colon primaries. It may be identified early in the asymptomatic phase by follow up monitoring or may present with similar symptoms to the primary lesion. Blood loss through the rectum, mucous discharge, altered bowel habit, and straining are common features of recurrent rectal cancer. Pain and urinary symptoms are features of localised pelvic recurrence. Recurrent intra-abdominal disease can present as small or large bowel obstruction, and recurrence at other sites may be indicated by focal features such as hepatic capsular pain, jaundice, dyspnoea, localised bone pain, or neurological symptoms. Systemic features of weight loss, anorexia, nausea, and asthenia are symptoms commonly associated with advanced colorectal cancer. The tumour is often palpable on rectal or abdominal examination, and malignant ascites may also be evident.

Box 8.1 Clinical presentation

- About 20% of colorectal cancer cases will present with advanced disease
- About 50% of patients treated with curative surgery will develop advanced disease
- About 80% of relapses will occur within three years of primary surgery
- About 50% of patients with advanced disease will present with liver metastasis
- About 20% of patients with advanced disease have disease confined to the liver

Referral

Despite clear evidence of the value of chemotherapy and the apparent willingness of cancer patients to have chemotherapy, in the United Kingdom only about 25% of patients with advanced disease are referred to an oncology tertiary centre for consideration of chemotherapy. Referral patterns and treatment policies for patients with advanced colorectal cancer vary widely in the United Kingdom. Currently many regions are in the throws of reorganising their cancer services as part of the implementation of the Calman report, *A Policy Framework for Commissioning Cancer Services*. It is envisaged that referrals to oncologists will increase considerably owing to the publication in 1997 of guidelines for managing colorectal cancer.

Box 8.2 Which patients should be referred for palliative chemotherapy*?

Patients in whom chemotherapy should be considered

- Able to carry out all normal activity without restriction
- Restricted in physically strenuous activity but able to walk about and carry out light work
- Able to walk about and capable of all self care but unable to carry out any work; out of bed or chair for more than 50% of waking hours

Patients unlikely to benefit from chemotherapy

- Capable only of limited self care; confined to bed or chair for more than 50% of waking hours
- Severely disabled; cannot carry out any self care; totally confined to bed or chair

*Based on the World Health Organization's criteria for functional performance status

Overview of management

The management of patients with advanced colorectal cancer involves a combination of specialist active treatment, symptom control measures, and psychosocial support. Active treatment comprises an individual plan (often combining palliative surgery), cytotoxic chemotherapy, and radiation therapy.

The outcome measures of the impact of active treatment have traditionally been survival, response, and toxicity. Alternative end points—for example, quality of life, convenience, acceptability to patients, and patients' preferences—assume greater importance in those with advanced disease, and they should now also be incorporated into the assessment of the relative worth of treatments.

Box 8.3 Current status of chemotherapy

- Many patients with advanced colorectal cancer die without having received chemotherapy
- Chemotherapy improves survival by an average of about six months, compared with supportive care alone
- Chemotherapy improves overall quality of life
- Stabilisation of disease with chemotherapy improves both survival and disease related symptoms
- Early chemotherapy treatment (rather than waiting until symptoms appear) prolongs survival

Surgery

Palliative surgical procedures for advanced colorectal cancer are commonly used to overcome obstructing lesions and to alleviate pelvic symptoms. The liver is the most frequent site of metastasis, and in selected patients with no extrahepatic metastases surgical resection offers the only hope of cure. Five year survival rates of 25-35% have been reported with this highly specialised procedure (Cady and Stone, 1991).

Radiotherapy

In advanced colon cancer, radiotherapy is rarely indicated. In locally advanced rectal disease, localised radiation may render some tumours resectable. Radiotherapy can also be effective in palliation of symptoms—it can improve pain, stop haemorrhage, and lessen straining. In the absence of distant metastases, radiation may afford long term control of the tumour. Pain from isolated bone metastases can also be alleviated with short courses of radiation.

Conventional chemotherapy

In patients with advanced colorectal cancer, chemotherapy is delivered with palliative rather than curative intent. For over four decades fluorouracil has been the mainstay of treatment for advanced colorectal cancer. Folinic acid is given intravenously before fluorouracil to enhance the fluorouracil's cytotoxicity. Large randomised trials of chemotherapy versus best supportive care have shown that fluorouracil based chemotherapy adds about 4-6 months to the remaining life of patients with advanced colorectal cancer. Chemotherapy delays the occurrence or progression of symptoms by about six months and improves symptoms, weight gain, and functional performance in about 40% of patients. Palliative chemotherapy in advanced colorectal cancer should not be restricted by chronological age but by fitness and activity level.

Is failure to respond a failure of treatment?
Less than a third of patients receive an objective tumour response—complete or partial—with fluorouracil based therapy. In a further 20-30% of patients, the disease is stabilised during chemotherapy. The patients with stable disease ("no change" category) also derive a symptomatic and survival advantage from chemotherapy.

Which regimen?
Current evidence supports the use of infusional fluorouracil regimens over bolus schedules in terms of both toxicity and efficacy, but infusional chemotherapy is more complex to administer, requiring permanent vascular access technology or admission to hospital. In the United Kingdom a 48 hour regimen of fluorouracil plus folinic acid repeated every 14 days is commonly used. Ideally, chemotherapy for advanced colorectal cancer should be given within the umbrella of a clinical trial to help resolve outstanding questions of optimal type, duration, and scheduling of therapy.

Tailoring treatment
The optimum duration of chemotherapy is unknown and is currently being tested in clinical trials. The current approaches are either to treat for a fixed period (usually six months) or to treat until progression occurs. Irrespective of which of these approaches is adopted, the overriding need is to monitor rigorously the effect of treatment in terms of response, palliative

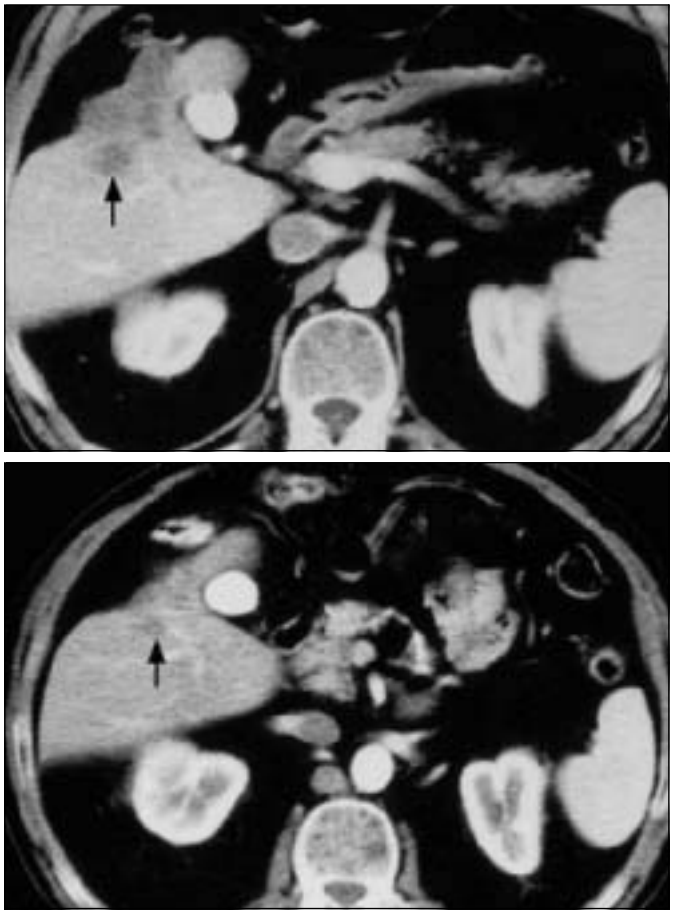


Figure 8.1 Abdominal computed tomogram showing a hepatic metastasis (arrow) before chemotherapy (top) and 17 weeks after chemotherapy (bottom); the later image shows a substantial reduction in the bulk of the hepatic tumour

Box 8.4 Definitions for assessing response and progression after chemotherapy

Complete response—Disappearance of all known disease, determined by two observations not less than four weeks apart

Partial response—Decrease of at least 50% of the sum of the products of the largest perpendicular diameters of all measurable lesions as determined by two observations not less than four weeks apart

No change—Less than 50% decrease and less than 25% increase in the sum of the products of the largest perpendicular diameters of all measurable lesions; no new lesions should appear

Progressive disease—More than 25% increase in the size of at least one lesion or appearance of a new lesion

Data from trials by the Nordic Gastrointestinal Tumour Therapy Group support the early use of chemotherapy, before the patient's condition deteriorates

benefit, and toxicity. This ensures that any toxicity or disease progression is recognised as soon as possible and that the appropriate individualised treatment or cessation of chemotherapy can be implemented without delay.

Chemotherapy toxicity

Chemotherapy for advanced colorectal cancer should be prescribed by experienced oncologists familiar with the toxicity profile of the drug regimens used. Despite concerns over toxicity, currently used infusional regimens are remarkably well tolerated. Management of toxicities in the community requires close liaison with the hospital team, and severe toxicity requires immediate admission. The most common effects of toxicity from chemotherapies for advanced colorectal cancer are diarrhoea, mucositis, asthenia, and neutropenia. Nausea, alopecia, and anorexia can also be experienced. Diarrhoea can be substantially relieved with oral antimotility drugs. Mucositis should be managed with antiseptic mouthwash and prophylactic or early treatment of oral candidiasis. Neutropenia is less common with current infusional regimens but must always be suspected in patients with fever. Prolonged treatment with fluorouracil can produce painful blistering erythema of palms and soles of the feet (palmar plantar erythrodysaesthesia), which often improves with pyridoxine.

Cost effectiveness

In 1995 Glimelius et al showed that the overall cost of early intervention with chemotherapy in patients with advanced colorectal cancer is similar to that of no treatment or delayed chemotherapy, indicating that chemotherapy as part of the management of the advanced disease is indeed cost effective. Inevitably, it is becoming increasingly difficult for the health service to fund modern drugs to treat advanced colorectal cancer. The NHS is struggling to fund the new chemotherapy treatments that are proved to extend life by only a few months or to improve the quality of life only.

Ambulatory and domiciliary chemotherapy

The emergence of primary care health teams, together with developing technology, has allowed for more complex care to be carried out in the community or at home.

Ambulatory infusional chemotherapy is administered via a small pump (battery assisted and disposable elastomeric infuser). The chemotherapy may be connected and disconnected at the hospital outpatient clinic by oncology nurses, or patients can be taught to do this themselves.

A feasibility study of home chemotherapy has been undertaken in Birmingham for patients with advanced colorectal cancer. This shows that a nurse led service (backed up by oncology medical and nursing staff from both primary and secondary health care) is safe and that patients and carers find home therapy of immeasurable value. Early analysis shows that the cost of this home service is similar to and often cheaper than the current hospital based service.

New drugs

In recent years the availability of several new drugs has revived interest in the treatment of advanced colorectal cancer. New treatments include alternative fluoropyrimidines, new thymidylate synthase inhibitors, new modulators of fluorouracil and also mechanistically new drugs.

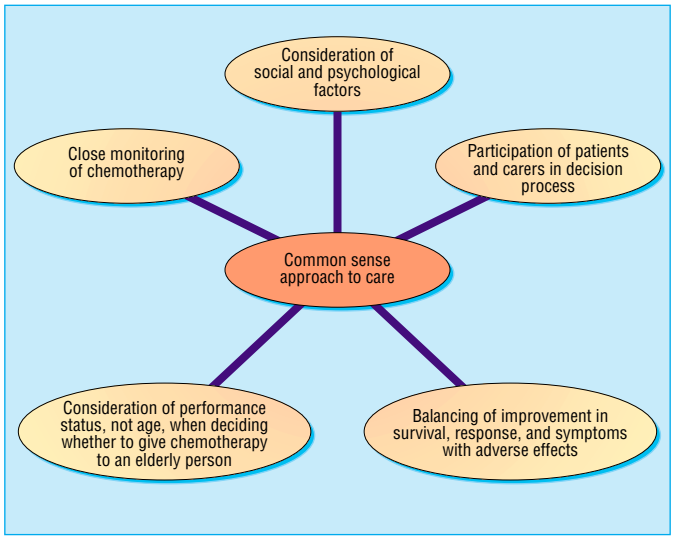


Figure 8.2 Basic elements of caring for patients with advanced colorectal cancer

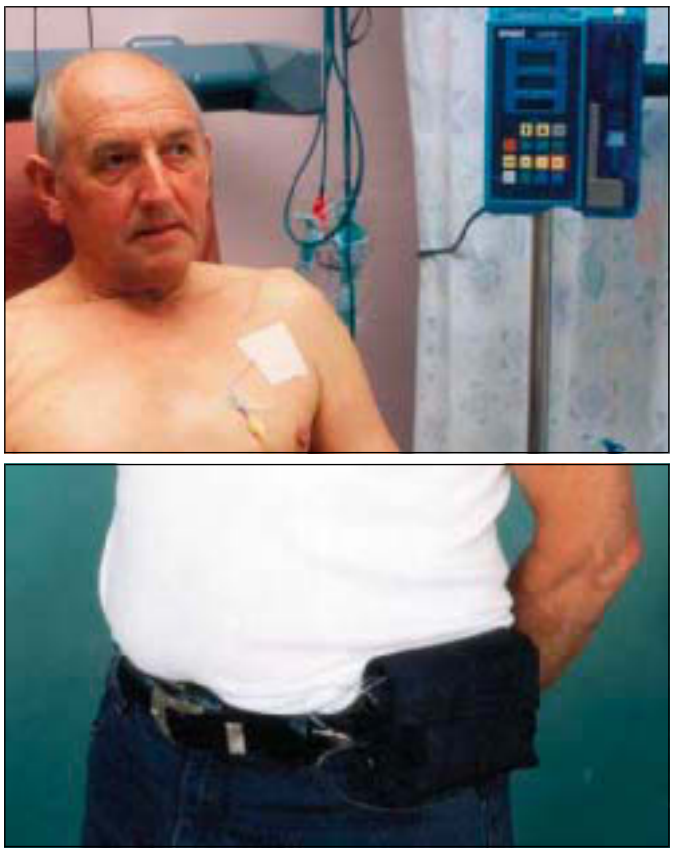


Figure 8.3 Patient receiving chemotherapy through central venous catheter in hospital outpatient department (top); and small, battery assisted pump, worn on the waist and used to deliver chemotherapy through a central venous catheter (bottom). Patients are free to perform many normal activities during "ambulatory" chemotherapy

Box 8.5 Current controversies in advanced colorectal cancer

- For how long should chemotherapy be given?
- Are new delivery routes for fluorouracil—for example, orally and by intrahepatic arterial administration—superior to conventional intravenous fluorouracil?
- Should newer agents with similar efficacy but more convenient intravenous regimens be used in place of fluorouracil?
- What is the optimum combination and sequence for fluorouracil based therapies and the new chemotherapy drugs?
- Is home chemotherapy viable?
- How are the new, more expensive drug therapies to be funded?

New thymidylate synthase inhibitors
Raltitrexed is a quinazoline analogue antifolate that gains entry to cells via the reduced folate carrier and is polyglutamated to a potent, long acting, specific inhibitor of thymidylate synthase. Its regimen—a short intravenous infusion every three weeks—has similar efficacy to that of fluorouracil plus folinic acid and is clearly more convenient, although potentially more toxic.

Oral fluorouracil prodrugs and modulators
Fluoropyrimidine analogues have been developed with reliable oral bioavailability. In addition, oral inhibitors of fluorouracil catabolism can facilitate oral dosing. Preliminary data show similar effectiveness and lower toxicity compared with fluorouracil. Given the convenience and potential cost savings, oral therapy may soon find a place in routine practice.

Irinotecan and oxaliplatin
Irinotecan is a camptothecin analogue that acts through the inhibition of a DNA unwinding enzyme, topoisomerase I, resulting in replication arrest with breaks in single strand DNA. It is useful in advanced colorectal cancer, even after resistance to fluorouracil has developed, and is associated with a survival benefit (about three months) compared with best supportive care. This drug can be associated with severe late onset diarrhoea, which must be treated immediately. Selection of patients, therefore, plays an important part in the safe use of this agent.

Oxaliplatin is a new platinum derivative analogue that crosslinks DNA and induces apoptotic cell death. It shows synergism with fluorouracil. The dominant toxic effect is cumulative neurotoxicity.

Fluorouracil plus either irinotecan or oxaliplatin is superior to fluorouracil alone as a first line treatment for advanced colorectal cancer, with improvement in progression-free survival and, in the case of irinotecan, overall survival. Questions about the optimum sequence and combination of these agents remain and are the subject of ongoing clinical trials.

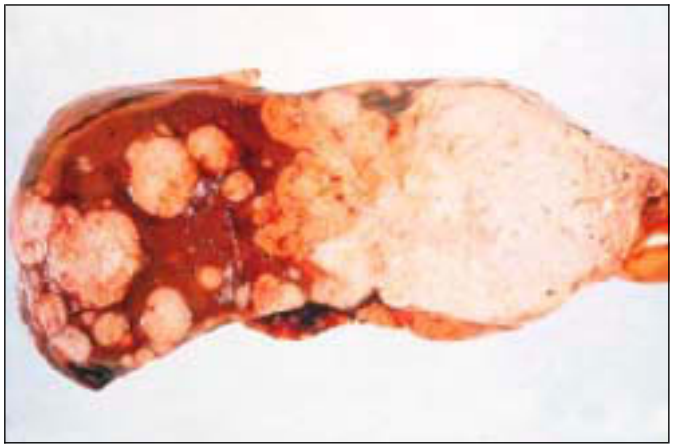


Figure 8.4 Liver with over 50% hepatic replacement by metastatic colorectal cancer

The Colorectal Forum is a worldwide educational service for healthcare professionals working with patients with colorectal cancer. Its website provides news on conferences and events, recommendations on management of advanced colorectal cancer, articles and visual images, reviews of recent publications, and the opportunity to debate controversial clinical issues. It can be accessed at www.colorectal-forum.org

Intrahepatic arterial chemotherapy

For patients with unresectable hepatic metastases, intrahepatic arterial chemotherapy should be considered. This approach greatly increases drug delivery to the liver and doubles the rate at which tumours shrink, with tolerable toxicity. Owing to the complexity of placing the delivery catheter, intrahepatic arterial chemotherapy is usually administered at specialist centres. Current trials should offer definitive proof of whether intrahepatic arterial chemotherapy offers survival benefits compared with conventional intravenous therapy.

Supportive care

All patients with advanced colorectal cancer need continual evaluation of symptoms and appropriate measures for controlling symptoms. Dietary advice and nutritional supplements can stop weight loss, and corticosteroids may be used for their anabolic effect. Psychosocial aspects of care should incorporate evaluation of and provision for the needs of both the patient and the family. Supportive care needs to be tailored to the individual's circumstances and should involve the close collaboration of locally available palliative care services (both in the community and in hospitals). The initial contact between the patient and the palliative team should ideally be made at the time of diagnosis rather than at a crisis point when urgent input from palliative care services is required.

Further reading

- *A policy framework for commissioning cancer services*. London: Department of Health, 1994. (Consultative document.)
- Clinical Outcomes Group. *Guidance on commissioning cancer services: improving outcomes in colorectal cancer*. London: NHS Executive, 1997.
- Cady B, Stone M. The role of surgical resection of liver metastases in colorectal carcinoma. *Semin Oncol* 1991;18:399-406.
- Nordic Gastrointestinal Tumour Adjuvant Therapy Group. Expectancy or primary chemotherapy in patients with advanced, asymptomatic colorectal cancer: a randomised trial. *J Clin Oncol* 1992;10:904-11.
- Glimelius B, Hoffman K, Graf W, Haglund U, Nyren O, Pahlman L, et al. Cost-effectiveness of palliative chemotherapy in advanced gastrointestinal cancer. *Ann Oncol* 1995;6:267-74.

9 Effectiveness of follow up

Colin McArdle

Population based studies show that for rectal cancer the incidence of local recurrence after apparently curative resection is about 20%. Local recurrence after surgery for colon cancer is less common. The liver is the commonest site of distant spread, followed by the lungs; brain and bone metastases are relatively rare. Most recurrences are within 24 months of surgery.

Symptomatic recurrence of colorectal cancer is seldom amenable to curative surgery

Aim of follow up

Traditionally surgeons have reviewed their patients at regular intervals after apparently curative resection. Recent surveys, however, have highlighted the lack of consensus among surgeons about the optimal modality and intensity of follow up; surveillance strategies range from a single postoperative visit to lifelong surveillance. Enthusiasts believe that intensive follow up and early intervention will lead to a reduction in the number of deaths from colorectal cancer; others point to the fact that the value of follow up is unproved. With so many tests available and no consensus on their value, it is not surprising that individual clinicians have tended to devise their own protocols.

Box 9.1 Aims of follow up

- Early detection and treatment of recurrent disease
- Detection of a second, or metachronous, tumour in the large bowel
- Provision of psychological support and advice
- Facilitation of audit

Results of meta-analysis

A meta-analysis in the mid-1990s did little to clarify the situation. The researchers evaluated the results of seven non-randomised studies (covering over 3000 subjects in total) that compared intensive follow up with minimal or no follow up. Clearly several potential biases could and did exist. In the intensive group, investigations included clinical examination, faecal occult blood testing, liver function tests, measurement of the carcinoembryonic antigen, sigmoidoscopy, and either colonoscopy or barium enema examination. Liver ultrasonography was performed in only three studies and even then infrequently. In the intensive group more asymptomatic recurrences were detected, more patients underwent "second look" laparotomy, and more patients had a second potentially curative resection; more metachronous tumours were also detected and resected. However, although there were fewer deaths in the group receiving intensive follow up, this difference did not reach significance.

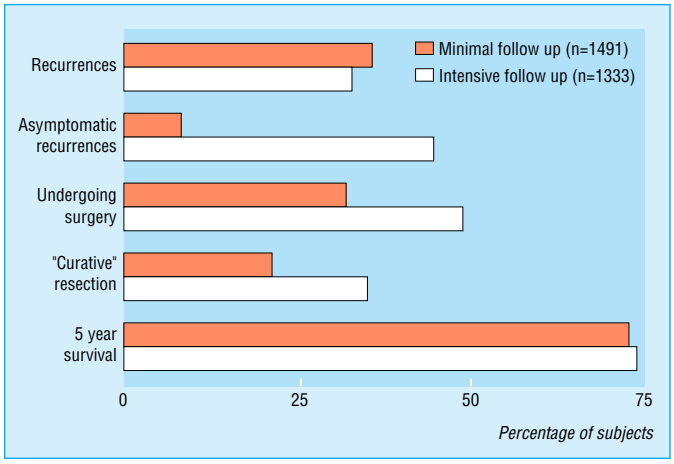


Figure 9.1 Results of meta-analysis of seven non-randomised trials that compared intensive with minimal or no follow up (Bruinvels et al, 1994)

Results of randomised clinical trials

Since the meta-analysis, four randomised trials of intensive follow up have been reported. Ohlsson and his colleagues randomised 107 patients to no follow up or to intensive follow up, similar to that described above. No liver imaging was performed routinely. No differences were found in recurrence rates or in overall or cancer specific mortality.

Mákelá and his associates compared conventional with intensive follow up in 106 patients. In the intensive group flexible sigmoidoscopy was performed every three months, ultrasonography every six months, and colonoscopy and abdominal computed tomography at yearly intervals. Recurrences were detected at an earlier stage (median 10 months *v* 15 months) in the intensive group. Despite this, no difference in survival was found between the two groups.

The results of the four randomised controlled trials of intensive follow up should be interpreted with caution. Despite consistently fewer deaths in the intensive group in each study, the numbers in each were small, and no study had sufficient power to detect a survival advantage

Kjeldsen and his colleagues randomised almost 600 patients to either six monthly follow up or to follow up visits at five and 10 years only. Investigations included chest *x* ray and colonoscopy; no routine liver imaging was performed. Recurrence rates were similar (26%) in both groups, but the recurrences in the intensive group were detected on average nine months earlier, often at an asymptomatic stage. More patients with local recurrence underwent repeat surgery with curative intent. No difference existed, however, in overall survival (68% *v* 70%) or cancer related survival.

More recently, Schoemaker and his colleagues evaluated the addition of annual chest radiography, colonoscopy, and computed tomography of the liver to a standard follow up based on clinical examination, faecal occult blood testing, liver function tests, and measurement of the carcinoembryonic antigen, with further investigations as clinically indicated. At five years, fewer patients in the intensive group had died, but the result was not significant. At the cost of 505 additional investigations, annual colonoscopy failed to detect any asymptomatic local recurrences; only one asymptomatic metachronous colon tumour was detected. Six hundred and eight additional liver computed tomograms detected only one asymptomatic patient with liver metastases who might have benefited from liver resection.

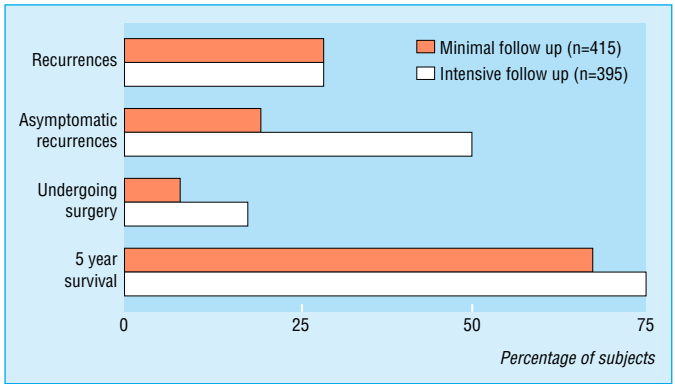


Figure 9.2 Combined results of three randomised trials of intensive follow up

Carcinoembryonic antigen

Carcinoembryonic antigen concentrations have also been used to predict recurrence. About three quarters of patients with recurrent colorectal cancer have a raised carcinoembryonic antigen concentration before developing symptoms.

An alternative approach therefore would be to monitor this concentration regularly during follow up and, in those patients showing a rising concentration, undertake second look laparotomy. However, although early non-randomised studies suggested that surgery that was prompted by this method resulted in more potentially curative repeat operations for recurrence, more recent studies have failed to show a survival advantage.

Moertel analysed outcome in patients included in trials of adjuvant therapy, according to whether the patient underwent carcinoembryonic antigen testing. Of 1017 patients whose concentrations were monitored, 417 (41%) developed recurrence. A comparison of those patients whose follow up included measurements of carcinoembryonic antigen with those whose follow up did not, failed to show any difference in disease-free survival. Among 29 laparotomies performed solely on the basis of a raised concentration of carcinoembryonic antigen, only one patient remained alive and disease-free after one year.

In the randomised study by Northover and his colleagues, 1447 patients undergoing potentially curative surgery were randomised to an intervention group or a control group. Carcinoembryonic antigen was measured in all patients at frequent intervals. In the intervention group, a rising antigen concentration prompted further investigation, including second look laparotomy, if appropriate.

Preliminary analysis showed no difference in survival between the two groups. The failure to show a survival advantage in the intervention group may be due to the fact that a rising antigen concentration is a relatively poor predictor of local recurrence; furthermore, even in patients with liver metastases a rising concentration is a relatively late phenomenon.

Table 9.1 Results of intensive follow up*

Follow up	Colonoscopy	Chest *x* ray	Liver CT
Standard (n = 158)	72	17	66
Intensive (n = 167)	577	650	674
No of extra investigations	505	633	608
No of asymptomatic recurrences resulting from extra investigations	0	0	10
No of cures resulting from extra investigations	0	1	1

CT = computed tomography.
*Data from Schoemaker et al, 1998 (see Further reading box).

Carcinoembryonic antigen concentrations have been used to predict recurrence of colorectal cancer, but recent evidence does not support this approach

Table 9.2 Results of "second look" surgery according to measurement of carcinoembryonic antigen (CEA)*

CEA concentration	No of patients	No (%) of "curative" resections	% of patients free of recurrence at 1 year
Raised	345	47 (14)	2.9
Normal	672	38 (6)	1.9
Not measured	200	23 (12)	2.0

*Data from Moertel et al, 1993.

Cost effectiveness

Concern is also increasing about the cost of follow up. A review of the published literature suggests a 28-fold difference in costs between the least intensive and most intensive, published, five year follow up protocols.

Box 9.2 Costs of follow up, suggested by recent study from Italy

- £2530 per patient over five years
- £9050 per recurrence detected
- £39 890 for each case undergoing further surgery
- £91 190 for each "cured" patient

Wrong target?

Clearly, follow up as currently practised is ineffective. Why, therefore, should we continue to follow patients up after apparently curative resection for colorectal cancer? There are several reasons. Firstly, we should do so to provide psychological support and advice; many patients welcome the reassurance that regular check up provides. Secondly, routine follow up facilitates audit of outcome measures after surgery, ensures quality control and facilitates evaluation of trials of new treatments and strategies.

There may, however, be a more fundamental reason that current follow up practices are ineffective. On theoretical grounds, attempts to identify potentially resectable local recurrences or metachronous tumours were never likely to have a significant impact on survival. Isolated resectable anastomotic recurrences are uncommon. Most local recurrences arise from residual disease left at the time of surgery and therefore, by definition, are unlikely to be amenable to further curative surgery. Metachronous tumours, although potentially amenable to surgery, are relatively uncommon.

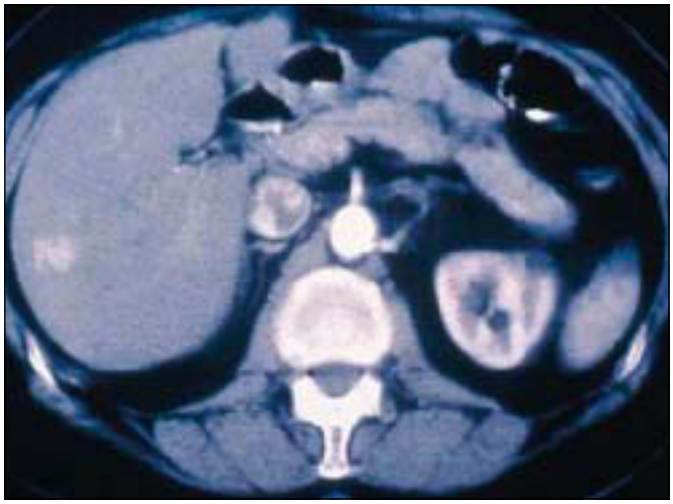


Figure 9.3 Contrast enhanced computed tomogram (arterial phase) showing solitary liver metastasis

Wrong intervention?

In contrast, liver metastases are much more common. Furthermore, these metastases are confined to the liver in about a quarter of patients.

Perhaps, therefore, the emphasis should shift towards the early detection of liver metastases. It is worth noting that in contemporary studies of liver resection, mortality is less than 5% and about 35% of patients survive five years. These figures are better than the results obtained after primary surgery for many types of gastrointestinal cancer. Furthermore recent studies have shown that patients with disseminated disease who receive systemic chemotherapy at an asymptomatic stage have higher response rates, better quality of life, and improved survival compared with those in whom the administration of chemotherapy is delayed until symptoms appear. Therefore if liver metastases were diagnosed in more patients at a point at which they were amenable to resection or chemotherapy, more long term survivors might be anticipated.

To date only two randomised studies have included liver imaging. In both these studies the numbers were small and liver imaging was infrequent. In neither study was a survival advantage noted. However, intensive liver imaging for the first three years after surgery may be more effective: at the Royal Infirmary in Glasgow more than 80% of patients who developed liver metastases as the initial site of recurrence were detected at an asymptomatic stage.

Table 9.3 Comparison of results of trial of early versus delayed chemotherapy in patients with advanced colorectal cancer

Treatment group	No of patients	Median symptom-free survival (months)	Median survival (months)	Survival at 1 year (%)
Early	92	10	14	55
Delayed	91	2	9	38

Early chemotherapy was given when patients were asymptomatic; delayed chemotherapy was given when patients were symptomatic.
Data from the Nordic Gastrointestinal Tumor Group, 1992.

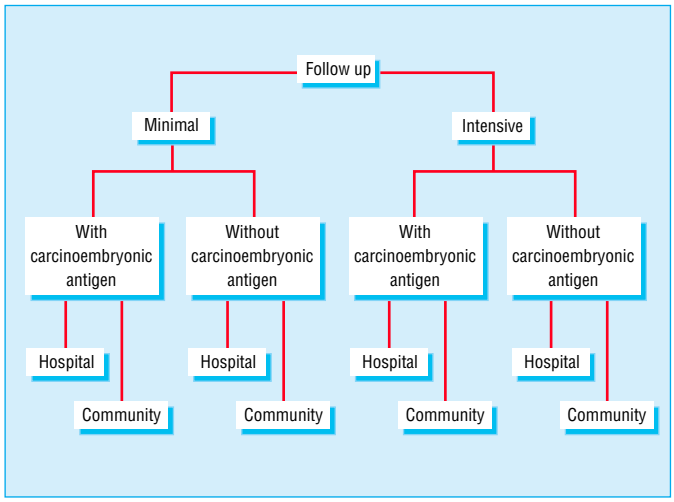


Figure 9.4 Suggested study outline to test three follow up strategies: intensive *v* minimalist; role of carcinoembryonic antigen; and general practitioner (community) coordinated *v* hospital coordinated

Hospital or community coordination of follow up?

Most patients with colorectal cancer are followed up in hospital. Yet overwhelming evidence from previous studies shows that few curable recurrences are detected at routine follow up based on history, physical examination, and routine blood tests. Few patients are followed up by their general practitioners, although

good evidence exists that, in other tumours at least, such follow up is as effective (or ineffective) as hospital follow up. Furthermore, provided that general practitioners have access to a "fast track" referral system for patients in whom they suspected recurrent disease, follow up coordinated by general practitioners might offer several advantages. It might be more acceptable to and convenient for patients and might reduce costs.

Perhaps it is time to reassess follow up. Formal studies to assess the value of these strategies might include:

- A comparison of the value of intensive versus minimalist follow up
- A re-evaluation of the role of carcinoembyronic antigen
- A comparison of the effectiveness of follow up that is coordinated by general practitioners rather than by hospitals.

Conclusion

Current methods of follow up, aimed at the early detection and treatment of local recurrence or metachronous tumours, have yet to be shown to be cost effective.

As liver metastases are common, a protocol that includes regular liver imaging to detect potentially resectable lesions may prove more effective. Further studies are needed to assess the value of this approach in patients undergoing apparently curative resection for colorectal cancer.

Further reading

- Bruinvels DJ, Stiggelbout AM, Kievit J, Van Houwelingen HC, Habbema JDF, van de Velde CJH. Follow-up of patients with colorectal cancer: a meta-analysis. *Ann Surg* 1994;219:174-82.
- Kjeldsen B, Kronberg O, Fenger C, Jorgensen O. A prospective randomised study of follow-up after radical surgery for colorectal cancer. *Br J Surg* 1997;84:666-9.
- Mákelá JT, Laitinen SO, Kairaluoma MI. Five year follow-up after radical surgery for colorectal cancer: results of a prospective randomised trial. *Arch Surg* 1995;130:1062-7.
- Ohlsson B, Breland U, Ekberg H, Graffner H, Tranberg K. Follow-up after curative surgery for colorectal carcinoma: randomised comparison with no follow-up. *Dis Colon Rectum* 1995;38:619-26.
- Schoemaker D, Black R, Giles L, Toouli J. Yearly colonoscopy, liver CT, and chest radiography do not influence 5-year survival of colorectal cancer patients. *Gastroenterology* 1998;114:7-14.

10 Innovative treatment for colon cancer

G A Chung-Faye, D J Kerr

Despite advances in treatment for colon cancer, the five year survival has not significantly altered over the past decade. Survival could improve in several key areas:

- Preventive measures—such as diet and chemoprevention with agents such as non-steroidal anti-inflammatory drugs
- Screening strategies—such as faecal occult blood testing and flexible sigmoidoscopy
- Optimisation of current chemotherapy and radiotherapy regimens and the development of more effective antineoplastic agents
- New therapeutic approaches—such as immunotherapy and gene therapy.

This article will focus on prevention with non-steroidal anti-inflammatory drugs and on new strategies for treating colon cancer.

Dietary modifications to reduce the incidence of colon cancer may be difficult to implement (dietary interventional studies have shown this to be the case for cardiovascular disease); the roles of screening, chemotherapy, and radiotherapy have been covered earlier in this series

Non-steroidal anti-inflammatory drugs

Evidence strongly suggests a protective effect of non-steroidal anti-inflammatory drugs in colon cancer. Several cohort and case-control studies have consistently shown dose related reductions of colorectal cancer in regular users of these drugs. Furthermore, patients with familial adenomatous polyposis who took the non-steroidal anti-inflammatory sulindac had reductions in the number and size of their polyps. Gene knockout studies in mice suggest that inhibition of the cyclo-oxygenase type 2 pathway by non-steroidal anti-inflammatory drugs may be important in the mechanism of action.

The only randomised controlled trial examining the effect of aspirin in primary prevention of colon cancer did not show any benefit after five years of aspirin use. A recent prospective cohort study suggested, however, that five years may be insufficient to show any benefit and that 10-20 years is needed to show an effect.

The predominant side effect from using non-steroidal anti-inflammatory drugs is the increased incidence of gastrointestinal bleeds. On the current evidence, the mortality risk from such bleeding would be outweighed by the reduction in mortality from colon cancer. To maximise the benefit to risk ratio, however, targeting individuals at high risk of colon cancer may prove more fruitful.

Non-steroidal anti-inflammatory drugs could be used as secondary prevention after surgical resection of colonic tumours, but this approach has yet to be tested in a large randomised controlled trial.

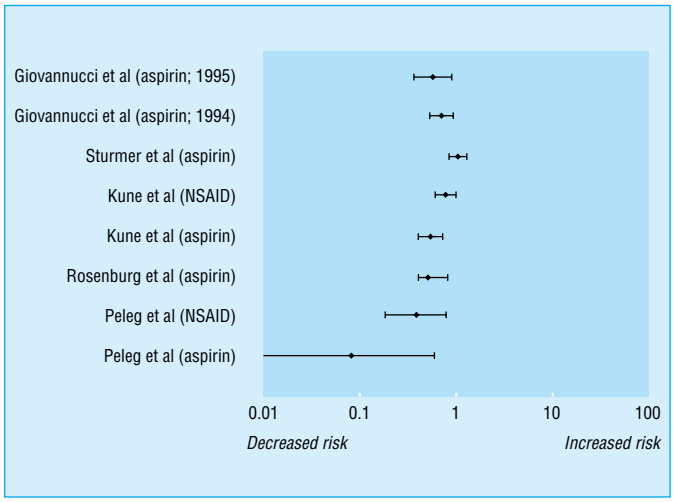


Figure 10.1 Use of non-steroidal anti-inflammatory drugs (NSAID) and relative risk of colorectal cancer

Immunotherapy

Many cancers can be destroyed by a tumour specific, cell mediated immune response, usually by CD8 (cytotoxic) lymphocytes. However, colorectal tumours are poorly immunogenic and may evade immune destruction by various mechanisms, such as tumour "tolerance." To overcome these problems, several immunostimulatory approaches have been advocated to augment the innate immune response against tumours.

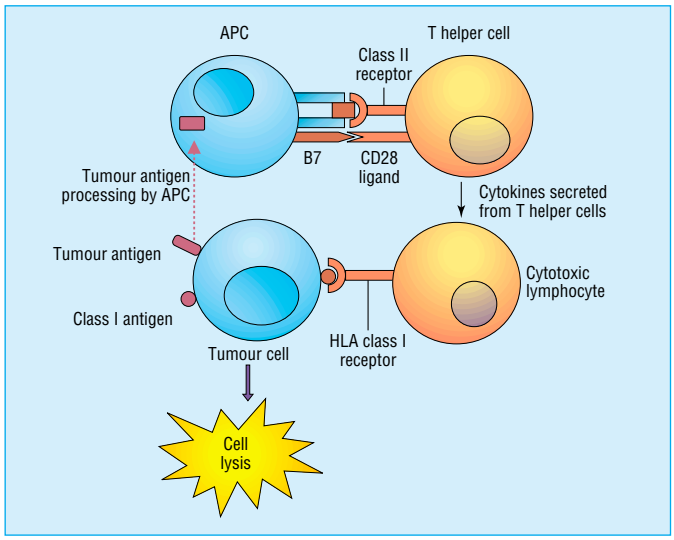


Figure 10.2 Cell mediated immunity against tumours. Tumour antigens are taken up and processed by antigen presenting cells (APC) and re-presented to class II receptors on T helper cells. This requires a costimulatory signal, B7, which binds to the CD28 ligand, causing T helper cell activation. This leads to secretion of cytokines, which in turn activates cytotoxic lymphocytes to bind to tumour cells via class I receptors and causes tumour lysis

Vaccination with autologous tumour cells
This approach uses cells derived from the patient's tumour to elicit a cell mediated immune response against the tumour. To increase the efficacy of this response, tumour cells are coadministered with an immunomodulatory adjuvant, such as BCG. This approach has been tested in three randomised, controlled trials in an adjuvant setting in colorectal cancer, after resection of the tumour. No serious side effects were encountered in any of the studies.

Vaccination against tumour associated antigens
An alternative approach is vaccination against a tumour associated antigen, such as the carcinoembryonic antigen, which is overexpressed in 90% of colon cancers. A phase I immunisation study of a recombinant vaccinia virus, encoding the gene for carcinoembryonic antigen, in patients with advanced colorectal cancer, showed HLA specific, cytolytic T cell responses to carcinoembryonic antigen epitopes in vitro. This study did not show any clinical benefit, but several trials are under way, using optimal vaccination approaches in patients with minimal residual disease where clinical responses may be observed.

Monoclonal antibodies directed against tumour antigens
Monoclonal antibodies against tumour antigens have been shown to elicit immune responses against the tumour, which may previously have induced immunogenic tolerance. The 17-1A antigen is a surface glycoprotein with a putative role in cell adhesion and is present in over 90% of colorectal tumours.

In a study among patients with Dukes's stage C colon cancer the patients were randomised to receive either surgery alone or surgery plus repeat administrations of a monoclonal antibody against the 17-1A antigen. Side effects of the treatment were infrequent, consisting mainly of mild constitutional and gastrointestinal symptoms. Four patients experienced an anaphylactic reaction, which required intravenous steroids but no hospital admission.

Gene therapy

Gene therapy represents a new treatment approach for colon cancer. It is at a developmental stage, and preclinical studies are only just being translated into clinical trials. Two gene therapy strategies are currently used, gene correction and enzyme-prodrug systems.

Gene correction
The most logical approach to gene therapy is the correction of a single gene defect, which causes the disease phenotype. In colon cancer, as in many other cancers, this goal is elusive as malignant transformation is usually accompanied by a series of genetic mutations. However, some of these mutations, such as the p53 gene mutation, are important for the propagation of the malignant phenotype, and the corollary is that correcting these mutations may inhibit the growth of tumour cells.

P53 gene
The p53 gene regulates the cell cycle and can cause growth arrest or apoptosis in response to DNA damage. Loss of p53 control leads to uncontrolled growth and is associated with more aggressive tumours. Restoration of wild-type p53 in p53 mutated tumours inhibits growth. In a phase I trial an adenovirus encoding wild-type p53 was delivered by hepatic artery infusion to 16 patients with p53 mutated colorectal liver metastases. This procedure was well tolerated, with the side

Box 10.1 Immunostimulatory approaches for augmenting the innate immune response against tumours

- Vaccination with autologous tumour cells
- Vaccination against tumour associated antigens, such as carcinoembryonic antigen
- Use of monoclonal antibodies directed against tumour antigens

Box 10.2 Vaccination with autologous tumour cells

Hoover et al, 1993
- 98 patients with colon or rectal cancer were randomised to surgery alone or to surgery plus vaccination with autologous tumour cells
- No significant improvement in the recurrence or the survival rate
- Subgroup analysis of patients with colon cancer showed a significant improvement in survival and disease-free survival in those who received vaccination (P = 0.02, P = 0.039 respectively)

Harris et al, 1994
- 412 patients with Dukes's stage B and C colon cancer were postoperatively randomised to vaccination with autologous tumour cells or to no further treatment
- No significant differences between treated and untreated groups

Vermorken et al, 1999
- 254 postoperative patients with stage II or III colon cancer were randomised to vaccination with autologous tumour cells or to no further treatment. Those randomised to receive vaccination received a 4th booster vaccine after six months (in contrast with the patients in the two previous studies, who received only three doses)
- In the those receiving vaccination, there was a significant reduction in recurrence (44%, 95% confidence interval 7% to 66%) and a reduction in overall survival, although this did not reach significance
- The main benefit was in stage II disease, with a non-significant reduction in recurrence in stage III disease; this was thought to be due to the increased tumour burden in more advanced stages

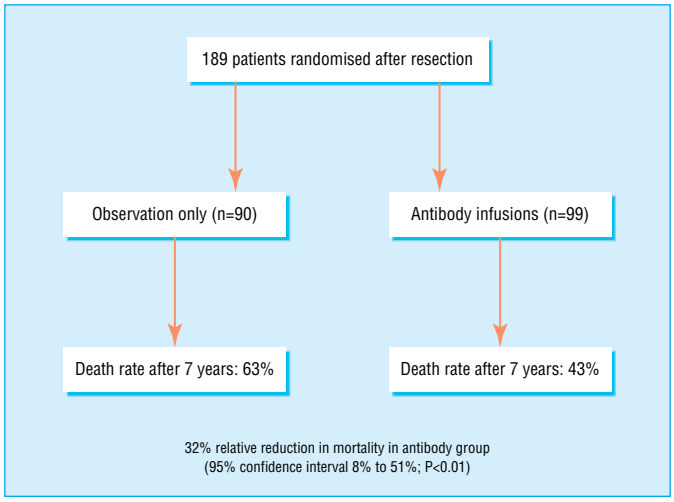


Figure 10.3 Results of a study of monoclonal antibody against 17-1A as adjuvant therapy in Dukes's C colon cancer (Riethmuller et al, 1998)

The p53 gene mutation is present in most colon cancers

effects of fever and transiently damaged liver function. Although gene expression was detected in subsequently resected tumours, no radiographic responses were seen at 28 days. This study has now proceeded to a phase II trial, in combination with intrahepatic floxuridine based chemotherapy, in which 11 out of 12 patients have had partial responses.

Virus directed enzyme-prodrug treatment

Enzyme-prodrug systems are used to localise the toxic drug effects to tumour cells. This involves gene transfer of an enzyme into tumour cells, which converts an inactive prodrug into a toxic metabolite, leading to cell death. An important feature of enzyme-prodrug systems is the "bystander effect," whereby surrounding cells (not expressing the enzyme) are also killed by active metabolites. Gene transfer is achieved by viral vectors, such as retroviruses or adenoviruses. One such enzyme-prodrug combination is the bacterial enzyme cytosine deaminase, which converts the antifungal agent fluorocytosine into the antineoplastic agent fluorouracil. Fluorouracil induces apoptosis by inhibition of the enzyme thymidylate synthase during DNA replication. In murine models with colon cancer xenografts expressing cytosine deaminase, 75% of mice were cured by administration of fluorocytosine, whereas no anti-tumour effect was seen with the maximally tolerated dose of fluorouracil.

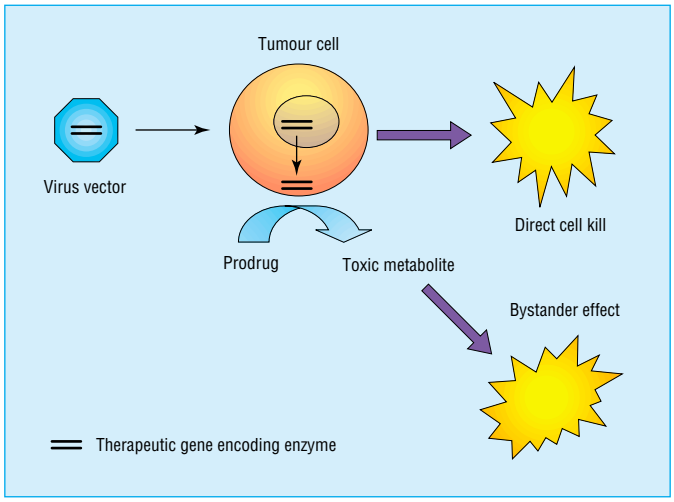


Figure 10.4 Virus directed enzyme-prodrug treatment

Two current phase I trials are using a virus directed enzyme-prodrugs approach for treating colorectal liver metastases by direct injection into the tumour of an adenovirus encoding a therapeutic enzyme. One study is using nitroreductase plus the intravenous prodrug CB1954; the other is using cytosine deaminase plus fluorocytosine

New therapeutic agents

The matrix metalloproteinases are a group of enzymes involved in the physiological maintenance of the extracellular matrix. They degrade the extracellular matrix and promote the formation of new blood vessels and are involved in tissue remodelling processes, such as wound healing and angiogenesis. Matrix metalloproteinases are overexpressed, however, in various tumours, including colorectal cancers, and have been implicated in facilitating tumour invasion and metastasis. The matrix metalloproteinase inhibitor, marimastat, has shown reductions in levels of tumour markers in phase I studies, and its clinical efficacy is currently being tested in phase III trials.

Conclusions

Non-steroidal anti-inflammatory drugs seem to be the most promising drug for prevention of colon cancer; case-control and prospective cohort studies strongly suggest they reduce the risk of colon cancer. This is further supported by studies in familial cancer patients and animal data. However, this effect of non-steroidal anti-inflammatory drugs is unproved in randomised controlled trials, and the issue remains to be addressed.

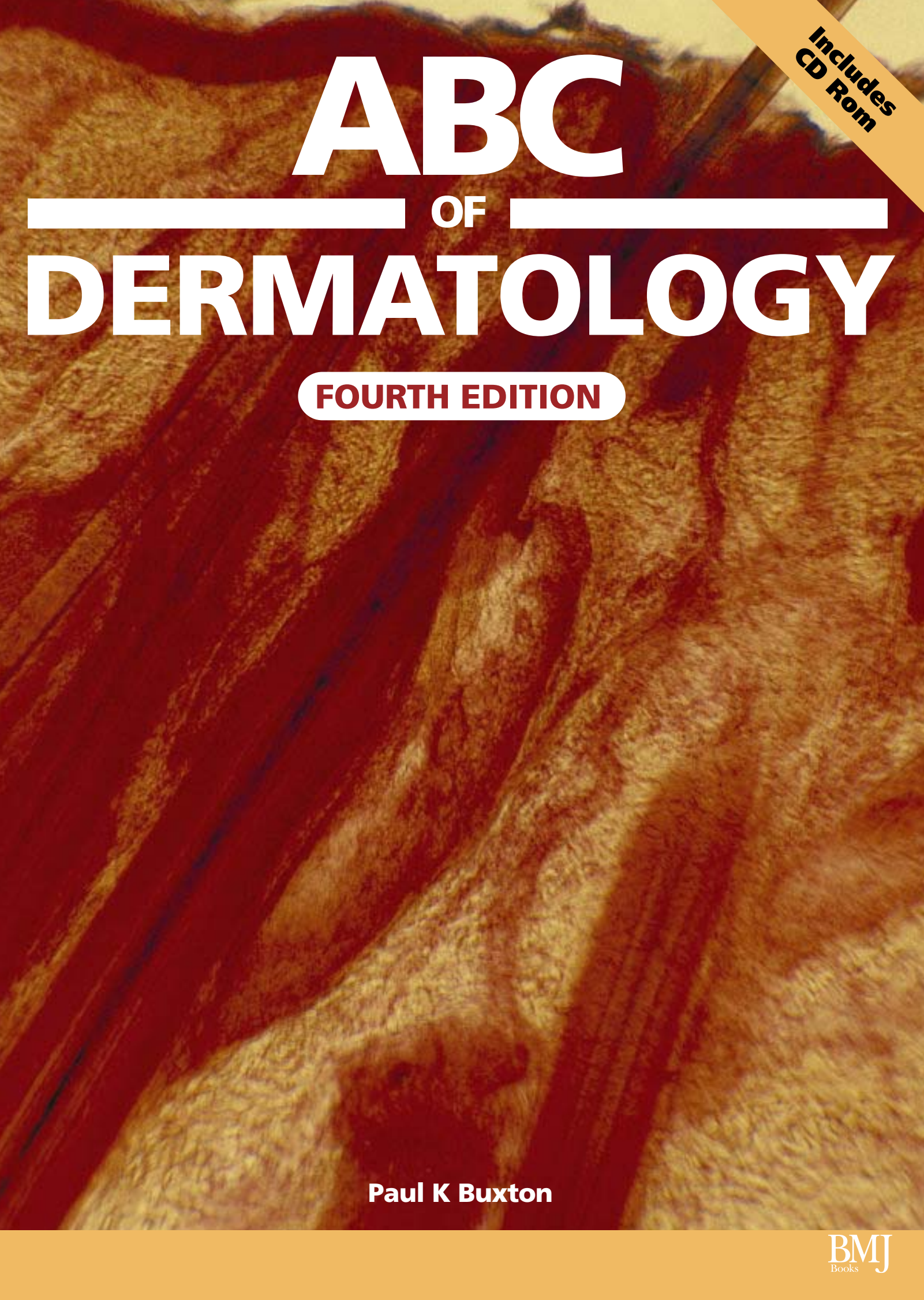
Immunotherapy seems to be well tolerated and effective in an adjuvant setting in colon cancer with limited residual disease. Its effect in stage II disease is comparable to that of adjuvant chemotherapy in Dukes's C colon cancer. In more advanced disease it may have a role in combination with chemotherapy, and this approach is being explored in ongoing trials.

Gene therapy for colon cancer is still at an early stage of development. Preclinical studies have prompted several phase I trials. However, significant problems remain, such as low efficiency in gene transfer and the inhibitory effect of the host immunity, which may be addressed by developments in vector technology. As our understanding of the molecular biology of cancer increases, gene therapy is likely to have an increasingly important role in the expanding array of treatment options for colon cancer.

Further reading

- Langman M, Boyle P. Chemoprevention of colorectal cancer. *Gut* 1998;43:578-85.
- Gann PH, Manson JE, Glynn RJ, Buring JE, Hennekens CH. Low-dose aspirin and incidence of colorectal tumors in a randomized trial. *J Natl Cancer Institute* 1993;85:1220-4.
- Vermorken JB, Claessen AME, van Tintern H, Gall HE, Ezinga R, Meijer S, et al. Active specific immunotherapy for stage II and stage III human colon cancer: a randomised trial. *Lancet* 1999;353:345-50.
- Riethmuller G, Holz E, Schlimok G, Schmiegel W, Raab R, Hoffken K, et al. Monoclonal antibody therapy for resected Dukes' C colorectal cancer: seven-year outcome of a multicenter randomized trial. *J Clin Oncol* 1998;16:1788-94.
- Roth JA, Cristiano RJ. Gene therapy for cancer: what have we done and where are we going? *J Natl Cancer Institute* 1997;89:21-39.
- Chung-Faye GA, Kerr DJ, Young LS, Searle PF. Gene therapy strategies for colon cancer. *Molecular Medicine Today* 2000;6:82-7.

Index


Includes CD Rom
ABC
OF
DERMATOLOGY
FOURTH EDITION
Paul K Buxton
BMJ Books

ABC OF DERMATOLOGY

Fourth Edition

ABC of Dermatology CD Rom

Features

ABC of Dermatology PDF eBook
- Bookmarked and hyperlinked for instant access to all headings and topics
- Fully indexed and searchable text—just click the "Search Text" button

Artwork slideshow
- Every diagram and photograph from the book, organised by chapter
- Hover over a image thumbnail and the caption will appear in a pop-up window
- Click on the image thumbnail to view at full-screen size, then use the left and right cursor keys to view the previous or next figure

PDA Edition sample chapter
- A chapter from *ABC of Dermatology*, adapted for use on handheld devices such as *Palm and Pocket PC*
- Click on the underlined text to view an image (or images) relevant to the text concerned
- Uses Mobipocket Reader technology, compatible with all PDA devices and also available for Windows
- Follow the on-screen instructions on the relevant part of the CD Rom to install Mobipocket for your device
- Full title available in this format for purchase as a download from http://www.pda.bmjbooks.com

BMJ Books catalogue
- Instant access to BMJ Books full catalogue, including an order form

Instructions for use

The CD Rom should start automatically upon insertion, on all Windows systems. The menu screen will appear and you can then navigate by clicking on the headings. If the CD Rom does not start automatically upon insertion, please browse using "Windows Explorer" and double-click the file "BMJ_Books.exe".

Tips

To minimise the bookmarks pane so that you can zoom the page to full screen width, simply click on the "Bookmarks" tab on the left of your screen. The bookmarks can be accessed again at any time by simply clicking this tab again.
To search the text simply click on "Search Text", then type into the window provided. You can stop the search at any time by clicking "Stop Search", and can then navigate directly to a search result by double-clicking on the specific result in the Search pane. By clicking your left mouse button once on a page in the PDF ebook window, you "activate" the window. You can now scroll through pages uses the scroll-wheel on your mouse, or by using the cursor keys on your keyboard.

Note: the ABC of Dermatology PDF eBook is for search and reference only and cannot be printed. A printable PDF version as well as the full PDA edition can be purchased from http://www.bmjbookshop.com

Troubleshooting

If any problems are experienced with use of the CD Rom, we can give you access to all content* via the internet. Please send your CD Rom with proof of purchase to the following address, with a letter advising your email address and the problem you have encountered:

ABC of Dermatology eBook access
BMJ Bookshop
BMA House
Tavistock Square
London
WC1H 9JR

*Unfortunately, due to technical limitations, this offer currently excludes the artwork slideshow

ABC OF DERMATOLOGY

Fourth Edition

PAUL K BUXTON
Consultant Dermatologist
Royal Infirmary, Edinburgh

BMJ Books

First published by the BMJ Publishing Group Ltd in 1988
Second edition 1993
Third edition 1998
Hot Climates edition 1999
Fourth edition 2003

BMJ Publishing Group Ltd, BMA House, Tavistock Square,
London WC1H 9JR

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0-7279-1696-3

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed and bound in Malaysia by Times Offset
Cover picture is a light micrograph of a vertical section through a human skull showing several hair follicles. With permission of Dr Clive Kocher/Science Photo Library

Contents

Contributors

R Balfour
General Practitioner, Edinburgh

R StC Barnetson
Professor of Dermatology, Department of Dermatology, Prince Albert Hospital, Camperdown, Australia

E Crawford
General Practitioner, Edinburgh

DJ Gawkrodger
Consultant Dermatologist, Royal Hallamshire Hospital, Sheffield

DWS Harris
Consultant Dermatologist, Whittington Hospital, London

RJ Hay
Dean, Faculty of Medicine and Health Sciences and Professor of Dermatology, Queens University, Belfast

D Kemmett
Consultant Dermatologist, Lothian University NHS Trust, Edinburgh

B Leppard
Professor, Regional Dermatology Training Centre, Moshi, Tanzania

JA Savin
Consultant Dermatologist, Lothian University NHS Trust, Edinburgh

MA Waugh
Consultant in Genitourinary Medicine, Leeds Teaching Hospitals NHS Trust, Leeds

AL Wright
Consultant Dermatologist, Bradford Royal Infirmary

Acknowledgements

Professor R StC Barnetson, University of Sydney, Australia, wrote the original chapter on the sun and the skin, which is included in this edition. Professor Barbara Leppard, Regional Dermatology Training Centre, Moshi, Tanzania, has contributed a chapter on tropical dermatology with her own illustrations and some from Professor Barnetson. Professor R Hay, St Johns Institute of Dermatology, UMDS, Guy's Hospital, London, extensively revised the section on bacterial and fungal infections and provided some illustrations. Dr JA Savin, Lothian University NHS Trust, Edinburgh, rewrote the section on genetics and skin disease. Dr MA Waugh, consultant in GU medicine, The Leeds Teaching Hospitals NHS Trust, provided material and illustrations on AIDS. Dr Robin Balfour and Dr Ewan Crawford, general practitioners in Edinburgh, provided contributions on dermatology in general practice.

Material from contributors to earlier editions has been retained, particularly that supplied by Dr DJ Gawkrodger, consultant dermatologist, Royal Hallamshire Hospital, Sheffield (autoimmunity), Dr DWS Harris, consultant dermatologist, Whittington Hospital, London (practical procedures), Dr D Kemmett, consultant dermatologist, Lothian University NHS Trust, Edinburgh (diseases of hair and scalp), Dr AL Wright, consultant dermatologist, Bradford Royal Infirmary (diseases of nails).

The illustrations come from the Fife hospitals, the Royal Infirmary Edinburgh and the author's own collection. Some specific illustrations have been donated by Dr JA Savin (flea bites on the ankle); Dr Peter Ball (rubella); Professor CV Ruckley (varicose veins); Dr GB Colver (spider naevus); Dr MA Waugh and Dr M Jones (AIDS); Dr PMW Copemen (dermatoses in black skin). Miss Julie Close made the diagrams of the nail and types of immune response. The illustrations for dermatology in general practice were produced by Sister Sheila Robertson, Dermatology Liaison Nurse in Fife and Julie Close. The text of the third edition, on which this one is based, was typed by Mrs Mary Henderson. I would also like to thank Pat Croucher, who proofread the third edition, for copy-editing the script for this edition with perception and patience. Sally Carter and the editorial staff at BMJ Books gave great help and support.

Finally thanks are due to all the hospital staff—and particularly the patients—without whom dermatology could not be practised at all.

Preface

The remit for the first edition of the *ABC of Dermatology* in 1987 was that it should concentrate on common conditions and give down to earth advice. The ABC format proved well suited for this and there has been a steady demand for the book since then. In this edition the same approach is maintained while taking into account advances in diagnosis and treatment. Research in genetics and immunology is providing ever-increasing insights into the mechanisms that underlie clinical changes, and has led to more accurate diagnosis and more rational treatment. Specialised techniques that may not be relevant to common conditions can be of the greatest importance to an individual patient with a rare disease. In epidermolysis bullosa, for example, the ability to differentiate accurately between the different types with electronmicroscopy and immunohistochemistry is of considerable significance. Generally research increases our understanding of *how* diseases arise, but we have to admit to ourselves and our patients that *why* they occur remains as elusive as ever.

In recent years the management of inflammatory skin conditions has become both more effective and less demanding for the patient. In addition there is greater recognition of the impact of skin diseases on the patient's life. Major advances in treatment include more effective and safer phototherapy and the use of immunosuppressive drugs that enable inflammatory dermatoses to be managed without the need to attend for dressings or admission to hospital. This is just as well, since dermatology inpatient beds are no longer available in many hospitals. As a consequence, more dermatology patients are managed in the community with a greater role for the community nurse and general practitioner or family doctor. Dermatology liaison nurses play a very important part in making sure that the patients are using their treatment effectively at home and in maintaining the link between the hospital department, the home situation, and the general practitioner. Self-help groups are a valuable resource of support for patients, and there is now much more information available to the public on the recognition and management of skin disease.

Progress has been made in increasing the awareness of the general public and the politicians (who control the resources for health care) of the importance of skin diseases. In countries with minimal medical services there are immense challenges—particularly the need for training medical workers in the community who can recognise and treat the most important conditions. This has a major impact on the suffering and disability from skin diseases. The International Foundation for Dermatology and the pioneering Regional Dermatology Training Centre in Moshi, Tanzania, have set an important lead in this regard.

All the chapters have been revised for this new edition and a number of new illustrations included. A new chapter on tropical dermatology, which was previously included in the "hot climates" Australasian edition, is incorporated. In addition, there is a chapter on dermatology in general practice. Colleagues with special areas of expertise have been generous in giving advice and suggestions for this edition, which I trust will be a means of introducing the reader to a fascinating clinical discipline, covering all age groups and relevant to all areas of medicine.

Edinburgh, 2003, Paul Buxton

1 Introduction

The object of this book is to provide the non-dermatologist with a practical guide to the diagnosis and treatment of skin conditions. One advantage of dealing with skin conditions is that the lesions are easily examined and can be interpreted without the need for complex investigations, although a biopsy may be required to make or confirm the diagnosis. An understanding of the microscopic changes underlying the clinical presentation makes this interpretation easier and more interesting.

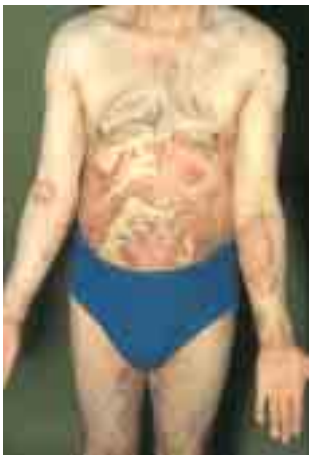
In the early chapters the relationship between the clinical presentation and the underlying pathological changes is discussed for a few important conditions, such as psoriasis. These are then used as a model for comparison with other skin diseases. This approach is suitable for skin conditions that present with characteristic lesions.

In other disorders a variety of causes may produce the same type of lesion. In this case it is more helpful to describe the characteristic clinical pattern that results. For example, similar inflamatory changes may result from drug allergy, autoimmune disease, or infection.

Tumours, acne, and leg ulcers are covered as separate subjects, as are diseases of the hair and nails.

The same condition is sometimes dealt with in more than one section, for example, fungal infections are discussed under "Rashes with epidermal changes" and again under "Fungal and yeast infections", giving different perspectives of the same disorder.

Skin lesions are sometimes an indication of internal disease and may be the first clinical sign. For example, the girl in the photograph presented with a rash on her face, made worse by sunlight. She then mentioned that she was aware of lassitude, weight loss, and vague musculoskeletal symptoms which, in conjunction with the appearance of the rash, suggested lupus erythematosus. This was confirmed by further investigations and appropriate treatment was initiated. Other dermatological associations with systemic disease are discussed in the relevant sections.



Psoriasis—large legions

Lupus erythematosus

The significance of skin disease

A large proportion of the population suffers from skin diseases, which make up about 10% of all consultations in primary care in the United Kingdom. However, community studies show that over 20% of the population have a medically significant skin condition and less than 25% of these consulted a doctor.

The skin is not only the largest organ of the body, it also forms a living biological barrier and is the aspect of ourselves we present to the world. It is therefore not surprising that there is great interest in "skin care", with the associated vast cosmetic industry. The impairment of the normal functions of the skin can lead to acute and chronic illness with considerable disability and sometimes a need for hospital treatment.

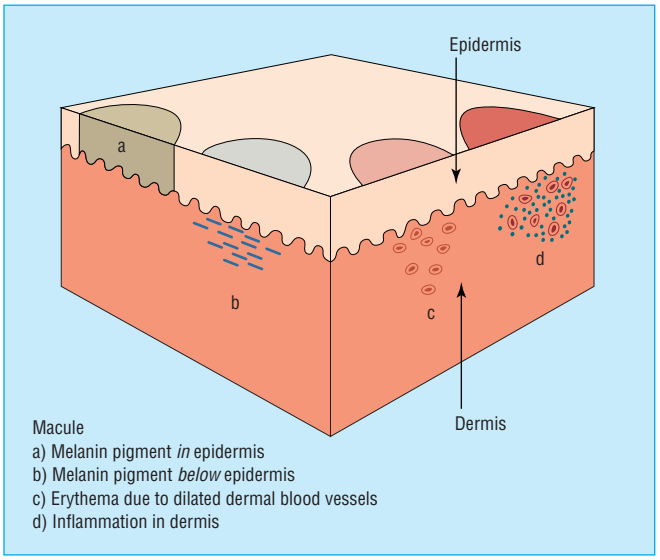
A wide variety of tumours, both benign and malignant, arise in the skin. Fortunately the majority are harmless and most moles never develop dysplastic change.

Most cancers arising in the skin remain localised and are only invasive locally, but others may metastasise. It is important therefore to recognise the features of benign and malignant tumours, particularly those, such as malignant melanoma, that

Skin tags—examples of benign tumours

can develop widespread metastases. Recognition of typical benign tumours saves the patient unneccessary investigations and the anxiety involved in waiting for results.

Although a wide range of internal diseases produce physical signs in the skin, most skin diseases do not themselves have serious physical effects. However there can be significant psychological effects and problems with personal relationships, employment, and sporting activity. It is therefore important to use what Dr Papworth called "wide angle lenses" in assessing the patient and their disease. So, in addition to concentrating on the skin changes, the overall health and demeanour of the patient should be taken into account. This also means making sure that there are no other signs, such as involvement of the nails, mucous membranes, or other parts of the skin. The general physical condition and psychological state of the patient should be assessed, with more specific examination if indicated.

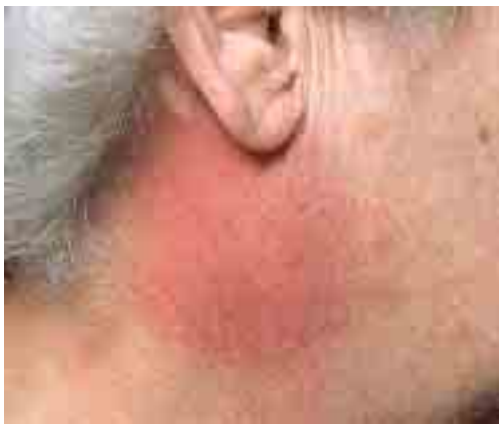


Section through skin

Descriptive terms

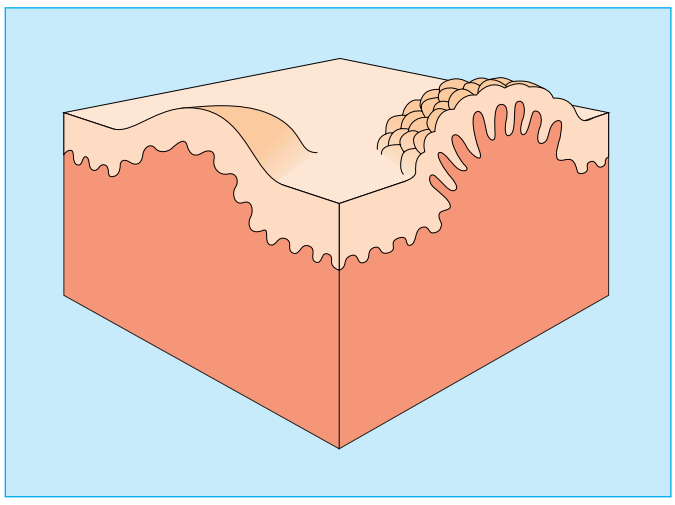
All specialties have their own common terms, and familiarity with a few of those used in dermatology is a great help. The most important are defined below.

Macule
Derived from the Latin for a stain, the term macule is used to describe changes in colour or consistency without any elevation above the surface of the surrounding skin. There may be an increase of melanin, giving a black or blue colour depending on the depth of the pigment. Loss of melanin leads to a white macule. Vascular dilatation and inflammation produce erythema.

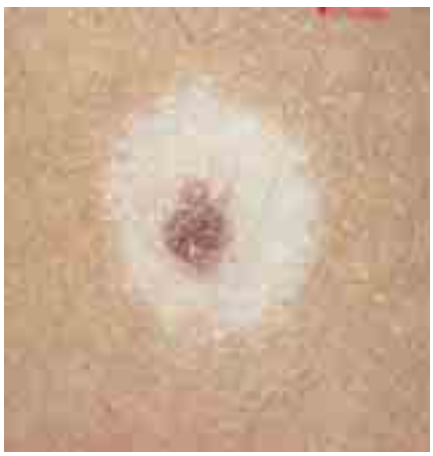


Eythema

Papules and nodules
A papule is a circumscribed, raised lesion, conventionally less than 1 cm in diameter. It may be due to either epidermal or dermal changes.



Section through skin with a papule

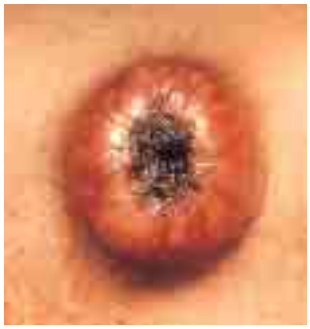


A papule surrounded by a depigmented macule

A nodule is similar to a papule but over 1 cm in diameter. A vascular papule or nodule is known as an haemangioma.

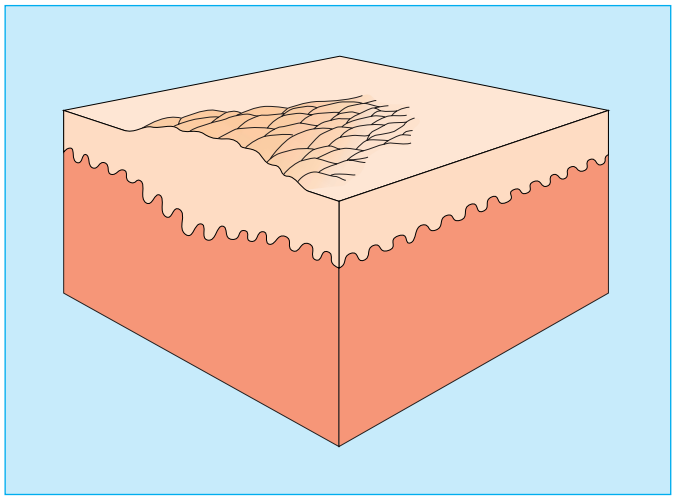
Plaque

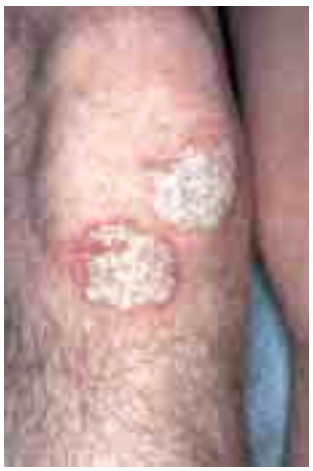
Plaque is one of those terms which conveys a clear meaning to dermatologists but is often not understood by others. To take it literally, one can think of a commemorative plaque stuck on the wall of a building, with a large area relative to its height and a well defined edge. Plaques are most commonly seen in psoriasis.



Papule

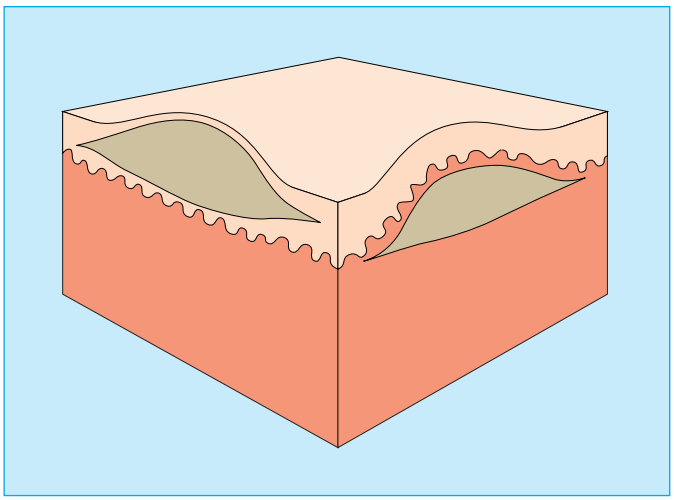
Haemangioma

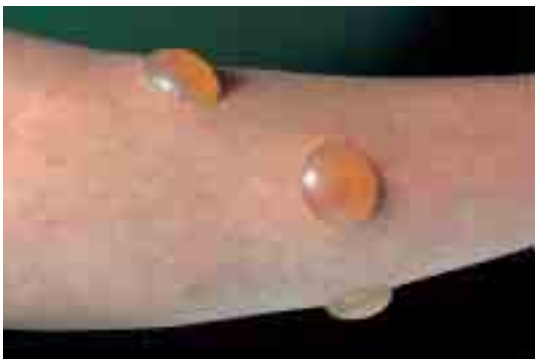

Section through skin with plaque


Plaques in psoriasis

Vesicles and bullae

Vesicles and bullae are raised lesions that contain fluid. A bulla is a vesicle larger than 0.5 cm. They may be superficial within the epidermis or situated in the dermis below it.


Section through skin showing situations of vesicle and bulla


Acute reaction to insect bite—bullae

Lichenification

Lichenification is another term frequently used in dermatology as a relic of the days of purely descriptive medicine. Some resemblance to lichen seen on rocks and trees does occur, with hard thickening of the skin and accentuated skin markings. It is most often seen as a result of prolonged rubbing of the skin in localised areas of eczema.


Lichen simplex

Nummular lesions
Nummular literally means a "coin-like" lesion. There is no hard and fast distinction from *discoid* lesions, which are flat disc-like lesions of variable size. It is most often used to describe a type of eczematous lesion.

Pustules
The term pustule is applied to lesions containing purulent material—which may be due to infection, as in the case shown—or sterile pustules, which are seen in pustular psoriasis.

Numular lesion as a response to a vaccination site in the arm

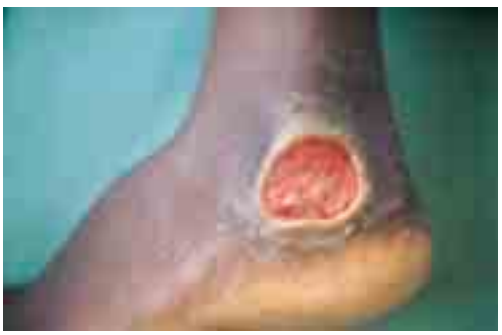
Pustule due to infection

Atrophy
Atrophy refers to loss of tissue which may affect the epidermis, dermis, or subcutaneous fat. Thinning of the epidermis is characterised by loss of the normal skin markings, and there may be fine wrinkles, loss of pigment, and a translucent appearance. There may be other changes as well, such as sclerosis of the underlying connective tissue, telangiectasia, or evidence of diminished blood supply.

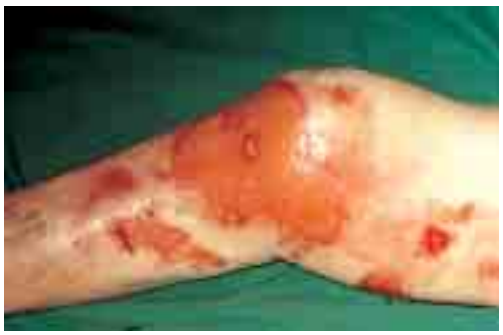
Ulceration
Ulceration results from the loss of the whole thickness of the epidermis and upper dermis. Healing results in a scar.

Erosion
An erosion is a superficial loss of epidermis that generally heals without scarring.

Epidermal atrophy



Tropical ulcer



Bullous pemphigoid causing erosion

Excoriation
Excoriation is the partial or complete loss of epidermis as a result of scratching.

Fissuring
Fissures are slits through the whole thickness of the skin.

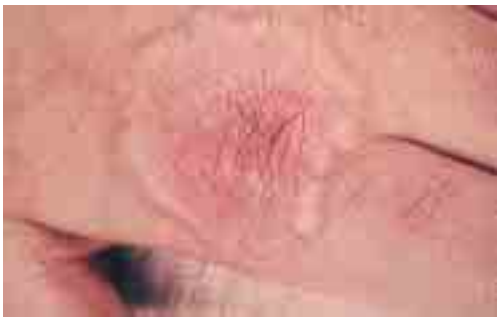
Excoriation of epidermis

Hyperkeratosis with fissures

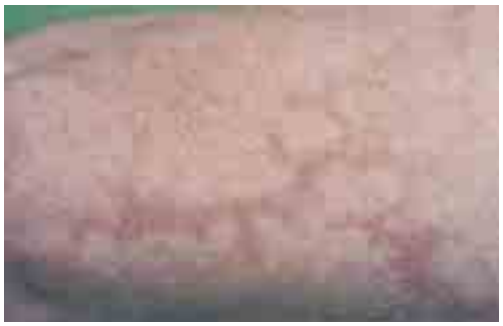
Desquamation
Desquamation is the peeling of superficial scales, often following acute inflammation.

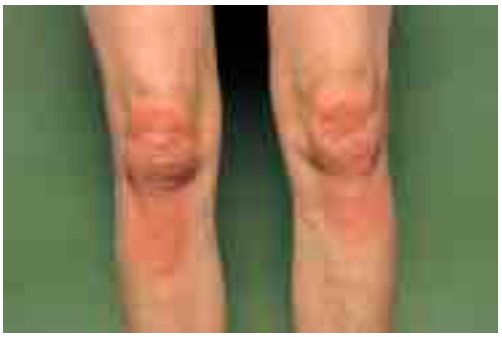
Annular lesions
Annular lesions are ring shaped lesions.

Reticulate
The term reticulate means “net-like”. It is most commonly seen when the pattern of subcutaneous blood vessels becomes visible.


Ring-shaped annular lesion


Desquamation


Reticulate pattern on skin


Psoriasis of both legs

Rashes

Approach to diagnosis
A skin rash generally poses more problems in diagnosis than a single, well defined skin lesion such as a wart or tumour. As in all branches of medicine a reasonable diagnosis is more likely to be reached by thinking firstly in terms of broad diagnostic categories rather than specific conditions.

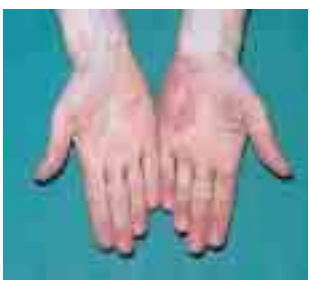
There may have been previous episodes because it is a constitutional condition, such as atopic eczema. In the case of contact dermatitis, regular exposure to a causative agent leads to recurrences that fit with the times of exposure and this is usually apparent from the history. Endogenous conditions such as psoriasis can appear in adults who have had no previous episodes. If there is no family history and several members of the household are affected, a contagious condition, such as scabies, should be considered. A common condition with a familial tendency, such as atopic eczema, may affect several family members at different times.

A simplistic approach to rashes is to clarify them as being from “inside” or “outside”. Examples of “inside” or endogenous rashes are atopic eczema or drug rashes, whereas fungal infection or contact dermatitis are “outside” rashes.

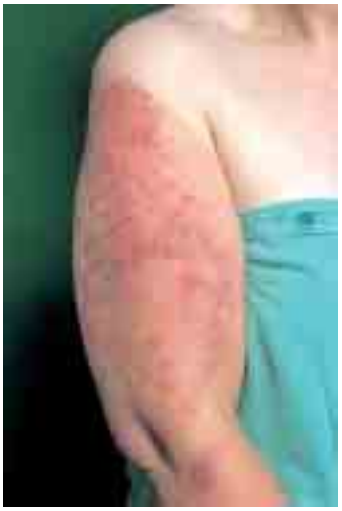
Symmetry
Most endogenous rashes affect both sides of the body, as in the atopic child or a man with psoriasis on his knees. Of course, not all exogenous rashes are asymmetrical. A seamstress who uses scissors in her right hand may develop an allergy to metal in this one hand, but a hairdresser or nurse can develop contact dermatitis on both hands.

Diagnosis of rash
- Previous episodes of the rash, particularly in childhood, suggest a constitutional condition such as atopic eczema
- Recurrences of the rash, particularly in specific situations, suggests a contact dermatitis. Similarly a rash that only occurs in the summer months may well have a photosensitive basis
- If other members of the family are affected, particularly without any previous history, there may well be a transmissible condition such as scabies


Contact dermatitis as a response to mascara


Irritant dermatitis

Distribution
It is useful to be aware of the usual sites of common skin conditions. These are shown in the appropriate chapters. Eruptions that appear only on areas exposed to sun may be entirely or partially due to sunlight. Some are due to a sensitivity to sunlight alone, such as polymorphous light eruption, or a photosensitive allergy to topically applied substances or drugs taken internally.



Allergic reaction producing photosensitivity

Morphology
The appearance of the skin lesion may give clues to the underlying pathological process.

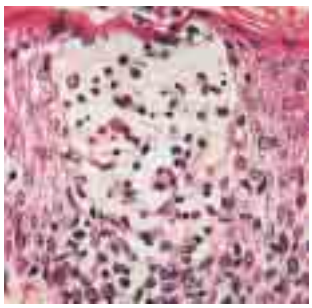
The surface may consist of normal epidermis overlying a lesion in the deeper tissues. This is characteristic of many types of erythema in which there is dilatation of the dermal blood vessels associated with inflammation. The skin overlying cysts or tumours in the dermis and deeper tissues is usually normal. Conditions affecting the epidermis will produce several visible changes such as thickening of the keratin layer and scales in psoriasis or a more uniform thickening of the epidermis in areas lichenified by rubbing. An eczematous process is characterised by small vesicles in the epidermis with crusting or fine scaling.

The margin of some lesions is very well defined, as in psoriasis or lichen planus, but in eczema it merges into normal skin.

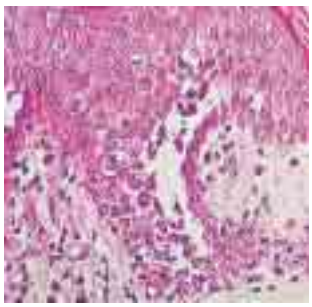
Blisters or vesicles occur as a result of (a) oedema between the epidermal cells or (b) destruction of epidermal cells or (c) the result of separation of the epidermis from the deeper tissues. Of course, more than one mechanism may occur in the same lesion. Oedema within the epidermis is seen in endogenous eczema, although it may not be apparent clinically, particularly if it is overshadowed by inflammation and crusts. It is also a feature of contact dermatitis.

Lesion in deeper tissue with normal epidermis

Small vesicles of eczema



Eczema—intraepidermal vesicle



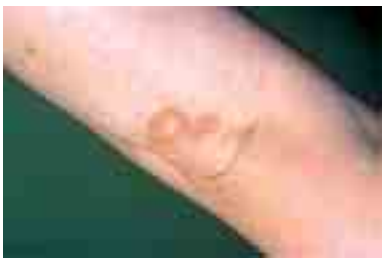
Pemphigus—destruction of epidermal cells

Pemphigoid—blister forming below epidermis

Blisters occur in:

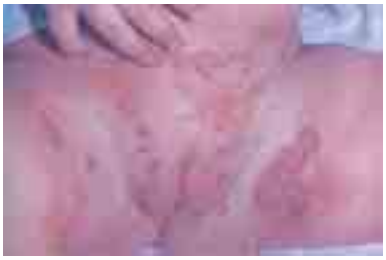
- *viral diseases* such as chickenpox, hand, foot and mouth disease, and herpes simplex
- *bacterial infections* such as impetigo
- eczema and contact dermatitis
- *primary blistering* disorders such as dermatitis herpetiformis, pemphigus and pemphigoid as well as metabolic disorders such as porphyria.

Herpes simplex



Impetigo

Pemphigoid



Impetigo

Bullae, blisters over 0.5 cm in diameter, may occur in congenital conditions (such as epidermolysis bullosa), lichen planus, and pemphigoid without much inflammation. However, those forming as a result of vasculitis, sunburn, or an allergic reaction may be associated with pronounced inflammation. In pustular psoriasis there are deeper pustules, which contain polymorphs but are sterile and show little inflammation. Drug rashes can appear as a bullous eruption.

Induration is thickening of the skin due to infiltration of cells, granuloma formation, or deposits of mucin, fat, or amyloid.

Inflammation is indicated by erythema, which may be accompanied by increased temperature if acute—for example, in cellulitis or erythema nodosum. There may be a chronic inflammatory infiltrate in, for example, conditions such as lichen planus or lupus erythematosus.

Assessment of the patient

As well as assessing the clinical changes, the effect of a skin condition on the patient's life and their attitude to it must always be taken into account. For example, severe pustular psoriasis of the hands can be devastating for a self employed electrician and total hair loss from the scalp very distressing for a 16 year old girl.

Fear that a skin condition may be due to cancer or infection is often present and reassurance should always be given whether asked for or not. If there is the possibility of a serious underlying cause that requires further investigation, it is part of good management to answer any questions the patient has and provide an explanation that he or she can understand. It is easy to forget this aspect of medical practice at times.

The significance of occupational factors must be taken into account. In some cases, such as an allergy to hair dyes in a hairdresser, it may be impossible for the patient to continue their job. In other situations the allergy can be easily avoided.

Patients understandably ask whether psoriasis can be cured and often want to know the cause. The cause is unknown and the best answer is that the tendency to develop psoriasis is part of a person's constitution and some factor triggers the development of the clinical lesions. Known factors include physical or emotional stress, local trauma to the skin (Koebner's phenomenon), infection (in guttate psoriasis), drugs (β blockers, lithium, and antimalarial drugs).

To illustrate the use of these basic concepts in the diagnosis of lesions in practice two common skin diseases are considered—*psoriasis*, which affects 1–2% of the population, and *eczema*, an even more common complaint. Both are rashes with distinctive epidermal changes. The difficulty arises with the unusual lesion: Is it a rarity or a variation of a common disease? What should make us consider further investigation? Is it safe to wait and see if it resolves or persists? The usual clinical presentations of psoriasis and eczema are also used as a basis for comparison with variations of the usual pattern and other skin conditions.

A relevant history should be taken in relation to occupational and environmental factors

- Where? Site of initial lesion(s) and subsequent distribution
- How long? Has condition been continuous or intermittent?
- Prognosis—Is it getting better or worse?
- Previous episodes—How long ago? Were they similar? Have there been other skin conditions?
- Who else? Are other members of the family affected? Or colleagues at work or school?
- Other features—Is there itching, burning, scaling, or blisters? Any association with drugs or other illnesses?
- Treatment—By prescription or over the counter? Have prescribed treatments actually been used?

The following points are helpful when examining skin lesions

Distribution

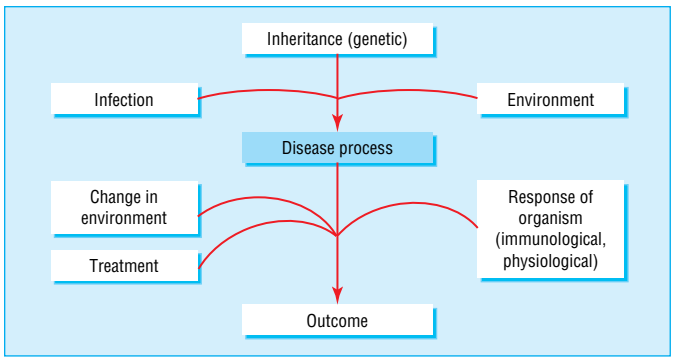
- This may give the essential clue, so a full examination is necessary. For example, there are many possible causes for dry thickened skin on the palms, and finding typical psoriasis on the elbows, knees, and soles may give the diagnosis

Morphology

- Are the lesions dermal or epidermal? Macular (flat) or forming papules? Indurated or forming plaques? With a well defined edge? Forming crusts, scabs, or vesicles?

Pattern

- This is the overall clinical picture of both morphology and distribution. For example, an indeterminate rash may be revealed as pityriasis rosea when the "herald patch" is found



Factors possibly affecting development of skin disease such as psoriasis

Further reading

Braun-Falco O, Plewig G, Wolff HH, Winkelmann RK. *Dermatology*. Berlin: Springer-Verlag, 1991

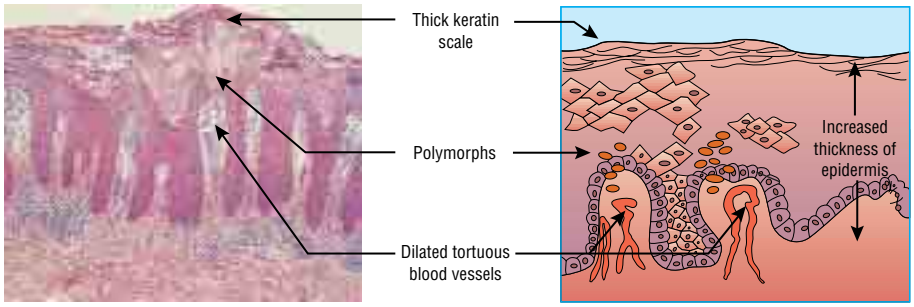
Champion RH, Burton JL, Ebling FJ. *Textbook of dermatology*. 6th ed. Oxford: Blackwell Scientific, 1998

Fitzpatrick TB, Freedberg IM, Eisen AZ, Austen KF, Wolff K. *Dermatology in general medicine*. 4th ed. New York: McGraw-Hill, 1993

Sams WM, Lynch PJ, eds. *Principles and practice of dermatology*, 2nd ed. New York: Churchill Livingstone, 1996

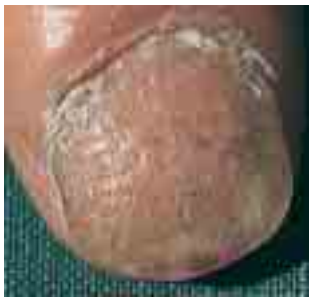
2 Psoriasis

The familiar pink or red lesions with a scaling surface and well defined edge are easily recognised. These changes can be related to the histological appearance:



Increased epidermal proliferation—nuclei found … throughout the epidermis

- The increased thickness of the epidermis, presence of nuclei above the basal layer, and thick keratin are related to increased epidermal turnover.
- Because the epidermis is dividing it does not differentiate adequately into normal keratin scales. These are readily removed to reveal the tortuous blood vessels beneath, appearing clinically as "Auspitz sign". The psoriatic plaque can be likened to a brick wall badly built by a workman in too much of a hurry—it may be high but it is easily knocked down.
- The polymorphs that migrate into the epidermis form sterile pustules in pustular psoriasis. These are most commonly seen on the palms and soles.
- The dilated blood vessels can be a main feature, giving the clinical picture of intense erythema.



Pitting of the nail

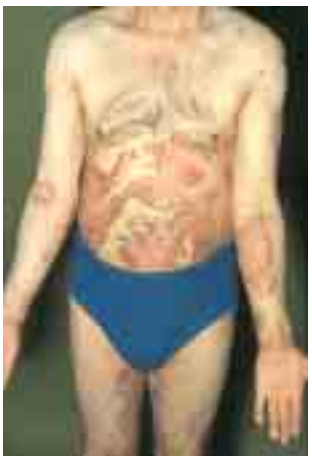
The equivalent changes in the nail cause thickening and "pits" 0.5–1.0 mm in diameter on the surface; these are thought to be due to small areas of psoriatic changes in the upper layer of the nail plate that then fall out. Onycholysis, in which the nail plate is raised up, also occurs in psoriasis.

Clinical appearance

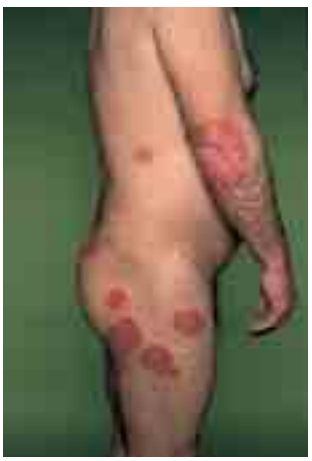
The main characteristics of psoriatic lesions, which reflect the pathological processes listed above, follow.

Plaques consisting of well defined raised areas of psoriasis. These may be few or numerous, covering large areas of the trunk and limbs. Sometimes there are large confluent lesions.

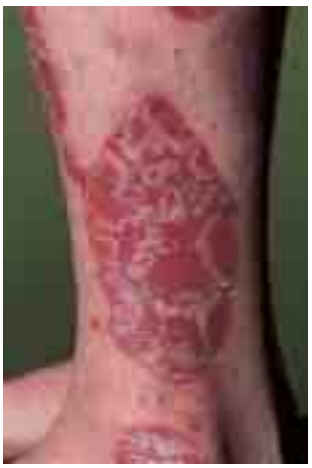
Small lesions



Large lesions



Plaques



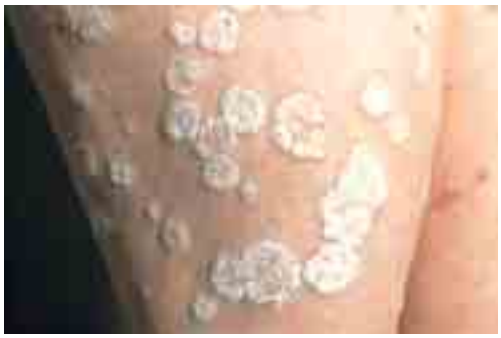
Plaques

Scaling may predominate, giving a thick plaque, which is sometimes likened to limpets on the sea shore, hence the name "rupioid". Scratching the surface produces a waxy appearance—the "tache de bougie" (literally "a line of candle wax").

Erythema may be conspicuous, especially in lesions on the trunk and flexures.

Pustules are rare on the trunk and limbs, but deep seated pustules on the palms and soles are fairly common. In the form of *palmo-plantar pustulosis* they may occur without psoriasiatic lesions elsewhere.

The *size* of the lesions varies from a few millimetres to very extensive plaques.



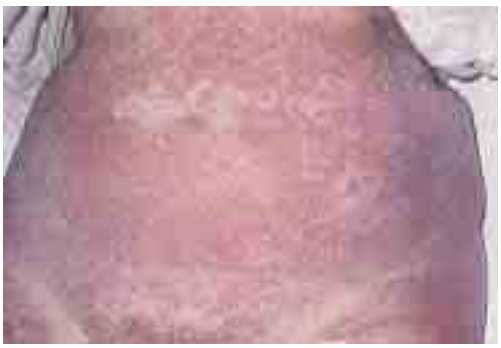
Rupoid lesions

The typical patient

Psoriasis usually occurs in early adult life, but the onset can be at any time from infancy to old age, when the appearance is often atypical.

The following factors in the history may help in making a diagnosis:

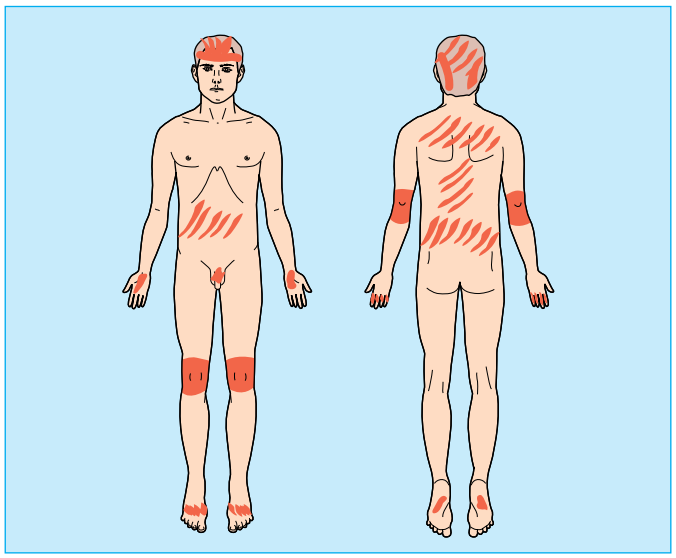
- There may be a family history—if one parent has psoriasis 16% of the children will have it, if both parents, the figure is 50%.
- The onset can occur after any type of stress, including infection, trauma, or childbirth.
- The lesions may first appear at sites of minor trauma—Koebner's phenomenon.
- The lesions usually clear on exposure to the sun.
- Typically, psoriasis does not itch.
- There may be associated arthropathy—affecting either the fingers and toes or a single large joint.



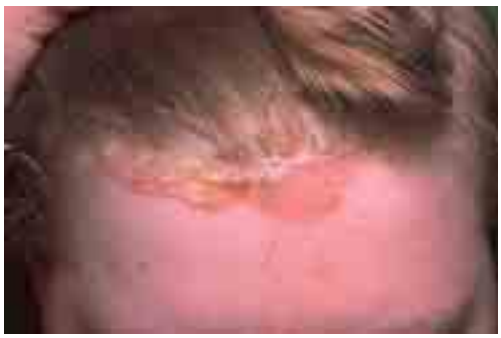
Widespread pustular psoriasis

Clinical presentation

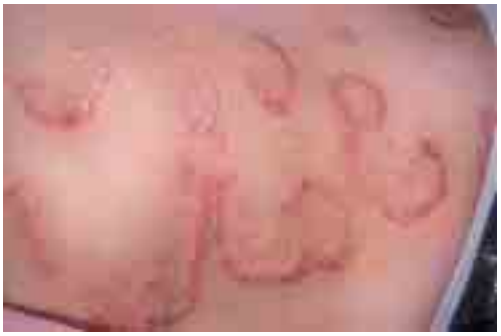
Patients usually present with lesions on the elbows, knees, and scalp. The trunk may have plaques of variable size and which are sometimes annular. Patients with psoriasis show Koebner's phenomenon with lesions developing in areas of skin trauma such as scars or minor scratches. Normal everyday trauma such as handling heavy machinery may produce hyperkeratotic lesions on the palms. In the scalp there is scaling, sometimes producing very thick accretions. Erythema often extends beyond the hair margin. The nails show "pits" and also thickening with separation of the nail from the nail bed (oncholysis).



Common patterns of distribution in psoriasis



Scalp psoriasis



Annular lesions

Koebner's phenomenon: psoriasis in surgical scar

Psoriasis of the nail

Psoriasis of the hand

Guttate psoriasis—from the Latin *gutta*, a drop—consists of widespread small pink macules that look like drops of paint. It usually occurs in adolescents and often follows an acute β haemolytic streptococcal infection. There may be much distress to both parent and child when a previously healthy adolescent erupts in spots. Fortunately it also resolves quite rapidly.

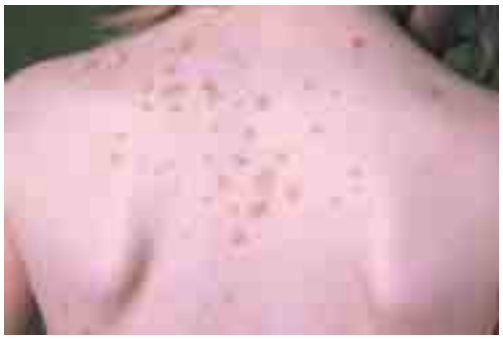
Pustular lesions occur as either chronic deep seated lesions or generalised pustular psoriasis.

Chronic deep seated lesions occur on the palms and soles with surrounding erythema which develops a brown colour and scaling. It is important to reassure the patient that, despite their appearance, these pustules are not infectious—they consist of sterile collections of polymorphs.

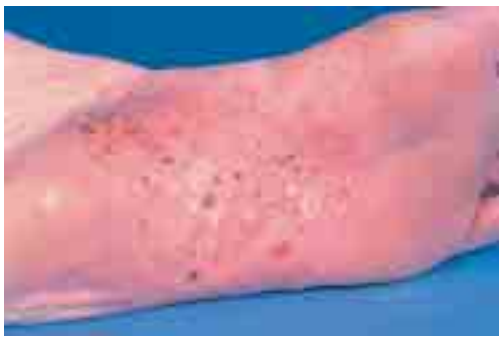
These lesions occur in an older age group than psoriasis, and psoriasis may not be present elsewhere. It is more common in smokers. *Acrodermatitis pustulosa* is a variant that occurs in a younger age group in which there are pustules and inflammation arround the nails and the fingertips.

Generalised pustular psoriasis is uncommon. Pustules develop in association with erythema. It may be precipitated by the use of steroids.

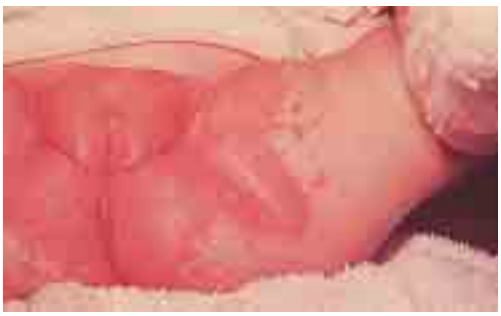
Flexural psoriasis produces well defined erythematous areas in the axillae and groins and beneath the breasts. Scaling is minimal or absent. It must be distinguished from a fungal infection and it is wise to send specimens for mycology if there is any doubt.



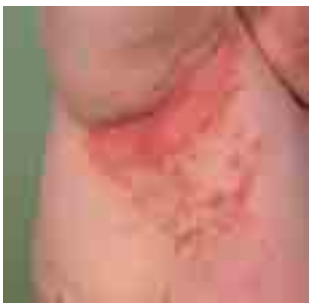
Guttate psoriasis



Pustules on the foot



Napkin psoriasis

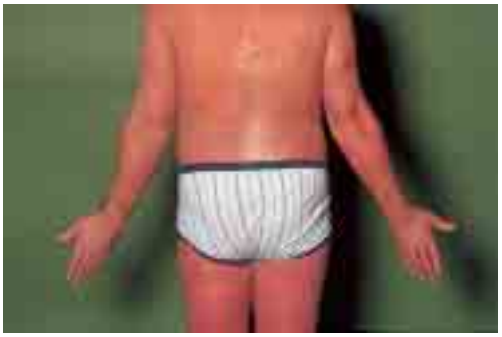


Flexural psoriasis

Napkin psoriasis in children may present with typical psoriatic lesions or a more diffuse erythematous eruption with exudative rather than scaling lesions.

Erythrodermic psoriasis is a serious, even life threatening, condition with erythema affecting nearly the whole of the skin. Diagnosis may not be easy as the characteristic scaling of psoriasis is absent, although this usually precedes the erythroderma. Less commonly the erythema develops suddenly without preceding lesions. There is a considerable increase in cutaneous blood flow, heat loss, metabolism, and water loss.

It is important to distinguish between the *stable*, chronic, plaque type of psoriasis, which is unlikely to develop exacerbations and responds to tar, dithranol, and ultraviolet treatment, and the more *acute* erythematous type, which is unstable and likely to spread rapidly. The use of tar, dithranol, or ultraviolet light can irritate the skin and will make it more widespread and inflamed.



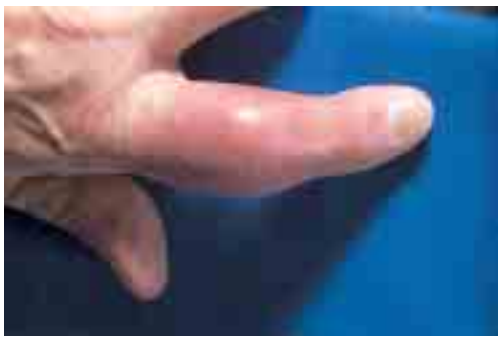
Erythrodermic psoriasis

Joint disease in psoriasis

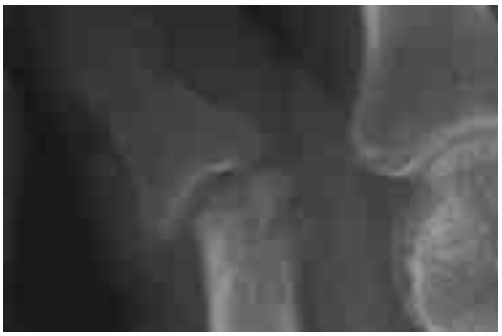
Patients with seronegative arthropathy of the non-rheumatoid type show double the normal (2%) incidence of psoriasis. Psoriatic arthropathy commonly affects the distal interphalangeal joints, sparing the metacarpophalangeal joints, and is usually asymmetrical. Radiological changes include a destructive arthropathy with deformity. Rheumatoid nodules are absent. The sex ratio is equal but a few patients develop a "rheumatoid-like" arthropathy, which is more common in women than in men. A third rare group have arthritic changes in the larger joints, where there is considerable resorption of bone. Other members of the families of those with psoriatic arthropathy are affected in 40% of cases.

There may be pustular psoriasis of the fingers and toes associated with arthropathy which can be sufficiently severe to immobilise the patient.

Both psoriatic arthropathy and Reiter's syndrome are associated with the presence of HLA B27. Reiter's syndrome is characterised by polyarthritis and the development of urethritis, inflammatory changes in the conjunctivae, and skin lesions including pustulosis hyperkeratosis of the soles.



Acute arthropathy

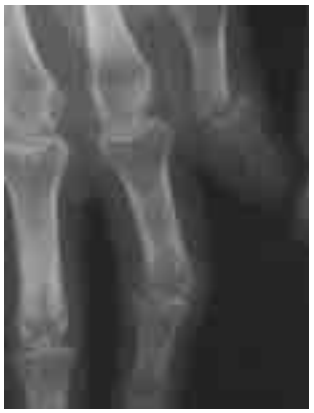


Acute arthropathy—*X* ray signs

Causes of psoriasis

The cause is unknown but there is an inherited predisposition. The strong genetic influence may result from a single dominant gene with poor penetrance or a number of genetic influences. Other factors such as local trauma, general illness and stress are also involved, so the cause of psoriasis is best regarded as being multifactorial. HLA-Cw6 is the phenotype most strongly associated with psoriasis, particularly the early onset variety in which hereditary factors seem to play the greatest part. There is an increase in HLA expression in psoriatic arthropathy.

Local trauma, acute illness, and stress may be factors in causing the appearance of clinical lesions. β Haemolytic streptococcal throat infection is a common precipitating factor in guttate psoriasis. Antimalarial drugs, lithium, and β blockers can make psoriasis worse. There is evidence that psoriasis occurs more readily and is more intractable in patients with a high intake of alcohol. Smoking is associated with palmo-plantar pustulosis.

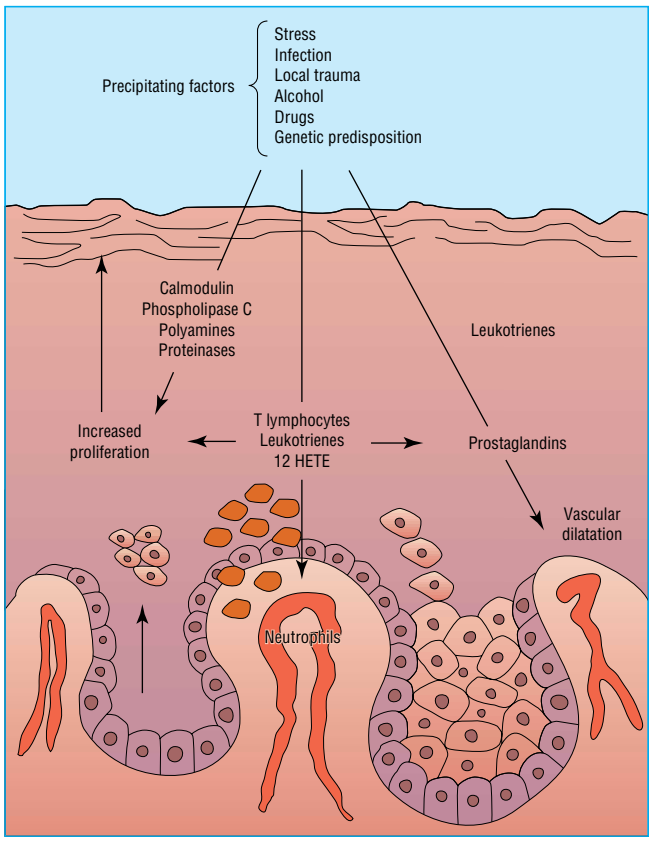


Acute arthropathy—*X* ray signs

There is evidence that both hormonal and immunological mechanisms are involved at a cellular level. The raised concentrations of metabolites of arachodonic acid in the affected skin of people with psoriasis are related to the clinical changes. Prostaglandins cause erythema, whereas leukotrienes (LTB4 and 12 HETE) cause neutrophils to accumulate. The common precursor of these factors is phospholipase A2, which is influenced by calmodulin, a cellular receptor protein for calcium. Both phospholipase A2 and calmodulin concentrations are raised in psoriatic lesions.

T helper lymphocytes have been found in the dermis as well as antibodies to the basal cell nuclei of psoriatic skin. In addition, dermal factors contribute to the development of psoriatic lesions.

The detailed treatment of psoriasis is covered in the next chapter. The only point to be made here is the importance of encouraging a positive attitude with expectation of improvement but not a permanent cure, since psoriasis can recur at any time. Some patients are unconcerned about very extensive lesions whereas to others the most minor lesions are a catastrophe.



Hormonal and immunological mechanisms and dermal factors involved in the development of psoriasis

Further reading

Farber EM. *Psoriasis*. Amsterdam: Elsevier, 1987
Fry L. *An atlas of psoriasis*. London: Parthenon Publications, 1992
Mier PD, Van de Kerkhof PC. *Textbook of psoriasis*. Edinburgh: Churchill Livingstone, 1986
Roenigk HH, Maibach HI. *Psoriasis*. Basle: Dekker, 1991

3 Treatment of psoriasis

It vanished quite slowly, beginning with the end of the tail, and ending with the grin, which remained some time after the rest of it had gone.
Lewis Carroll, *Alice in Wonderland*

To ignore the impact of the condition on the patient's life is to fail in treating psoriasis. Like the Cheshire cat that Alice met, it tends to clear slowly and the last remaining patches are often the hardest to clear. This is frustrating enough, but there is also the knowledge that it will probably recur and need further tedious courses of treatment, so encouragement and support are an essential part of treatment.

In an attempt to quantify the impact of psoriasis on the life of the individual patient the Psoriasis Disability Index (PDI) has been developed. This takes the form of a questionnaire and covers all aspects of the patient's work, personal relationships, domestic situation, and recreational activities. It can be helpful in assessing the effectiveness of treatment as perceived by the patient.

Patients understandably ask whether psoriasis can be cured and often want to know the cause. The cause is unknown and the best answer is that the tendency to develop psoriasis is part of an affected person's constitution and some factor triggers the development of the clinical lesions. Known factors include physical or emotional stress, local trauma to the skin (Koebner's phenomenon), infection (in guttate psoriasis), drugs (β blockers, lithium, and antimalarial drugs).

Treatment comprises ointments and pastes, systemic drugs, or various forms of ultraviolet light. The treatment should suit the type of psoriasis. The age and health of the patient, social and occupational factors need to be taken into consideration. The motivation of the individual patient is also important.

The preparations mentioned in the text are listed in the formulary in chapter 26. It is estimated that 80% of patients with psoriasis do not consult a doctor, as the lesions are minimal.

Preparation applied to affected area (left). Application of stockinette (right)

Bandages being applied to larger areas (left). Patient now prepared for contact treatment (right)

Treatment of psoriasis

Type of psoriasis	Treatment	Alternative treatment
Stable plaque psoriasis	Tar preparations Calcipotriol + topical steroids Tacalcitol Ultraviolet B (TL 01)	Short contact dithranol
Extensive stable plaques	As above. If not responding: Ultraviolet B (TL01) psoralen with ultraviolet A + etretinate	Methotrexate Ciclosporin A Tacrolimus
Widespread small plaque	Ultraviolet B	Tar
Guttate psoriasis	Emollients then ultraviolet B	Weak tar preparations
Facial psoriasis	1% hydrocortisone ointment	
Flexural psoriasis	Local mild to moderate strength steroids + antifungal	
Pustular psoriasis of hands and feet	Moderate to potent strength topical steroids	Acitretin
Acute erythrodermic, unstable, or general pustular psoriasis	Inpatient treatment Short term local steroids for acutely inflamed lesions	Methotrexate Acitretin Ciclosporin or other immunosuppressants

Local treatment

Local treatments entail the use of ointments and pastes, usually containing tar in various forms. It is much easier to apply them in hospital than at home if patients can make the time for hospital visits. Inpatient treatment can be more intensive and closely regulated; it also has the advantage of taking the patient completely away from the stresses of the everyday environment. In some units a "five day ward" enables patients to return home at weekends, which is particularly important for parents with young children.

Coal tar preparations are safe and effective for the stable plaque-type psoriasis but will irritate acute, inflamed areas. However, tar may not be strong enough for thicker hyperkeratotic lesions. Salicylic acid, which helps dissolve keratin, can be used in conjunction with tar for thick plaques. Refined coal tar extracts can be used for less severe areas of psoriasis.

Ichthammol, prepared from shale rather than coal tar, is less irritating and has a soothing effect on inflamed skin. It is therefore useful for "unstable" or inflamed psoriasis, when tar would not be tolerated.

Dithranol, obtained originally from the Goa tree in south India, is now made synthetically. It can easily irritate or burn the skin, so it has to be used carefully and should be kept from contact with normal skin as far as possible. For hospital treatment pastes are used and the lesions surrounded by petroleum jelly to protect the normal skin. Dithranol creams can be used at home—they are applied for 30 minutes and then washed off. A low concentration (0·1%) is used initially and gradually increased to 1% or 2% as necessary. All dithranol preparations are irritants and produce a purple-brown staining that clears in time. If used in the scalp dithranol stains red or fair hair purple.

Emollients soften dry skin and relieve itching. They are a useful adjunct to tar or dithranol.

Corticosteroid preparations produce an initial clearing of psoriasis, but there is rapid relapse when they are withdrawn and tachyphylaxis (increasing amounts of the drug having a diminishing effect) occurs. Strong topical steroids should be avoided. Only weak preparations should be used on the face but moderately potent steroids can be used elsewhere: (a) if there are only a few small lesions of psoriasis; (b) if there is persistent chronic psoriasis of the palms, soles, and scalp (in conjunction with tar paste, which is applied on top of the steroid at night); and (c) in the treatment of psoriasis of the ears, flexures, and genital areas. In flexural psoriasis secondary infection can occur and steroid preparations combined with antibiotics and antifungal drugs should be used, such as Terra-Cortril with nystatin and Trimovate.

Systemic corticosteroids should not be used, except in life threatening erythroderma, because of the inevitable "rebound" that occurs when the dose is reduced. The management of psoriasis in patients taking steroids for an unrelated condition may require inpatient or regular outpatient attendances to clear the skin lesions.

Calcipotriol and *tacalcitol*, vitamin D analogues, are calmodulin inhibitors used topically for mild or moderate plaque psoriasis. They are non-staining creams that are easy to use but can cause irritation. Sometimes a plateau effect is seen with the treatment becoming less effective after an initial response. If so, other agents, such as tar preparations, have to be used as well to clear the lesions completely. It is important not to exceed the maximum recommended dose so as to prevent changes in calcium metabolism.

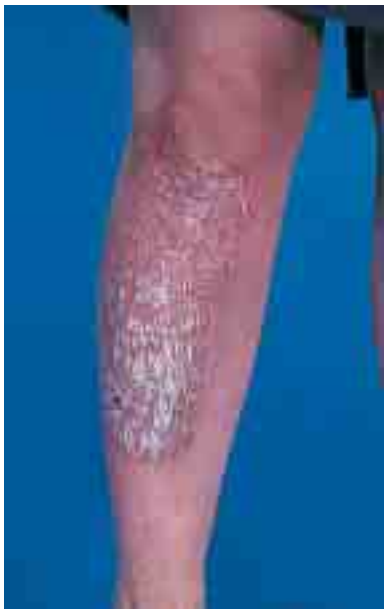
Short contact dithranol

Indications
- Stable plaque psoriasis on the trunk and limbs

Suitable preparations
- Those available are in a range of concentrations such as Dithocream (0·1%, 0·25%, 0·5%, 1·0%, 2·0%) or Anthranol (0·4%, 1·0%, 2·0%)

Method
- Start with the lowest concentration and increase strength every five to six days if there are no problems
- Apply cream to affected areas and then wash it off completely 20–30 minutes later
- Apply a bland emollient immediately after treatment

Cautions
- Do not apply to inflamed plaques, flexures, or the face
- Avoid contact of the dithranol with clothing and rinse the bath well after use to avoid staining
- Never leave the cream on for longer than 30 minutes (60 minutes is not twice as effective)



Psoriasis suitable for short contact dithranol treatment

Ultraviolet treatment (phototherapy)

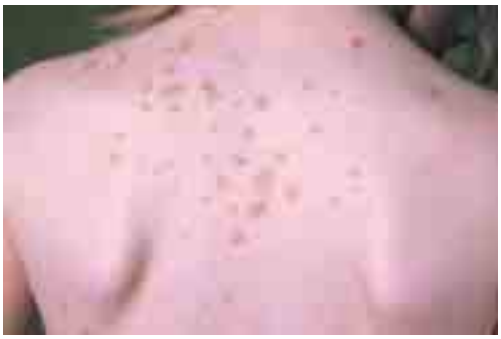
Ultraviolet B is short wavelength ultraviolet light and is used for widespread thin lesions or guttate psoriasis. The dose has to be accurately controlled to give enough radiation to clear the skin without burning. Recently, "narrow waveband" ultraviolet B treatment has been developed, which increases the therapeutic effect and diminishes burning. It can be used instead of psoralen with ultraviolet A in many cases.

Ultraviolet A is long wavelength ultraviolet light, which activates psoralens in the skin. This results in diminished DNA synthesis and hence reduced epidermal turnover. The combination of psoralen with ultraviolet A is known as PUVA therapy: a dose of 8-methoxypsoralen (8MOP), 0·6–0·8 mg/kg body weight, is taken one to two hours before treatment. 5-Methoxypsoralen is also used, particularly in patients develop itching or nausea with 8MOP.

Other long term cumulative side effects of ultraviolet treatment include premature ageing of the skin, lentigenes, and eventually cutaneous malignancies. For this reason the total cumulative dose is kept below 1000 Joules.

After medical assessment treatment is given two or three times a week, with gradually increasing doses of ultraviolet A. Once the psoriasis has cleared maintenance treatments can be continued once every two or three weeks. Protective goggles are worn during treatment with ultraviolet A and dark glasses for 24 hours after each treatment. The glasses are tested for their effectiveness in screening ultraviolet A light.

A variable degree of erythema and itching may occur after treatment. Longer term side effects include a slight risk of epitheliomas developing, premature ageing of the skin, and cataract formation (which can be prevented by wearing ultraviolet A filtering goggles during and after treatment). The total cumulative dosage is carefully monitored and kept as low as possible to reduce the risk of side effects.



Guttate psoriasis suitable for ultraviolet B treatment

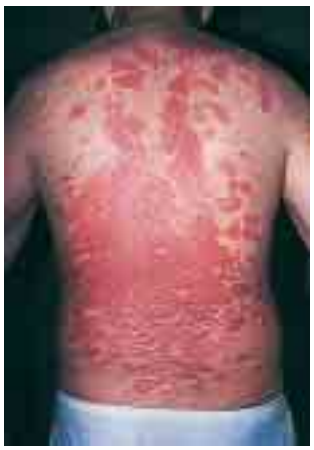
Ultraviolet B cabinet

Systemic treatment

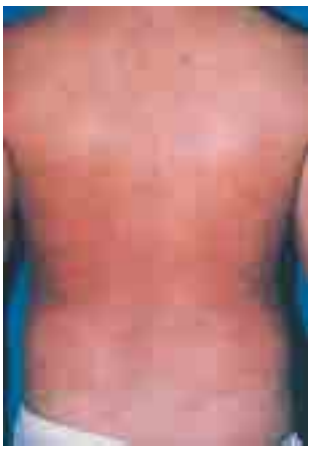
Extensive and inflamed psoriasis that is resistant to local treatment may require systemic treatment. A number of antimetabolite drugs (such as azathioprine and hydroxyurea) and immunosuppressive drugs (such as ciclosporin A) are effective, but the most widely used are methotrexate and acitretin.

Methotrexate inhibits folic acid synthesis during the S phase of mitosis and diminishes epidermal turnover in the lesions of psoriasis. Because it is hepatotoxic liver function has to be assessed initially and at regular intervals during treatment. The dosage must be monitored, and when a total of 1·5 g is reached a liver biopsy is indicated to exclude significant liver damage.

Although it is rare, bone marrow suppression can occur insidiously and rapidly in some patients. In order to detect this an initial test dose is followed by a full blood count. If this gives normal results a weekly dose of 7·5–15 mg is used. As it is excreted in the urine, the dose must be reduced if renal function is impaired. Aspirin and sulphonamides diminish plasma binding.



Before phototherapy

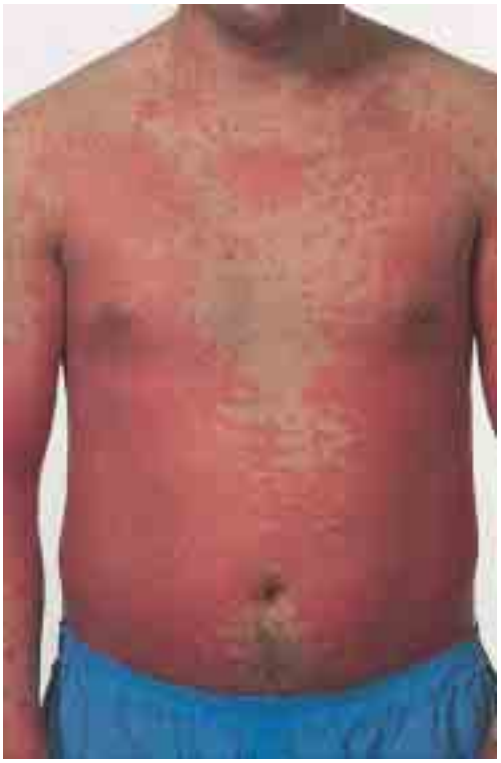


After phototherapy

Methotrexate may interact with barbiturates, para-aminobenzoic acid, phenytoin, probenecid, phenylbutazone, oral contraceptives, and colchicine.

Acitretin is a vitamin A derivative that can be prescribed only in hospital in the United Kingdom. It is useful in pustular psoriasis and has some effect on other types of psoriasis. However, the effect is increased when combined with PUVA. Minor side effects include drying of the mucous membranes, crusting in the nose, itching, thinning of the hair, and erythema of the palms and nail folds. These are usually not severe and settle when treatment stops. More serious side effects include hepatotoxicity and raised lipid concentrations. Liver function tests and serum lipid (cholesterol and triglyceride) concentrations have to be carefully monitored. Etretinate is teratogenic and should only be taken by women during reproductive years if effective contraception is used during treatment and for two years afterwards, as the half life is 70–100 days.

Ciclosporin A is an immunosuppressant widely used following organ transplantation. It is effective in suppressing the inflammatory types of psoriasis. Blood tests should be carried out before starting treatment, particularly serum creatinine, urea, and electrolytes, as ciclosporin A can interfere with renal function.



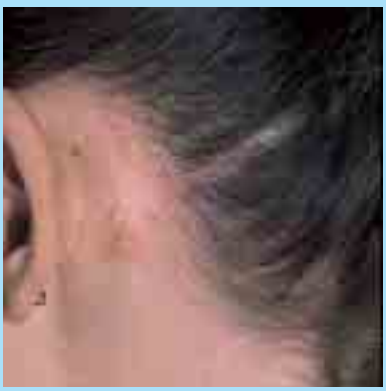
Erythematous psoriasis suitable for methotrexate treatment, having failed to respond to phototherapy

Further reading

Lowe NJ. *Managing your psoriasis*. London: Master Media, 1993
Lowe NJ. *Practical psoriasis therapy*, 2nd ed. St Louis: Mosby, 1992

Psoriasis of the scalp

This condition can be very difficult to clear, particularly if there are thick scales



Scalp psoriasis

- 3% salicylic acid in a suitable base and left on for four to six hours or overnight and then washed out with a tar shampoo
- Dithranol preparations are effective but will tint blonde or red hair purple
- Steroid preparations can be used to control itching

4 Eczema and dermatitis

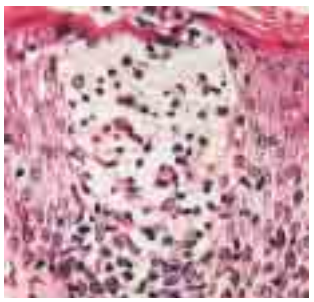
The terms eczema and dermatitis are interchangeable, covering a wide variety of conditions from the child with atopic eczema to the adult with an allergy to cement. If patients are told they have dermatitis they may assume that it is related to their employment with the implication that they may be eligible for compensation. It is not unusual for industrial workers to ask "Is it dermatitis, doctor?", meaning "is it due to my job?"

Eczema

Clinical appearance

Eczema is an inflammatory condition of the skin characterised by groups of vesicular lesions with a variable degree of exudate and scaling. In some cases dryness and scaling predominate, with little inflammation. In more acute cases there may be considerable inflammation and vesicle formation, in keeping with the Greek for "to boil out", from which the word eczema is derived. Sometimes the main feature may be blisters that become very large.

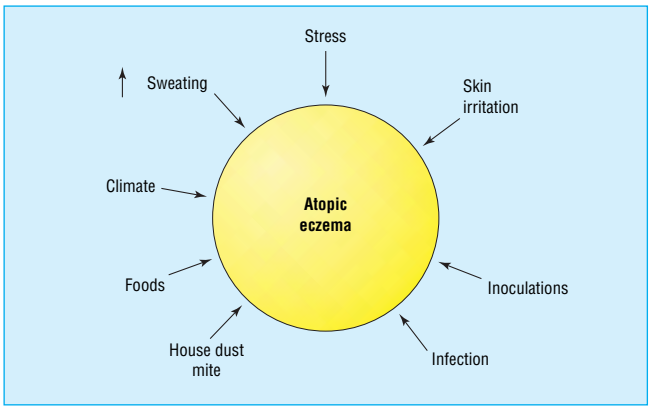
Eczema commonly itches and the clinical appearance may be modified by scratching, which with time may produce lichenification (thickening of the skin with increased skin markings). Also as a result of scratching the skin surface may be broken and have excoriations, exudate, and secondary infection.



Pathology of eczema

Pathology

The characteristic change is oedema between the cells of the epidermis, known as *spongiosus*, leading to formation of vesicles. The whole epidermis becomes thickened with an increased keratin layer. A variable degree of vasodilatation in the dermis and an inflammatory infiltrate may be present.



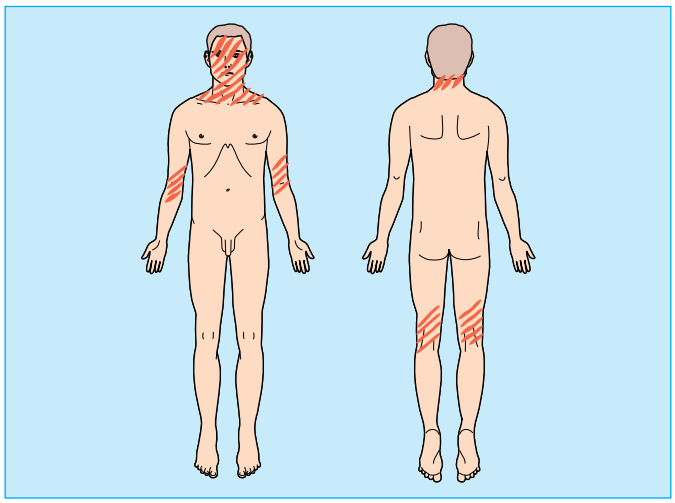
Factors leading to development of atopic eczema

Types of eczema

The many causes of eczema are not consistently related to the distribution and clinical appearance. In general there are either external factors acting on the skin producing inflammatory changes or it is an endogenous condition. It is important to remember there can be more than one cause—for example, in atopic eczema or varicose eczema on the ankle an allergic reaction may develop to the treatments used.

Atopic eczema affects mainly the flexor surfaces of the elbows and knees as well as the face and neck. To a variable degree it can affect the trunk as well.

The typical patient with atopic eczema is a fretful, scratching child with eczema that varies in severity, often from one hour to the next. In the older child or adult, eczema is more chronic and widespread and its occurrence is often related to stress. Atopic eczema is common, affecting 3% of all infants, and runs a chronic course with variable remissions. It normally clears during childhood but may continue into adolescence and adult life as a chronic disease. It is often associated with asthma and rhinitis. Sufferers from atopic eczema often have a family history of the condition.

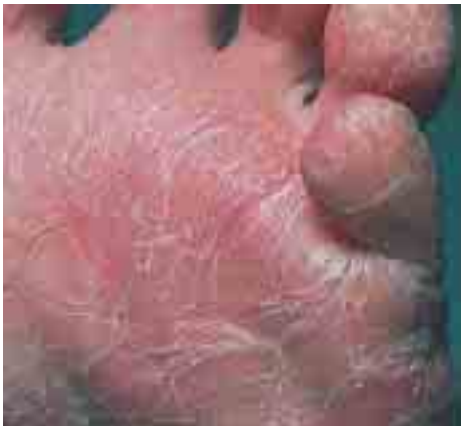


Atopic eczema—distribution

Variants of atopic eczema are pityriasis alba—white patches on the face of children with a fair complexion—and chronic juvenile plantar dermatosis—dry cracked skin of the forefoot in children. This does not affect the interdigital spaces and is not due to a fungal infection.

Eczema herpeticum. Children with atopic eczema are particularly prone to herpes virus infection, which may be life

Atopic eczema



Plantar dermatosis

Nummular eczema

threatening. Close contact with adults with "cold sores" should therefore be avoided.

Nummular eczema appears as coin shaped lesions on legs and trunk.

Stasis eczema occurs around the ankles, where there is impaired venous return.

Paget's disease of the breast. Whereas bilateral eczema of the nipples and areolae occur in women, any unilateral, persistent, areas of dermatitis in this region may be caused by Paget's disease, in which there is underlying carcinoma of the ducts. In such cases a biopsy is essential.

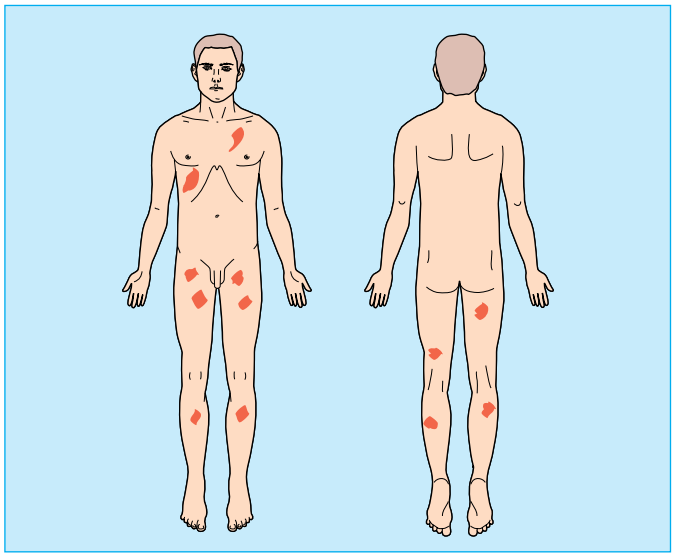
Lichen simplex is a localised area of lichenification produced by rubbing.

Neurodermatitis is a term often used synonymously with lichen simplex. It is also used to describe generalised dryness and itching of the skin, usually in those with atopic eczema.

Asteatotic eczema occurs in older people with a dry, "crazypaving" pattern, particularly on the legs.

Pompholyx is itching vesicles on the fingers, with lesions on the palms and soles in some patients.

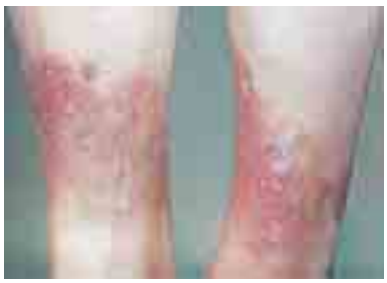
Infection can modify the presentation of any type of eczema or contact dermatitis.



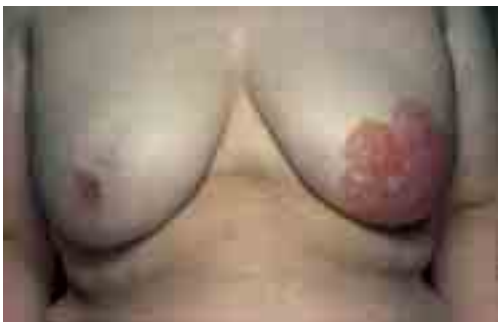
Nummular eczema—distribution

Classification of eczema

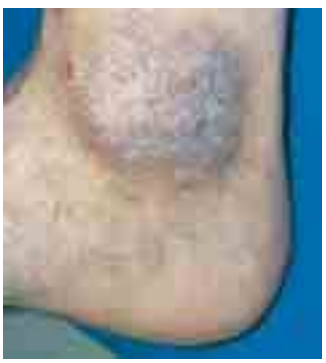
Endogenous (constitutional) eczema	Exogenous (contact) eczema	Secondary changes
Atopic	Irritant	Lichen simplex
Nummular or discoid	Allergic	Neurodermatitis
Pompholyx	Photodermatitis	Asteatosis
Stasis		Pompholyx
Seborrhoeic (discussed later)		Infection



Stasis eczema



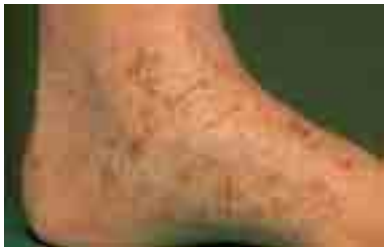
Paget's disease of the nipple



Lichen simplex


Asteatosis


Pompholyx

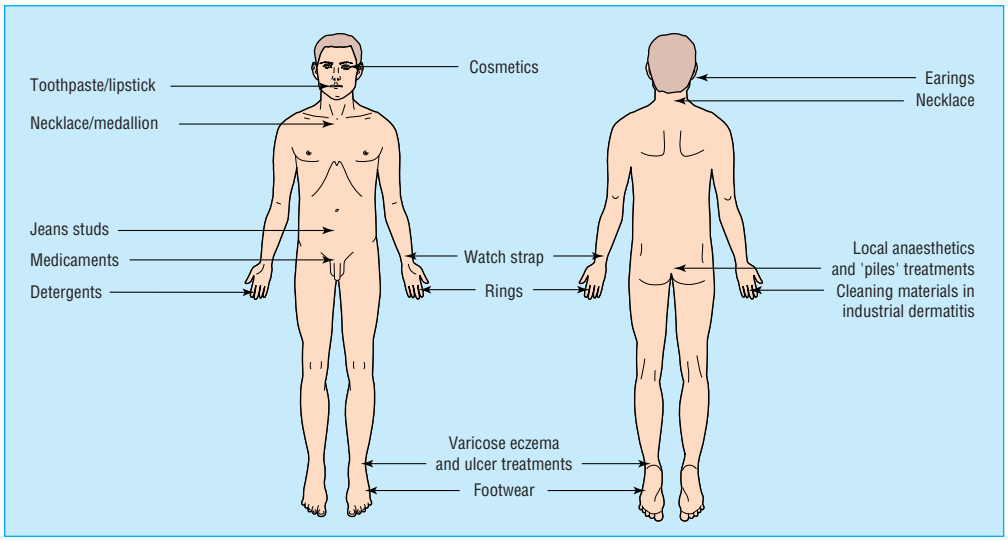

Infected eczema

Contact dermatitis

The skin normally performs its function as a barrier very effectively. If this is overcome by substances penetrating the epidermis an inflammatory response may occur leading to epidermal damage. These changes may be due to either (a) an allergic response to a specific substance acting as a sensitiser or (b) a simple irritant effect. An understanding of the difference between these reactions is helpful in the clinical assessment of contact dermatitis.

Common sources of allergic contact dermatitis

- Jewellery, clothing, wristwatch, scissors, cooking utensils
- Cement, leather
- Hair dyes, tights, shoes
- Rubber gloves and boots
- Creams, ointments, cosmetics
- Nickel—and cobalt occasionally
- Chromate
- Paraphenylenediamine—used in hair dyes
- Rubber preservative chemicals
- Preservatives (parabenz, quarternium), balsam of Peru, fragrances, lanolin, neomycin, benzocaine in medicated ointments



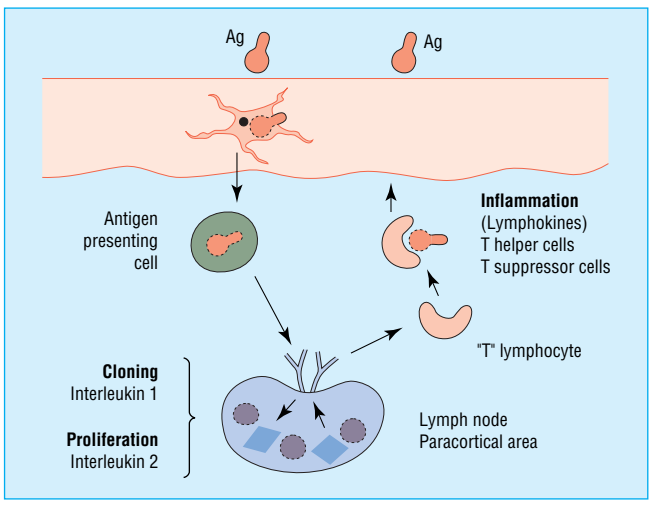
Contact dermatitis—common sources

Allergic contact dermatitis

The characteristics of allergic dermatitis are:

- Previous exposure to the substance concerned.
- 48–96 hours between contact and the development of changes in the skin.
- Activation of previously sensitised sites by contact with the same allergen elsewhere on the body.
- Persistence of the allergy for many years.

The explanation of the sequence of events in a previously sensitised individual is as follows: The antigen penetrates the epidermis and is picked up by a Langerhans cell sensitised to it. It is then transported to the regional lymph node where the paracortical region produces a clone of T cells specifically programmed to react to that antigen. The sensitised T cells accumulate at the site of the antigen and react with it to produce an inflammatory response. This takes 48 hours and is amplified by interleukins that provide a feedback stimulus to the production of further sensitised T cells.

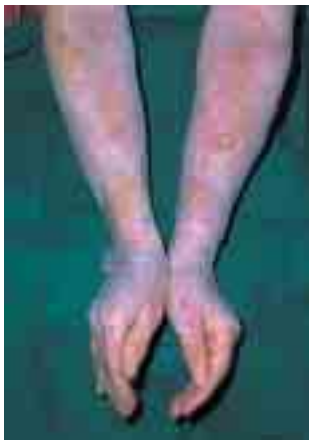


Immunological response leading to contact dermatitis

Allergic contact dermatitis can be illustrated by the example of an individual with an allergy to nickel who has previously reacted to a wrist watch. Working with metal objects that contain nickel leads to dermatitis on the hands and also a flare up at the site of previous contact with the watch. The skin clears on holiday but the dermatitis recurs two days after the person returns to work.

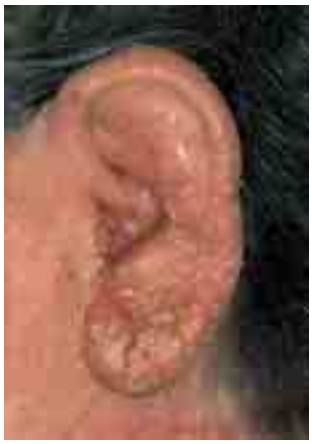
Sensitisers in leg ulcer treatments

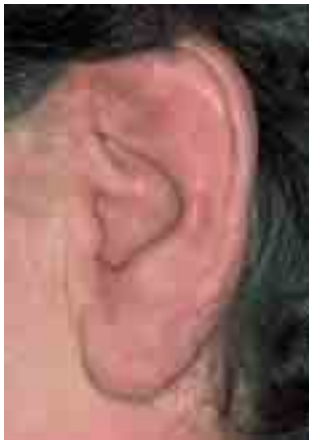
- Neomycin
- Lanolin (wool alcohol)
- Formaldehyde
- Tars
- Chinaform (the "C" of many proprietary steroids)



Allergic response to sulfapyridine

Allergic response to Solarcaine and sun





Allergic response to topical neomycin (left). After stopping ointment (right)

Irritant contact dermatitis

This has a much less defined clinical course and is caused by a wide variety of substances with no predictable time interval between contact and the appearance of the rash. Dermatitis occurs soon after exposure and the severity varies with the quantity, concentration, and length of exposure to the substance concerned. Previous contact is not required, unlike allergic dermatitis where previous sensitisation is necessary.

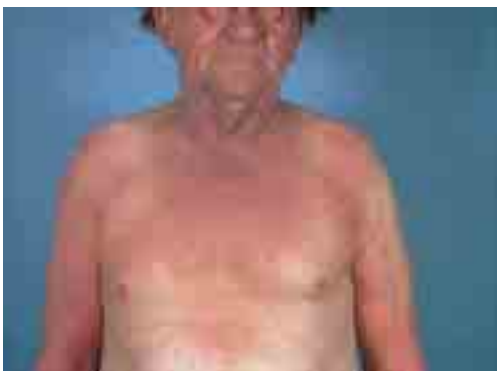
Allergic response to dithranol

Photodermatitis

Photodermatitis, caused by the interaction of light and chemical absorbed by the skin, occurs in areas exposed to light. It may be due to (a) drugs taken internally, such as sulphonamides, phenothiazines, and dimethylchlortetracycline, or (b) substances in contact with the skin, such as topical antihistamines, local anaesthetics, cosmetics, and antibacterials.

Morphology

The clinical appearance of both allergic and irritant contact dermatitis may be similar, but there are specific changes that help in differentiating them. An acute allergic reaction tends to produce erythema, oedema, and vesicles. The more chronic lesions are often lichenified. Irritant dermatitis may present as slight scaling and itching or extensive epidermal damage resembling a superficial burn, as the child in the illustration shows.



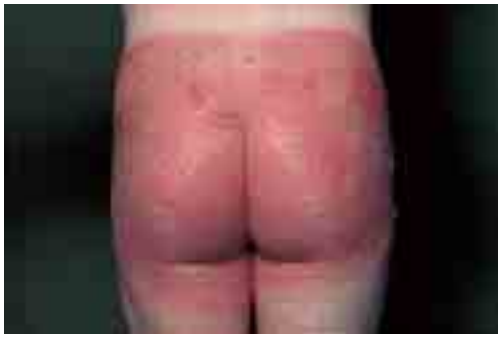
Photodermatitis

Pathology

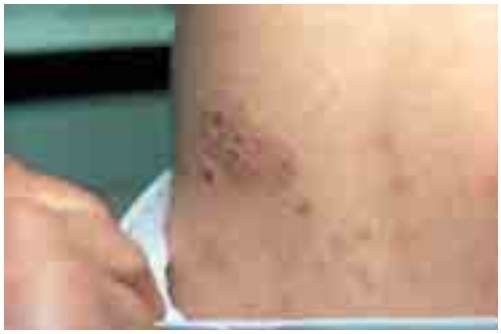
The reaction to specific allergens leads to a typical eczematous reaction with oedema separating the epidermal cells and blister formation. In irritant dermatitis there may also be eczematous changes but also non-specific inflammation, thickening of keratin, and pyknotic, dead epidermal cells.

The distribution of the skin changes is often helpful. For example, an itchy rash on the waist may be due to an allergy to rubber in the waistband of underclothing or a metal fastener. Gloves or the rubber lining of goggles can cause a persisting dermatitis. An irritant substance often produces a more diffuse eruption, as shown by the patient who developed itching and redness from dithranol.

An allergy to medications used for treating leg ulcers is a common cause of persisting dermatitis on the leg.



Acute irritant dermatitis



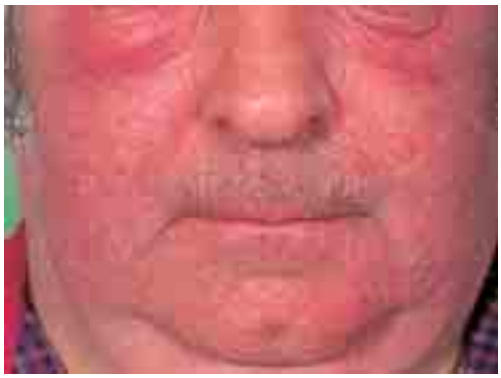
Allergic response to elastic in underpants

Allergic reaction to cosmetics

Substances commonly causing allergic occupational dermatitis

- Chromate—in cement and leather
- Biocides, for example, formaldehyde and isothiazolinones, used in cutting oils in engineering
- Epoxy resins (uncured monomers)
- Rubber chemicals
- Hair dressing chemicals—particularly dyes and setting lotions
- Plant allergens

Allergic reaction to epoxy resin



Acutely infected eczema

Patch testing

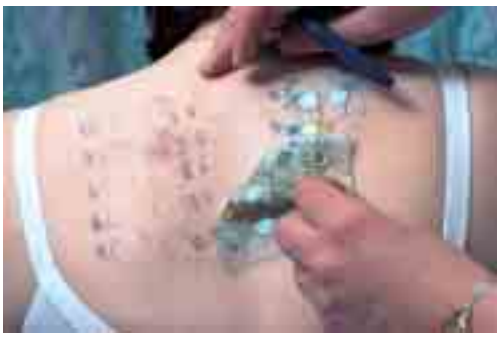
Patch testing is used to determine the substances causing contact dermatitis. The concentration used is critical. If it is too low there may be no reaction, giving a false negative result, and if it is too high it may produce an irritant reaction, which is interpreted as showing an allergy (false positive). Another possible danger is the induction of an allergy by the test substance. The optimum concentration and best vehicle have been found for most common allergens, which are the basis of the "battery" of tests used in most dermatology units.

The test patches are left in place for 48 hours then removed, the sites marked, and any positive reactions noted. A further examination is carried out at 96 hours to detect any further reactions.

It is most important not to put a possible causative substance on the skin in a random manner without proper dilution and without control patches. The results will be meaningless and irritant reactions, which are unpleasant for the patient, may occur.

Test patches in place

Positive reactions marked



Patches being removed after 48 hours

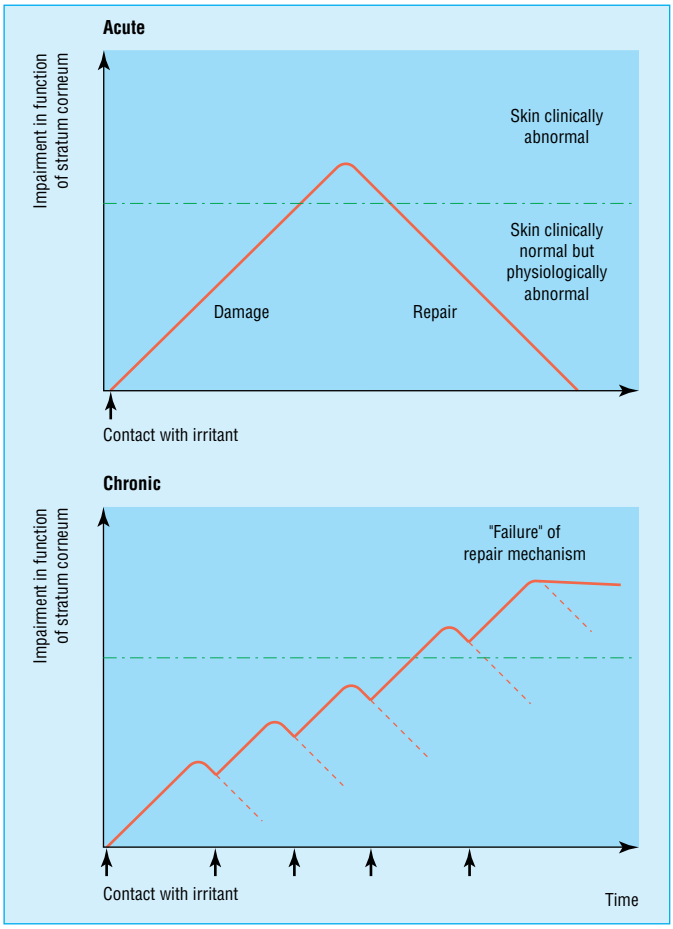
Occupational dermatitis

Dermatitis, which is simply inflammation of the skin, can arise as a result of:

- Inherited tendency to eczema (atopy)
- Contact dermatitis, which may be either irritant or allergic
- Infection.

In the workplace, all three factors may contribute to dermatitis. For example, a student nurse or trainee hairdresser is exposed to water, detergents, and other factors that will exacerbate any pre-existing eczema. In addition, there may be specific allergies and, as a result of the broken skin, secondary infection can occur making the situation even worse. The following points are helpful in determining the role of occupational causes.

- If the dermatitis first occurred during employment or with a change of employment and had not been present before, then occupational factors are more likely.
- If the condition generally clears during holidays and when away from the workplace, this suggests an occupational cause, but chronic irritant dermatitis may persist when the patient is away from the workplace.
- If there is exposure to substances that are known to induce dermatitis and protective measures are inadequate at the workplace then an occupational cause is likely.
- If secondary infection is present, this can keep a dermatitis active even when away from the workplace and sometimes allergen exposure continues at home; for example, an allergy



Progression of acute and chronic dermatitis

to rubber gloves at work will also occur when rubber gloves are used for domestic work at home.

Whatever the cause of the dermatitis, the end result may seem the same clinically, because the inflammation and blisters of atopic eczema may be indistinguishable from an allergic reaction to rubber gloves. Generally, contact dermatitis is more common on the dorsal surface of the hands whereas atopic eczema occurs on the palms and sides of the fingers.

Irritant contact dermatitis can occur acutely as mentioned above and there is usually a definite history of exposure to irritating chemicals.

Chronic irritant dermatitis can be harder to assess as it develops insidiously in many cases. Often it starts with episodes of transient inflammation that clear up, but with each successive episode the damage becomes worse with an escalation of inflammatory changes that eventually become chronic and fixed.

Once chronic damage has occurred the skin is vulnerable to any further irritation, so the condition may flare up in the future even after removal of the causative factors. Individuals with atopic eczema are particularly liable to develop chronic irritant dermatitis and secondary infection is an additional factor.

Allergic contact dermatitis occurs as an allergic reaction to specific substances. As this involves a cell mediated response the inflammatory reaction occurs about two days after exposure and once the allergy is present further exposure will inevitably produce a reaction. Some substances are much more likely to produce an allergy, such as epoxy resin monomer, than others, such as cement, which characteristically requires exposure over many years before an allergy develops. In addition to the capacity of the substance to produce an allergic reaction, individuals also vary considerably in the capacity to develop allergies.

Immediate type sensitivity is sometimes seen as a reaction to food protein and sensitivity to latex gloves. This can produce a very severe reaction, particularly in atopic individuals.

Treatment of occupational dermatitis

The exact cause of the dermatitis should be identified as far as possible. It is important to ascertain exactly what an individual's job entails; for example, a worker in a plastics factory had severe hand dermatitis but the only positive result on patch testing was to nickel. On visiting the factory it became clear that the cause was a nickel plated handle that he used several thousand times a day and not the plastic components that the machine was making. It is also important to assess the working environment because exposure to damp and irritants (for example, on an oil rig or in a coal mine) can irritate the skin.

If occupational factors are suspected, then a full assessment and investigation in a dermatology department is important as the patient's future working life may be at stake

The itching skin (pruritus)

It is sometimes very difficult to help a patient with a persistently itching skin, particularly if there is no apparent cause. Pruritus is a general term for itching skin, whatever the reason.

Systemic causes

- Endocrine diseases—diabetes, myxoedema, hyperthyroidism
- Metabolic diseases—hepatic failure, chronic renal failure
- Haematological—polycythaemia, iron deficiency anaemia
- Malignancy—lymphoma, reticulosis, carcinomatosis
- Psychological—anxiety, parasitophobia
- Tropical infection—filariasis, hookworm
- Drugs—alkaloids

Itching with skin manifestations

Eczema is associated with itching due to the accumulation of fluid between the epidermal cells that are thought to produce stretching of the nerve fibres. As a result of persistent scratching there is often lichenification which conceals the original underlying areas with eczema. Exposure to irritants and persistent allergic reactions can produce intense itching and should always be considered.

"Allergic reactions" due to external agents often cause intense itching. Systemic allergic reactions such as a fixed drug eruption, erythema multiforme, and vasculitis are less likely to cause pruritus.

Psoriasis, which characteristically has hyperkeratotic plaques, usually does not itch but sometimes there can be considerable itching. Occasionally this is due to secondary infection of breaks in the skin surface.

Lichen planus presents with groups of flat-topped papules which often cause an intense itch. Blistering disorders of the skin may itch.

In *herpes simplex* there is usually burning and itching in the early stages.

Investigations

- Skin scrapings for mycology
- Patch testing for allergies
- Full blood count, erythrocyte sedimentation rate, liver and renal function tests
- Urine analysis
- Stools for blood and parasites

In *herpes zoster* there may be a variable degree of itching, but this is overshadowed by the pain and discomfort of the fully developed lesions.

By contrast, *bullous impetigo* causes few symptoms, although there may be extensive blisters. Itching is usually not present.

Dermatitis herpetiformis is characterised by intense persistent and severe itching that patients often describe as being unendurable. Usual measures such as topical steroids and antihistamines have little if any effect.

By contrast, the blisters of *pemphigoid* do not itch although the earlier inflamed lesions can be irritating.

Parasites. Fleas and mites cause pruritic papules in groups. The patient may not realise that they may have been acquired after a walk in the country or encountering a dog or cat. *Nodular prurigo* may develop after insect bites and is characterised by persistent itching, lichenified papules, and nodules over the trunk and limbs. The patient attacks them vigorously and promotes a persisting "itch–scratch–itch" cycle which is very difficult to break.

Parasitophobia is characterised by the patient reporting the presence of small insects burrowing into the skin which persists despite all forms of treatment. The patient will produce small flakes of skin, fibres of clothing, and pieces of dust, usually in carefully folded pieces of paper, for examination. These should always be examined and the patient gently informed that no insect could be found but this will not be believed. Treatment is therefore very difficult and sometimes recourse has to be had to psychotropic drugs (see page 106).

Infestations with lice cause irritation and a scabies mite can cause widespread persistent pruritus, even though only a dozen or so active scabies burrows are present. It is always acquired by close human contact and the diagnosis may be missed unless an adequate history of personal contacts and a thorough clinical examination is carried out. However, a speculative diagnosis of scabies should be avoided.

Itching with no skin lesions

If no dermatological lesions are present generalised pruritus or itchy skin may indicate an underlying internal cause. In elderly patients, however, the skin may itch simply because it is dry. Hodgkin's disease may present with pruritus as a sign of the internal malignancy long before any other manifestations. A 35 year old ambulance driver attended the dermatology clinic with intense itching but a normal skin and no history of skin disease. His general health was good and both physical examination and all blood tests were normal. However, a chest *x* ray examination showed a mediastinal shadow that was found to be due to Hodgkin's lymphoma. Fortunately this was easily treated. Other forms of carcinoma rarely cause pruritus.

Metabolic and endocrine disease

Biliary obstruction and *chronic renal disease* cause intense pruritus. *Thyroid disease* can be associated with an itching skin. In hyperthyroidism the skin seems normal but in hypothyroidism there is dryness of the skin causing pruritus.

Blood diseases. Polycythaemia and iron deficiency are sometimes associated with itching skin.

Treatment

Treatment of the cause must be carried out when possible. Calamine lotion cools the skin with 0·5% menthol or 1% phenol in aqueous cream. Camphor-containing preparations and crotamiton (Eurax) are also helpful. Topical steroid ointments and occlusive dressings may help to prevent scratching and may help to break the itch–scratch–itch cycle. Emollients should be used for dry skin.

Topical local anaesthetics may give relief but intolerance develops and they can cause allergic reactions. Sedative antihistamines at night may be helpful. In liver failure cholestyramine powder may help to relieve the intense pruritus, as this is thought to be due to bile salts in the skin.

Antihistamines can be helpful both for their antipruritic effect and because many are sedative and enable the itching patient to sleep.

Pruritus ani is a common troublesome condition and the following points may be helpful:

- Advise gentle cleaning once daily and patients should be advised to avoid excessive washing.
- Avoid harsh toilet paper, especially if coloured, because cheap dyes irritate and cause allergies. Olive oil and cotton wool can be used instead.
- Weak topical steroids will help to reduce inflammation, with zinc cream or ointment as a protective layer on top.
- Anal leakage from an incompetent sphincter, skin tags, or haemorrhoids may require surgical treatment.
- There may be an anxiety or depression and prutitus ani itself can lead to irritability and depression.

Pruritus vulvae is a persistent irritation of the vulva which can be most distressing and is most common in postmenopausal women. It is important to eliminate any factors that may be preventing resolution. These include:

- Secondary infection with pyogenic bacteria or yeasts
- Eczema or contact dermatitis
- Lichen sclerosus atrophicus.

The adjacent vaginal mucosa should be examined to exclude an intraepithelial neoplasm or lichen planus. Treatment includes suitable antiseptic preparations such as 2% eosin, regular but not excessive washing, emollients, and topical steroids, bearing in mind the possibility of infection.

Further reading

Adams RM. *Occupational skin disease*, 2nd ed. Philadelphia: Saunders, 1990

Arndt KA. *Manual of dermatological therapeutics*, 5th ed. New York: Little, 1995

Atherton DJ. *Eczema in childhood. The facts.* Oxford: Oxford University Press, 1994

Cronin E. *Contact dermatitis.* Edinburgh: Churchill Livingstone, 1980

Fisher AA. *Contact dermatitis*, 3rd ed. Baltimore: Williams and Wilkins, 1986

Foussereau J, Benezra JE, Maibach H. *Occupational contact dermatitis.* Copenhagen: Munksgaard, 1982

Schwanitz HJ. *Atopic palmoplantar eczema.* Berlin: Springer-Verlag, 1988

5 Treatment of eczema and inflammatory dermatoses

Treat the patient, not just the rash. Many patients accept their skin condition with equanimity but others suffer much distress, especially if the face and hands are affected. Acceptance by the doctor of the individual and his or her attitudes to the disease goes a long way to helping the patient live with the condition.

The common inflammatory skin diseases can nearly always be improved or cleared, but it is wise not to promise a permanent cure.

Be realistic about the treatment people can apply in their own homes. It is easy to unthinkingly give patients with a widespread rash a large amount of ointment to apply twice daily, which is hardly used because: (a) they have a busy job or young children and simply do not have time to apply ointment to the whole skin; (b) they have arthritic or other limitations of movement and can reach only a small part of the body; (c) the tar or other ointment is smelly or discolours their clothes. Most of us have been guilty of forgetting these factors at one time or another.

Dry skin tends to be itchy, so advise minimal use of soap. Emollients are used to soften the skin, and the simpler the better. Emulsifying ointment BP is cheap and effective but rather thick. By mixing two tablespoons in a kitchen blender with a pint of water, the result is a creamy mixture that can easily be used in the bath. A useful preparation is equal parts of white soft paraffin and liquid paraffin. Various proprietary bath oils are available and can be applied directly to wet skin. There are many proprietary emollients.

Wet weeping lesions should generally be treated with creams rather than ointments (which remain on the surface).

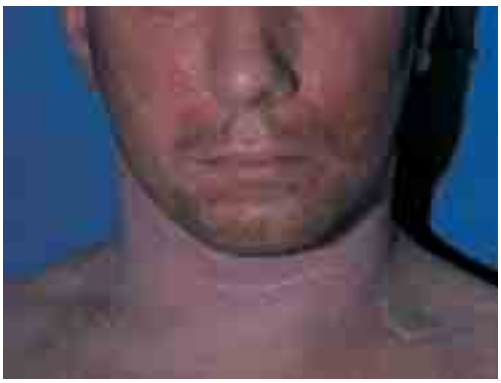
Steroid ointments are effective in relieving inflammation and itching but are not always used effectively. Advise patients to use a strong steroid (such as betamethasone or fluocinolone acetonide) frequently for a few days to bring the condition under control; then change to a weaker steroid (dilute betamethasone, fluocinolone, clobetasone, hydrocortisone) less frequently. Strong steroids should not be continued for long periods, and, as a rule, do not prescribe any steroid stronger than hydrocortisone for the face. Strong steroids can cause atrophy of the skin if used for long periods, particularly when applied under occlusive dressings. On the face they may lead to florid telangiectasia and acne-like pustules. Avoid using steroids on ulcerated areas. Prolonged use of topical steroids may mask an underlying bacterial or fungal infection.

Immunosuppressants are a valuable adjunct in severe cases not responding to topical treatment and antibiotics. Ciclosporin is usually given on an intermittent basis, with careful monitoring for side effects. Azathioprine is also used, provided the thiopurine methyl transferase (TPMT) level is normal.

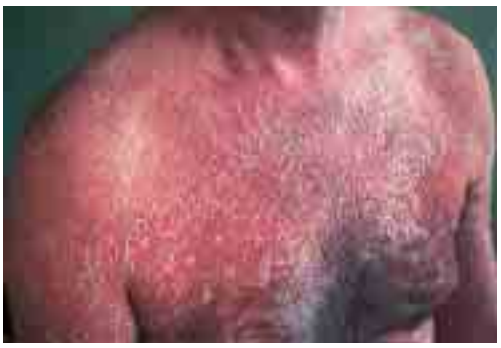
Tacrolimus is an immunosuppressant that has recently become available in two strengths as an ointment. It promises to be a successful treatment but is relatively expensive.

Treatment guidelines

- Treat the patient, not just the rash
- Complete cure may not be possible
- Be realistic about the problems of applying treatments at home
- Make sure the patient understands how to carry out the treatment
- Advise using emollients and minimal soap
- Provide detailed guidance on using steroids



Weeping eczema



Acute erythema

Specific treatment

Wet, inflamed, exuding lesions

(1) Use wet soaks with plain water, normal saline, or aluminium acetate (0·6%). Potassium permanganate (0·1%) solution should be used if there is any sign of infection.

(2) Use wet compresses rather than dry dressings ("wet wraps").
(3) Steroid *creams* should be used as outlined above. Greasy ointment bases tend to float off on the exudate.
(4) A combined steroid–antibiotic cream is often needed as infection readily develops.
(5) Systemic antibiotics may be required in severe cases. Take swabs for bacteriological examination first.

Dry, scaling, lichenified lesions

(1) Use emollients.
(2) Use steroid *ointments*, with antibiotics if infection is present.
(3) A weak coal tar preparation or ichthammol can be used on top of the ointments. This is particularly useful at night to prevent itching. 1–2% coal tar can be prescribed in an ointment. For hard, lichenified skin salicylic acid can be incorporated and the following formulation has been found useful in our department:
   (a) Coal tar solution BP 10%, salicylic acid 2%, and unguentum drench to 100%.
   (b) 1% ichthammol and 15% zinc oxide in white soft paraffin is less likely to irritate than tar and is suitable for children.
(4) In treating psoriasis start with a weaker tar preparation and progress to a stronger one.
(5) For thick, hyperkeratotic lesions, particularly in the scalp, salicylic acid is useful. It can be prescribed as 2–5% in aqueous cream, 1–2% in arachis oil, or 6% gel.

It is often easiest for the patient to apply the preparation to the scalp at night and wash it out the next morning with a tar shampoo.

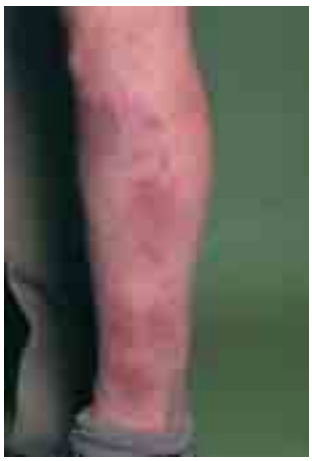
Infection

Remember that secondary infection may be a cause of persisting lesions.

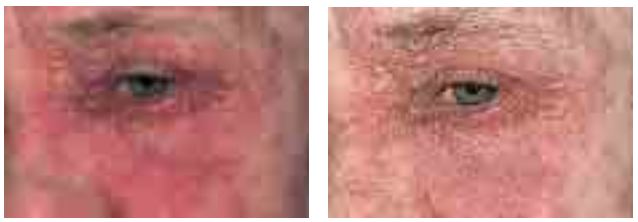
Hand dermatitis

Hand dermatitis poses a particular problem in management and it is important that protection is continued after the initial rash has healed because it takes some time for the skin to recover its barrier function. Ointments or creams should be reapplied each time the hands have been washed.

It is useful to give patients a list of simple instructions such as those shown in the box on the right.



Lichenified eczema



Infected eczema: before (left) and after (right) treatment

Hand dermatitis: hints on management

- Hand washing: use tepid water and soap without perfume or colouring or chemicals added; dry carefully, especially between fingers
- When in wet work: wear cotton gloves under rubber gloves (or plastic if you are allergic to rubber); try not to use hot water and cut down to 15 minutes at a time if possible; remove rings before wet or dry work; use running water if possible
- Wear gloves in cold weather and for dusty work
- Use only ointments prescribed for you
- Things to avoid on unprotected skin:
  Shampoo
  Peeling fruits and vegetables, especially citrus fruits
  Polishes of all kinds
  Solvents, for example, white spirit, thinners, turpentine
  Hair lotions, creams, and dyes
  Detergents and strong cleansing agents
  "Unknown" chemicals
- Use "moisturisers" or emollients which have been recommended by your doctor to counteract dryness

Further reading

Atherton DJ. *Diet and children with eczema.* London: National Eczema Society, 1986
Launer JM. *A practical guide to the management of eczema for general practitioners.* London: National Eczema Society, 1988
Mackie RM. *Eczema and dermatitis.* London: Martin Dunitz, 1983
Orton C. *Eczema relief: a complete guide to all remedies—alternative and orthodox.* London: Thorsons, 1990

6 Rashes with epidermal changes

Familiarity with the clinical features of psoriasis and eczema, which all clinicians see from time to time, provides a basis for comparison with other rather less common conditions. The characteristics that each condition has in common with psoriasis and eczema are highlighted in the relevant tables.

Lichen planus

Like psoriasis, the lesions are well defined and raised. They also occur in areas of trauma—the Koebner's phenomenon. There is no constant relation to stress.

Unlike psoriasis, there is no family history. Itching is common. The distribution is on the flexor aspects of the limbs, particularly the ankles and wrists, rather than on the extensor surfaces, as in psoriasis. It also occurs on the trunk. However, localised forms of lichen planus can occur on the shin, palm, and soles or elsewhere.

Nail involvement is less common than in psoriasis. There may be thinning and atrophy of part or all of a nail and these often take the form of a longitudinal groove, sometimes with destruction of the nail plate. The oral mucosa is commonly affected with a white, net-like appearance and sometimes ulceration.

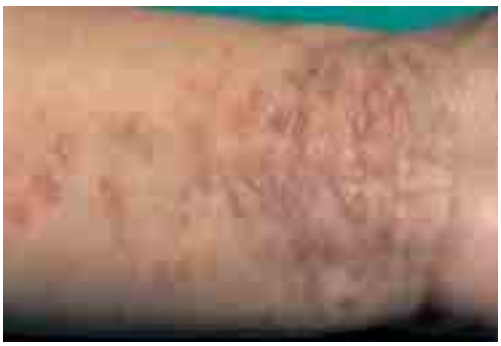
The typical flat topped lesions have a shiny hyperkeratotic lichenified surface with a violaceous colour, interrupted by milky white streaks—Wickham's striae.

Less commonly, very thick hypertrophic lesions occur and also follicular lesions. Lichen planus is one cause of localised alopecia on the scalp as a result of hair follicle destruction.

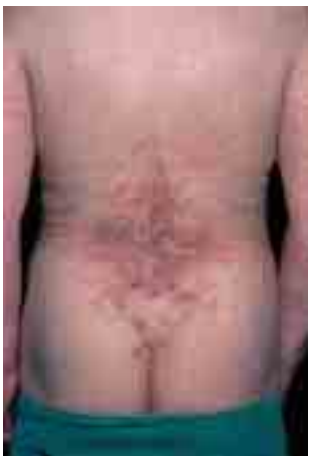
Lichen planus usually resolves over many months to leave residual brown or grey macules. In the oral mucosa and areas subject to trauma ulceration can occur.

Characteristics of lichen planus

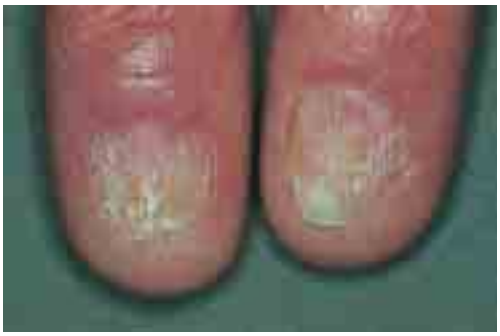
Clinical features of psoriasis	Clinical features of eczema
Possible family history	Possible family history
Sometimes related to stress	Sometimes worse with stress
Itching—rare	Usually itching
Extensor surfaces and trunk	Flexor surfaces and face
Well defined, raised lesions	Poorly demarcated lesions
Hyperkeratosis	Oedema, vesicles, lichenification
Scaling, bleeding points beneath scales	Secondary infection sometimes present
Koebner's phenomenon	
Nails affected	
Scalp affected	
Mucous membranes not affected	



Lichen planus—wrist



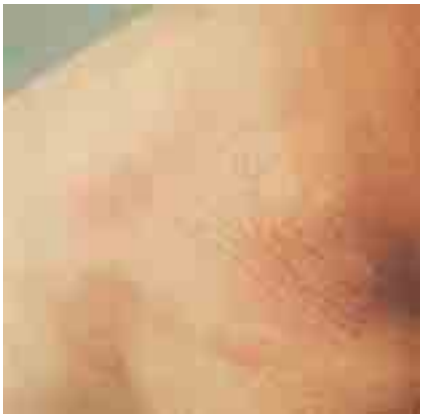
Lichen planus (left). Lichen planus—oral mucosa (right)



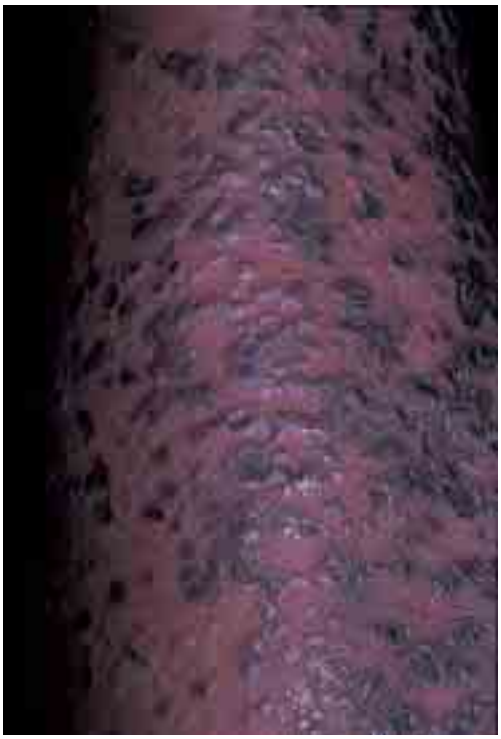
Lichen planus—nails

Treatment
There is usually a gradual response to topical steroids, but in very extensive and inflamed lesions systemic steroids may be needed. Localised hypertrophic lesions can be treated with intralesional injections.

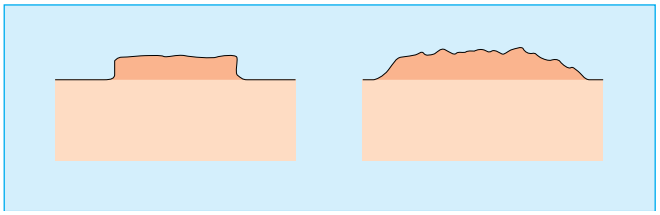
Similar rashes
Lichenified eczema
This is also itchy and may occur on the ankles and wrists. The edge of the lesion is less well defined and is irregular. The flat topped, shiny papules are absent.



Lichenified eczema



Lichen planus—skin



Guttate psoriasis—section through lesion (left); lichenified eczema—section through lesion (right)

Guttate psoriasis
Guttate psoriasis is not as itchy as lichen planus. Scaling erythematous lesions do not have the lichenified surface of lichen planus.

Pityriasis lichenoides
The lesions have a mica-like scale overlying an erythematous papule.

Drug eruptions
Rashes with many features of lichen planus can occur in patients taking:

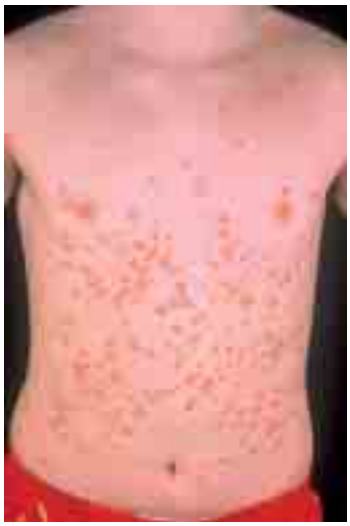
- Chloroquine
- Chlorpropamide
- Chlorothiazide

} The three "C"s

- Anti-inflammatory drugs
- Gold preparations

It also occurs in those handling colour developers.

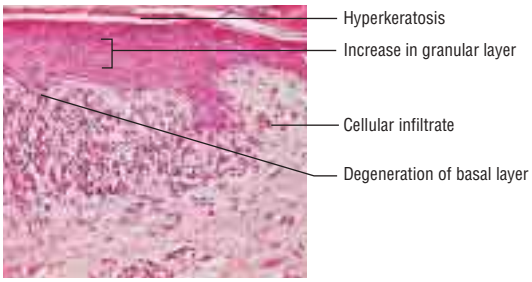
Treatment
The main symptom of itching is relieved to some extent by moderately potent steroid ointments. Very hypertrophic lesions may respond to strong steroid preparations under polythene occlusion. Careful intralesional injections may be effective in persistent lesions. In very extensive, severe lichen planus systemic steroids may be indicated.



Guttate psoriasis

Lichen planus
- Flexor surfaces
- Mucous membranes affected
- Itching common
- Violaceous colour
- Wickham's striae
- Small discrete lesions
- Lichenified

Pathology of lichen planus
As expected from the clinical appearance, there is hypertrophy and thickening of the epidermis with increased keratin. The white streaks seen clinically occur where there is pronounced thickness of the granular layer and underlying infiltrate. Degenerating basal cells may form "colloid bodies". The basal layer is being eaten away by an aggressive band of lymphocytes, the remaining papillae having a "saw toothed" appearance.



Lichen planus—pathology

Seborrhoeic dermatitis

Seborrhoeic dermatitis has nothing to do with sebum or any other kind of greasiness. There are two distinct types, adult and infantile.

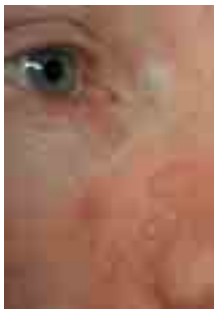
Adult seborrhoeic dermatitis
The adult type is more common in men and in those with a tendency to scaling and dandruff in the scalp. There are several commonly affected areas:

- Seborrhoeic dermatitis affects the central part of the face, scalp, ears, and eyebrows. There may be an associated blepharitis, giving some red eyes and also otitis externa.
- The lesions over the sternum sometimes start as a single "medallion" lesion. A flower-like "petaloid" pattern can occur. The back may be affected as well.
- Lesions also occur in well defined areas in the axillae and groin and beneath the breasts.

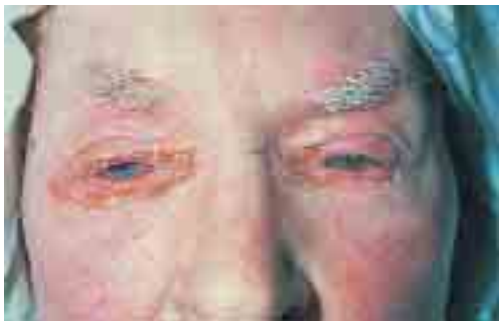
Typically the lesions are discrete and erythematous and they may develop a yellow crust. The lesions tend to develop from the hair follicles. It is a persistent condition that varies in severity.

Clinically and pathologically the condition has features of both psoriasis and eczema. There is thickening of the epidermis with some of the inflammatory changes of psoriasis and the intercellular oedema of eczema. Parakeratosis—the presence of nuclei above the basement layer—may be noticeable. Recently, increased numbers of *Pityrosporum ovale* organisms have been reported.

Treatment
Topical steroids produce a rapid improvement, but not permanent clearing. Topical preparations containing salicylic acid, sulphur, or ichthammol may help in long term control. Triazole antifungal drugs by mouth have been reported to produce clearing and can be used topically. These drugs clear yeasts and fungi from the skin, including *P. ovale*, which is further evidence for the role of this organism.

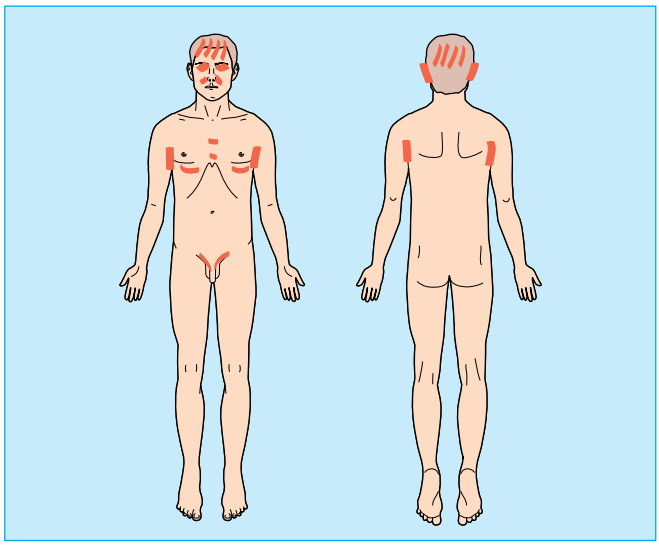

Seborrhoeic dermatitis affecting centre of face


Seborrheoic dermatitis under breasts


Seborrhoeic dermatitis

Characteristics of seborrhoeic dermatitis

Clinical features of psoriasis	Clinical features of eczema
Possible family history	Possible family history
Sometimes related to stress	Sometimes worse with stress
Itching—rare	Usually itching
Extensor surfaces and trunk	Flexor surfaces and face
Well defined, raised lesions	Poorly, demarcated lesions
Hyperkeratosis	Oedema, vesicles, lichenification
Scaling, bleeding points beneath scales	Secondary infection sometimes, present
Koebner's phenomenon	
Nails affected	
Scalp affected	
Mucous membranes not affected	


Seborrhoeic dermatitis—distribution pattern

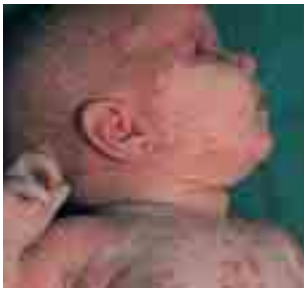
Infantile seborrhoeic dermatitis
In infants less than six months old a florid red eruption occurs with well defined lesions on the trunk and confluent areas in the flexures associated with scaling of the scalp. There is no consistent association with the adult type of seborrhoeic dermatitis. It has been suggested that infantile seborrhoeic dermatitis is a variant of atopic eczema. It is said to be more common in bottle fed infants. A high proportion of affected infants develop atopic eczema later but there are distinct differences.

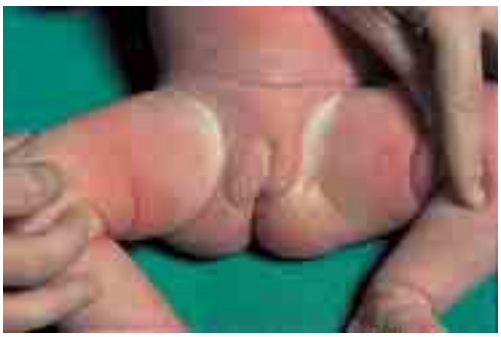
Itching is present in atopic eczema but not in seborrhoeic dermatitis.

The *clinical course* of atopic eczema is prolonged with frequent exacerbation, whereas seborrhoeic dermatitis clears in a few weeks and seldom recurs.

Treatment comprises emollients, avoiding soap, and applying hydrocortisone combined with an antibiotic plus nystatin (for example, Terra-Cortril plus nystatin cream). Hydrocortisone can be used on the scalp.

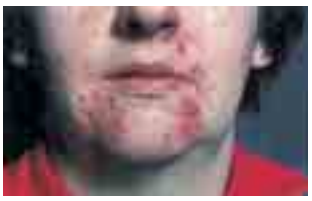
Allergy
IgE concentrations are often raised in atopic eczema and food allergy is common, but not in seborrhoeic dermatitis.





Infantile Seborrhoeic dermatitis

Perioral dermatitis
Perioral dermatitis is possibly a variant of seborrhoeic dermatitis, with some features of acne. Papules and pustules develop around the mouth and chin. It occurs mainly in women.



Perioral dermatitis

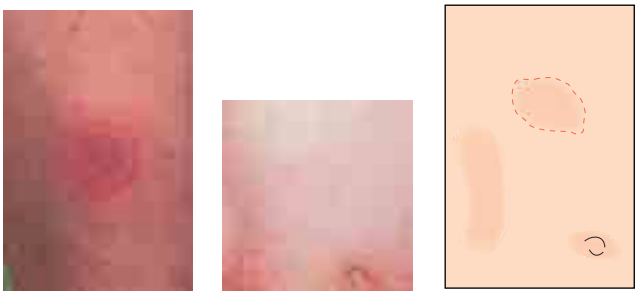
Pityriasis rosea

The word "pityriasis" is from the Greek for bran, and the fine bran-like scales on the surface are a characteristic feature. The numerous pale pink oval or round patches can be confused with psoriasis or discoid eczema. The history helps because this condition develops as an acute eruption and the patient can often point to a simple initial lesion—the herald patch.

There is commonly slight itching. Pityriasis rosea occurs mainly in the second and third decade, often during the winter months. "Clusters" of cases occur but not true epidemics. This suggests an infective basis. There may be prodromal symptoms with malaise, fever, or lymphadenopathy. Numerous causes have been suggested, from allergy to fungi; the current favourite is a virus infection.

The typical patient is an adolescent or young adult, who is often more than a little concerned about the sudden appearance of a widespread rash. The lesions are widely distributed, often following skin creases, and concentrated on the trunk with scattered lesions on the limbs. The face and scalp may be affected.

Early lesions are red with fine scales—usually 1–4 cm in diameter. The initial herald patch is larger and may be confused with a fungal infection. Subsequently the widespread eruption develops in a matter of days or, rarely, weeks. As time goes by the lesions clear to give a slight pigmentation with a collarette of scales facing towards the centre.



Pityriasis rosea—herald lesions

Similar rashes
Discoid eczema presents with itching and lesions with erythema, oedema, and crusting rather than scaling. Vesicles may be present. The rash persists unchanged.

A *drug eruption* can sometimes produce similar lesions.

In *guttate psoriasis* the lesions are more sharply defined and smaller (0·5–1·0 cm) and have waxy scales.

Characteristics of pityriasis rosea

Clinical features of psoriasis	Clinical features of eczema
Possible family history	Possible family history
Sometimes related to stress	Sometimes worse with stress
Itching—rare	Usually itching
Extensor surfaces and trunk	Flexor surfaces and face
Well defined, raised lesions	Poorly demarcated lesions
Hyperkeratosis	Oedema, vesicles, lichenification
Scaling, bleeding points beneath scales	Secondary infection sometimes present
Köebner's phenomenon	
Nails affected	
Scalp affected	
Mucous membranes not affected	

Pathology
Histological changes are non-specific, showing slight inflammatory changes in the dermis, oedema, and slight hyperkeratosis.

Pityriasis lichenoides

Pityriasis lichenoides is a less common condition occurring in acute and chronic forms.

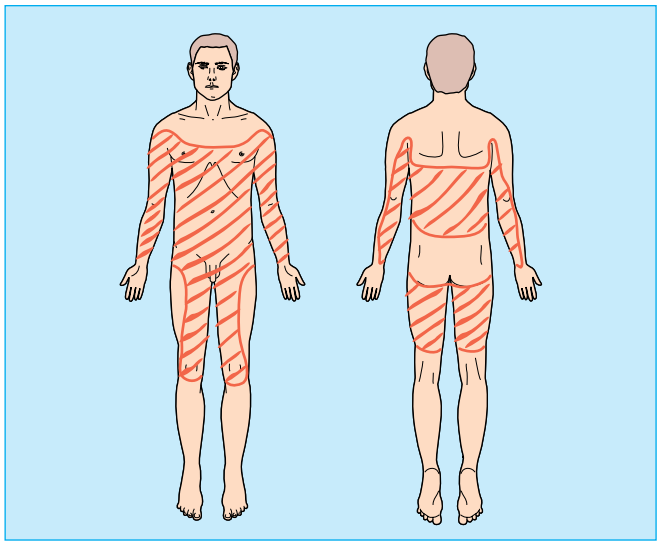
The *acute* form presents with widespread pink papules which itch and form crusts, sometimes with vesicle formation suggestive of chickenpox. There may be ulceration. The lesions may develop in crops and resolve over a matter of weeks.

The *chronic* form presents as reddish brown papules—often with a "mica"-like scale that reveals a smooth, red surface underneath, unlike the bleeding points of psoriasis. In lichen planus there is no superficial scale and blistering is unusual.

The distribution is over the trunk, thighs, and arms, usually sparing the face and scalp.

The underlying pathology—vascular dilatation and a lymphocytic infiltrate with a keratotic scale—is in keeping with the clinical appearance. The cause is unknown.

Treatment is with topical steroids. Ultraviolet light treatment is also helpful.



Pityriasis lichenoides—distribution pattern

Pityriasis versicolor

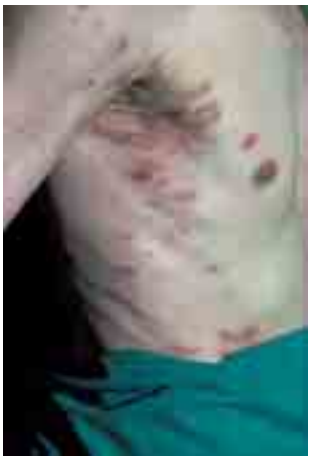
Pityriasis versicolor is a skin eruption that usually develops after sun exposure with white macules on the tanned skin but pale brown patches on the covered areas, hence the name versicolor, or variable colour. The lesions are: (a) flat; (b) only partially depigmented—areas of vitiligo are totally white; and (c) do not show inflammation or vesicles.

The causative organism is a yeast, *Pityrosporum orbiculare*, that takes advantage of some unknown change in the epidermis and develops a proliferative, stubby, mycelial form, *Malassezia furfur*. This otherwise incidental information can be simply put to practical use by taking a superficial scraping from a lesion on to a microscope slide—add a drop of potassium hydroxide or water with a coverslip. The organisms are readily seen under the microscope as spherical yeast forms and mycelial rods, resembling "grapes and bananas" ("spaghetti and meatballs" in the United States).

Treatment is simple: selenium sulphide shampoo applied regularly with ample water while showering or bathing will clear the infection. The colour change may take some time to clear.

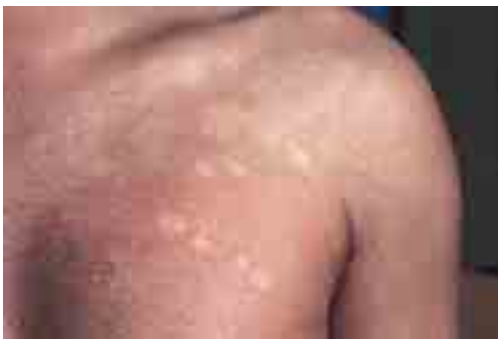
Characteristics of pityriasis lichenoides

Clinical features of psoriasis	Clinical features of eczema
Possible family history	Possible family history
Sometimes related to stress	Sometimes worse with stress
Itching—rare	Usually itching
Extensor surfaces and trunk	Flexor surfaces and face
Well defined, raised lesions	Poorly demarcated lesions
Hyperkeratosis	Oedema, vesicles, lichenification
Scaling, bleeding points beneath scales	Secondary infection sometimes present
Koebner's phenomenon	
Nails affected	
Scalp affected	
Mucous membranes not affected	



Pityriasis lichenoides showing acute erythematous rash

A mica scale pityriasis lichenoides



Pityriasis versicolor skin lesions

Ketoconazole shampoo is an effective alternative. Oral terbinafine, which is very effective in other fungal infections, has no effect.

Desquamating stage of generalised erythema

Any extensive acute erythema, from the erythroderma of psoriasis to a penicillin rash, commonly shows a stage of shedding large flakes of skin—desquamation—as it resolves. If only this stage is seen it can be confused with psoriasis.

Desquamation

Localised lesions with epidermal changes

Psoriasis, seborrhoeic dermatitis, atopic eczema, and contact dermatitis can all present with localised lesions.

Psoriasis may affect only the flexures, occur as a genital lesion, or affect only the palms. The lack of itching and epidermal changes with a sharp edge help in differentiation from infective or infiltrative lesions.

Seborrhoeic dermatitis can occur in the axillae or scalp with no lesions of other areas.

In *atopic eczema* the "classical" sites in children—flexures of the elbows and knees and the face—may be modified in adults to localised vesicular lesions on the hands and feet in older patients. Some atopic adults develop severe, persistent generalised eczematous changes.

Contact dermatitis is usually localised, by definition, to the areas in contact with irritant or allergen. Wide areas can be affected in reactions to clothing or washing powder, and sometimes the reaction extends beyond the site of contact.

Flexural seborrhoeic dermatitis

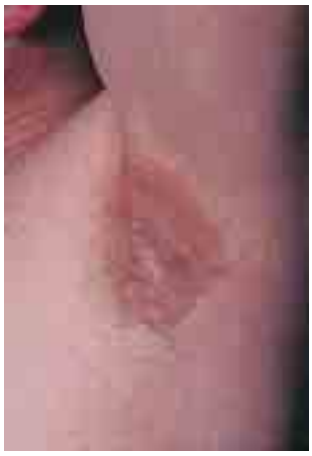
Fungal infections

Apart from athlete's foot, toenail infections, and tinea cruris (most commonly in men), "ringworm" is in fact not as common as is supposed. The damp, soggy, itching skin of athlete's foot is well known. An itching, red diffuse rash in the groin differentiates tinea cruris from psoriasis. However, erythrasma, a bacterial infection, may be confused with seborrhoeic dermatitis and psoriasis—skin scrapings can be taken for culture of *Corynebacterium minutissimum* or, more simply, coral pink fluorescence shown with Wood's light. The scaling macules from dog and cat ringworm (*Microsporum canis*) itch greatly, whereas the indurated pustular, boggy lesion (kerion) of cattle ringworm is quite distinctive.

Fungal infection of the axillae is rare; a red rash here is more likely to be due to erythrasma or seborrhoeic dermatitis.

Tinea cruris is very unusual before puberty and is uncommon in women.

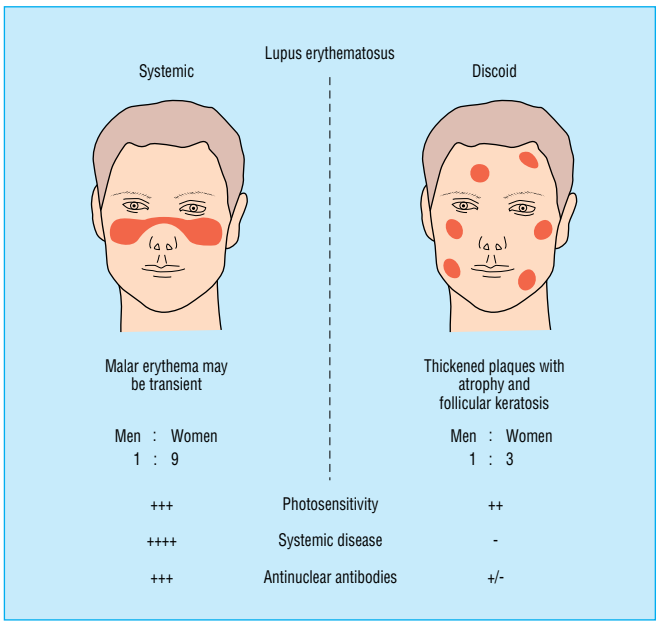
In all cases of suspected fungal infection skin scrapings should be taken on to black paper, in which they can be folded and sent to the laboratory. Special "kits" are available, which contain folded black paper and Sellotape strips on slides for taking a superficial layer of epidermis.



Corynebacterium minutissimum (erythrasma)

Lupus erythematosus

There are two forms of this condition: discoid, which is usually limited to the skin, and systemic, in which the skin lesions are associated with renal disease, arthritis, and other disorders. There is also a subacute type with limited systemic involvement.



Characteristics of systemic and discoid lupus erythematosus

Systemic lupus erythematosus, which is much more common in women than men, can be an acute, fulminating, multisystem disease that requires intensive treatment, or a more chronic progressive illness. Characteristically there is malar erythema with marked photosensitivity and a butterfly pattern. It may be transient. There may be scalp involvement as well with alopecia and also telangiectasia of the perifungal blood vessels. Mouth ulcers may also be present. Systemic involvement may cause nephritis, polyarteritis, leukopenia, pleurisy, myocarditis, and central nervous system involvement.

Systemic lupus erythematosus can present in many forms and imitate other diseases. The facial rash can resemble rosacea, cosmetic allergy, or sun sensitivity. Systemic involvement may present with lassitude, weight loss, anaemia, arthritis, renal failure, dyspnoea, or cardiac signs among others.

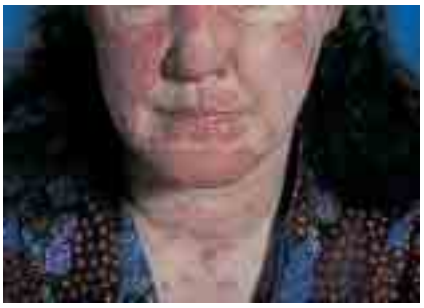
Criteria for making a diagnosis of systemic lupus erythematosus have been established, of which at least four must be present.

In the subacute variety there is less severe systemic involvement, with scattered lesions occurring on the face, scalp, chest, and arms.

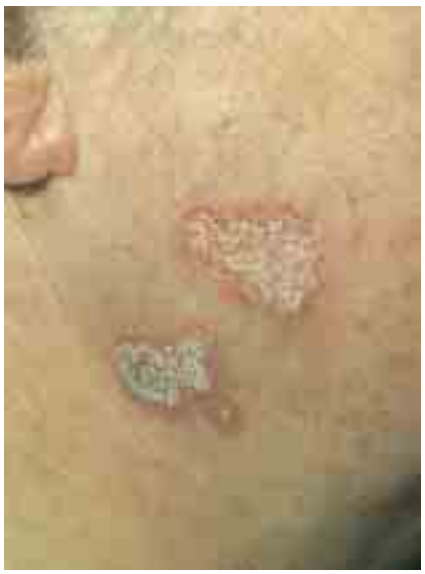
Treatment is with systemic steroids, with immunosuppressive agents if necessary. Antimalarial drugs, such as hydroxychloroquine, are more effective in the subacute type.

In discoid lupus erythematosus there are well defined lesions with a combination of atrophy and hyperkeratosis of the hair follicles giving a "nutmeg grater" appearance. They occur predominantly on the cheeks, nose, and forehead. It is about three times as common in women as men, which is a lower ratio than in the systemic variety. There is a tendency for the skin lesions to gradually progress and to flare up on sun exposure. It is rare for progression to the systemic type to occur.

Treatment is with moderate to very potent topical steroids and hydroxychloroquine by mouth, together with suitable sun screens.



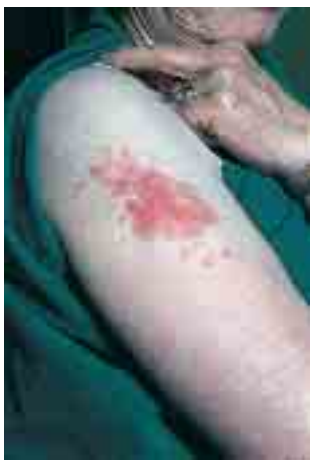
Systemic lupus erythematosus—subacute



Discoid lupus erythematosus

Fixed drug eruptions

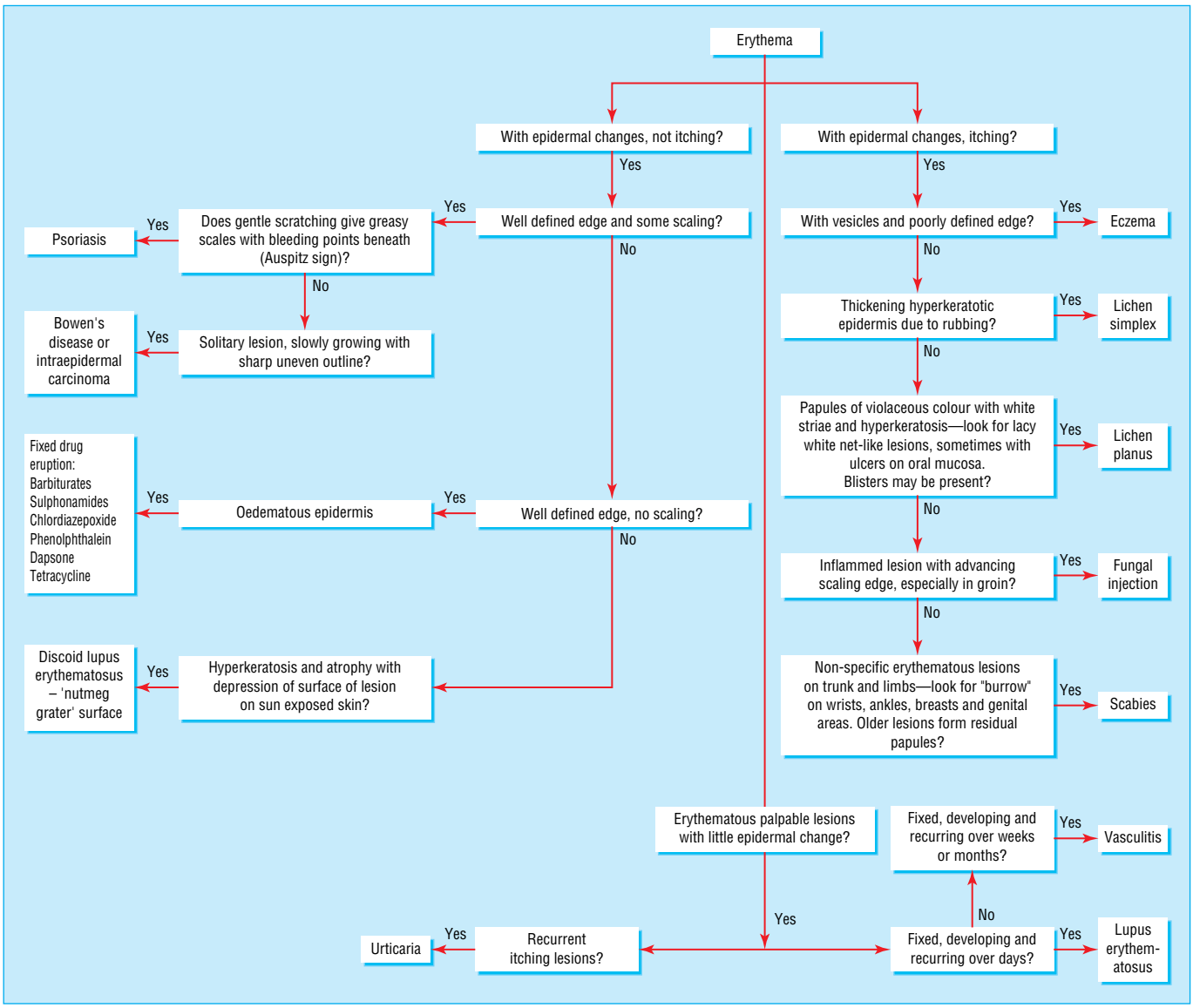
Generalised drug eruptions are considered under erythema, but there is a localised form recurring every time the drug is used. There is usually a well defined, erythematous plaque, sometimes with vesicles. Crusting, scaling, and pigmentation occur as the lesion heals. It is usually found on the limbs, and more than one lesion can occur.



Fixed drug eruption

Criteria for diagnosing systemic lupus erythematosus

- Malar rash
- Discoid plaques
- Photosensitivity
- Mouth ulcers
- Arthritis
- Serositis
- Renal disease
- Neurological disease
- Haematological changes
- Immunological changes
- Antinuclear antibodies



Causes of epidermal rashes

7 Rashes arising in the dermis

The erythemas

Complex reactions occurring in the capillaries and arterioles of the skin cause erythema, which is simply redness of the skin. This may present as flat macules or as papules, which are raised above the surrounding skin. The lesions may be transient or last for weeks, constant or variable in distribution, with or without vesicles.

It is possible to recognise specific patterns within this plethora of clinical signs, but even the most experienced dermatologist may be reduced to making a general diagnosis of "toxic" erythema. The best we can do therefore is to recognise the common types of erythema and list the possible causes. It is then a matter of deciding on the most likely underlying condition or group of conditions—for example, bacterial infection or autoimmune systemic disease.

Morphology and distribution

Because there can be the same cause for a variety of erythematous rashes detailed descriptions are of limited use. None the less, there are some characteristic patterns.

Morbilliform

The presentation of measles is well known, with the appearance of Koplik's spots on the mucosa, photophobia with conjunctivitis, and red macules behind the ears, spreading to the face, trunk, and limbs. The prodromal symptoms and conjunctivitis are absent in drug eruptions. Other viral conditions, including those caused by echoviruses, rubella, infectious mononucleosis, and erythema infectiosum, may have to be considered.

Scarlatiniform

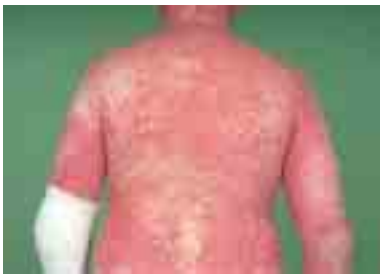
These rashes are similar to that in scarlet fever, when an acute erythematous eruption occurs in relation to a streptococcal infection. Characteristically erythema is widespread on the trunk. There is intense erythema and engorgement of the pharyngeal lymphoid tissue with an exudate and a "strawberry" tongue. Bacterial infections can produce a similar rash, as can drug rashes, without the systemic symptoms.

Figurate erythemas

These are chronic erythematous rashes forming annular or serpiginous patterns. There may be underlying malignancy or connective tissue disease.

Erythema multiforme

Erythema multiforme is sometimes misdiagnosed because of the variety of lesions and number of possible precipitating causes; some of these are listed in the box on the right.



Erythema from antibiotics

Causes of "toxic" erythema

Drugs
- Antibiotics, barbiturates, thiazides

Infections
- Any recent infections such as streptococcal throat infection or erysipelas; spirochaetal infections; viral infections

Systemic causes
- Pregnancy; connective tissue disease; malignancy

Erythema multiforme: precipitating causes

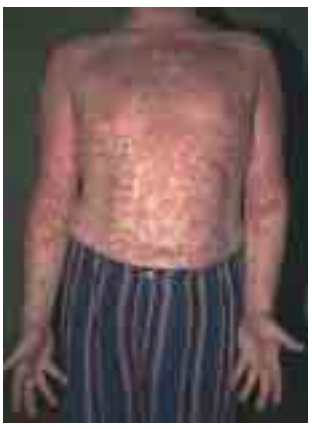
Infections
- Herpes simplex—the commonest cause
- Mycoplasma infection
- Infectious mononucleosis
- Poliomyelitis (vaccine)
- Many other viral and bacterial infections
- Any focal sepsis
- BCG inoculaton

Collagen disease
- Systemic lupus erythematosus
- Polyarteritis nodosa

Neoplasia
- Hodgkin's disease
- Myeloma
- Carcinoma

Chronic inflammation
- Sarcoidosis
- Wegener's granuloma

Drugs
- Barbiturates
- Sulphonamides
- Penicillin
- Phenothiazine and many others



Erythema multiforme

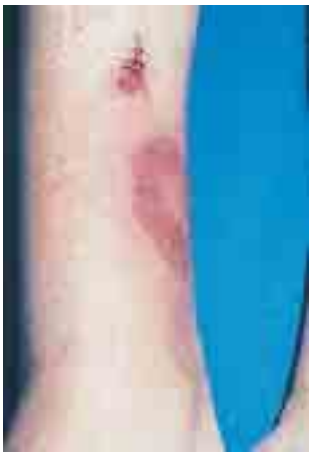
Erythema multiforme

Clinical picture

The usual erythematous lesions occur in crops on the limbs and trunk. Each lesion may extend, leaving a cyanotic centre, which produces an "iris" or "target" lesion. Bullae may develop in the lesions and on the mucous membranes. A severe bullous form, with lesions on the mucous membranes, is known as the Stevens–Johnson syndrome. There may be neural and bronchial changes as well. Barbiturates, sulphonamides, and other drugs, are the most common cause.

Histologically there are inflammatory changes, vasodilatation, and degeneration of the epidermis.

A condition that may be confused is Sweet's syndrome, which presents as acute plum coloured raised painful lesions on the limbs—sometimes the face and neck—with fever. It is more common in women. The alternative name, "acute febrile neutrophilic dermatosis", describes the presentation and the pathological findings of a florid neutrophilic infiltrate. There is often a preceding upper respiratory infection. Treatment with steroids produces a rapid response but recurrences are common.

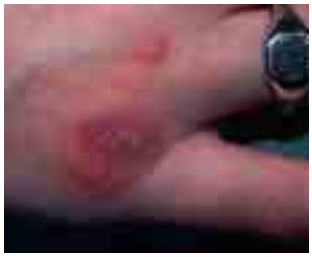


Annular lesions of erythema multiforme

Erythema induratum

Erythema induraturn occurs on the lower legs posteriorly, usually in women, with diffuse, indurated dusky red lesions that may ulcerate. It is more common in patients with poor cutaneous circulation. Epithelioid cell granulomas may form.

This erythema was originally described in association with tuberculous infection elsewhere in the body (Bazin's disease). It represents a vasculitic reaction to the infection, and when there is no tuberculous infection another chronic infection may be responsible.



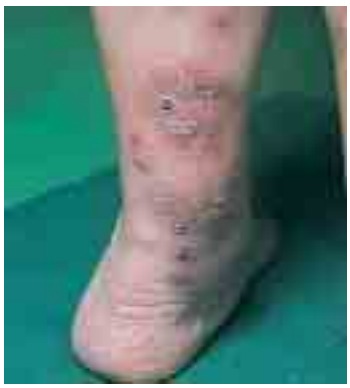
Annular lesions of erythema multiforme

Blistering lesions of erythema multiforme

Erythema nodosum

Erythema nodosum occurs as firm, gradually developing lesions, predominantly on the extensor aspect of the legs. They are tender and progress over four to eight weeks from an acute erythematous stage to residual lesions resembling bruises.

Single or multiple lesions occur, varying in size from 1 to 5 cm. The lesions are often preceded by an upper respiratory tract infection and may be associated with fever and arthralgia. Infections (streptococcal, tuberculous, viral, and fungal) and sarcoidosis are the commonest underlying conditions. Drugs can precipitate erythema nodosum, the contraceptive pill and the sulphonamides being the commonest cause. Ulcerative colitis, Crohn's disease, and lymphoma may also be associated with the condition.

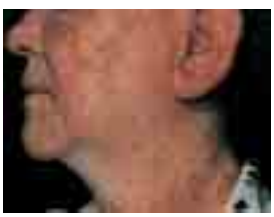


Erythema induratum

Rashes due to drugs

There is an almost infinite variety of types of drug reaction.

External contact with drugs can cause a contact dermatitis presenting with eczematous changes. This occurs commonly with neomycin and bacitracin. Chloramphenicol and sulphonamides from ophthalmic preparations can also cause dermatitis around the eyes. Penicillin is a potent sensitiser so is not used for topical treatment.



Topical neomycin allergy

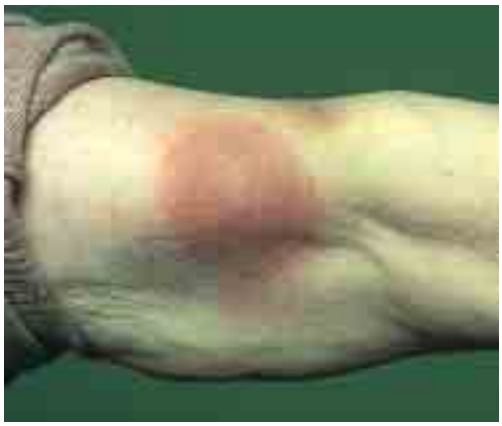
Same patient after withdrawing neomycin

Erythema nodosum

Drugs used systemically can cause a localised fixed drug eruption or a more diffuse macular or papular erythema, symmetrically distributed. In the later stages exfoliation, with shedding scales of skin, may develop. Antibiotics, particularly penicillins, are the most common cause. They also cause erythema multiforme as already mentioned.

Penicillins are the most common cause of drug rashes, which range from acute anaphylaxis to persistent diffuse erythematous lesions. Joint pains, fever, and proteinuria may be associated, as in serum sickness.

Ampicillin often produces a characteristic erythematous maculopapular rash on the limbs seven to 20 days after the start of treatment. Such rashes occur in nearly all patients with infectious mononucleosis who are given ampicillin.



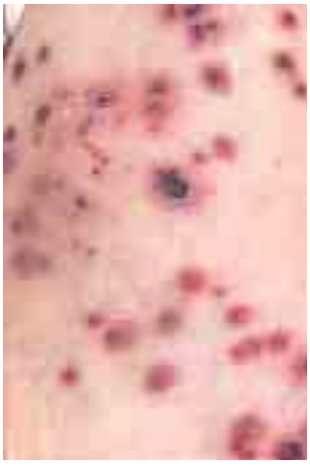
Fixed drug eruption

Vasculitis

Inflammation associated with immune complexes in the capillaries and small blood vessels is part of the pathological changes of many of the conditions described above. The term vasculitis is also used clinically to describe a variable clinical picture with red macules and papules and with necrosis and bruising in severe cases. In children purpura is more prominent and these cases are classified as Henoch–Schönlein purpura. The legs and arms are usually affected. Skin signs are preceded by malaise and fever with arthropathy and there may be associated urticaria. As a high proportion of cases are associated with systemic lesions, it is essential to check for renal, skeletal, gastrointestinal, and central nervous system disease. In children with Henoch–Schönlein purpura nephritis is common.

Vasculitis

Acute vasculitis with necrosis



Necrotising angiitis

Drug reactions

Blistering eruptions
- Barbiturates
- Sulphonamides
- Iodines or bromides
- Chlorpropamide
- Salicylates
- Phenylbutazone

Lichen planus like reactions
- Chloroquine
- Chlorothiazide
- Chlorpropamide

Photosensitivity (seen on areas exposed to light)
- Thiazide diuretics
- Sulphonamides
- Tetracyclines

Vasculitis
- Inflammation around dilated capillaries and small blood vessels
- A common component of the erythemas
- May occur as red macules and papules with necrotic lesions on the extremities
- In children a purpuric type (Henoch–Schönlein purpura) occurs in association with nephritis
- Systemic lesions may occur, with renal, joint, gastrointestinal, and central nervous system involvement

Purpura

Is seen on the skin as a result of:
- Thrombocytopenia—platelet deficiency
- Senile purpura—due to shearing of capillaries as a result of defective supporting connective tissue
- Purpura in patients on corticosteroid treatment—similar to senile purpura
- Schamberg's disease—brown macules and red spots resembling cayenne pepper on the legs of men
- Associated vasculitis

Some conditions associated with vasculitis
- Infection—streptococcal, hepatitis
- Drugs—numerous, including sulphonamides, penicillin, phenothiazine, phenacitin
- Chemicals—insecticides, weed killers, phenolic compounds
- Connective tissue diseases—systemic lupus erythematosus, rheumatoid arthritis
- Lymphoma and leukaemia
- Dysproteinaemias

Urticaria

In this condition itching red weals develop; they resemble the effects of stinging nettle (*Urtica dioica*) on the skin. The condition may be associated with allergic reactions, infection, or physical stimuli, but in most patients no cause can be found. Similar lesions may precede, or be associated with, vasculitis (urticarial vasculitis), pemphigoid, or dermatitis herpetiformis.

The histological changes may be very slight but usually there is oedema, vasodilatation, and a cellular infiltrate of lymphocytes, polymorphs, and histiocytes. Various vasoactive substances are thought to be involved, including histamine, kinins, leukotrienes, prostaglandins, and complement.

Angio-oedema is due to oedema of the subcutaneous tissues; it can occur rapidly and may involve the mucous membranes. Hereditary angio-oedema is a rare form with recurrent severe episodes of subcutaneous oedema, swelling of the mucous membranes, and systemic symptoms. Laryngeal oedema is the most serious complication.

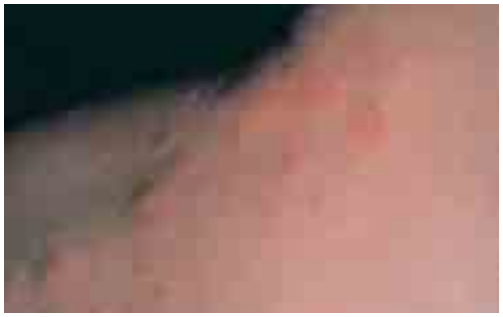
The *physical urticarias*, which account for about 25% of cases, include dermatographism and the pressure, cold, heat, solar, cholinergic, and aquagenic urticarias.

Dermatographism is an exaggerated release of histamine from stroking the skin firmly with a hard object, such as the end of a pencil. *Pressure urticaria* is caused by sustained pressure from clothing, hard seats, and footwear; it may last some hours. *Cold urticaria* varies in severity and is induced by cold, particularly by cold winds or by the severe shock of bathing in cold water. It appears early in life—in infancy in the rare familial form. In a few cases abnormal serum proteins may be found. *Heat urticaria* is rare, but warm environments often make physical urticaria worse. *Solar urticaria* is a rare condition in which sunlight causes an acute urticarial eruption. Tolerance to sun exposure may develop in areas of the body normally exposed to sun. There is sensitivity to a wide spectrum of ultraviolet light. *Cholinergic urticaria* is characterised by the onset of itching urticarial papules after exertion, stress, or exposure to heat. The injection of cholinergic drugs induces similar lesions in some patients. *Aquagenic urticaria* occurs on contact with water, regardless of the temperature.

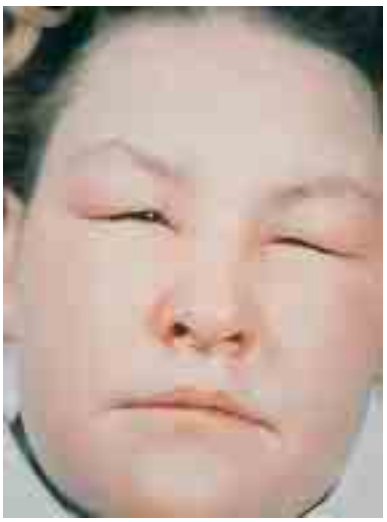
Non-physical urticaria may be acute in association with allergic reactions to insect bites, drugs, and other factors. Chronic recurrent urticaria is fairly common. Innumerable causes have been suggested but, to the frustration of patient and doctor alike, it is often impossible to identify any specific factor.

Treatment of urticaria

- Eliminate possible causative factors, such as aspirin, and by a diet free from food additives
- Antihistamines. Also, H_2 blockers, for example, cimetidine
- Adrenaline can be used for acute attacks, particularly if there is angio-oedema of the respiratory tract
- Systemic corticosteroids should not be used for chronic urticaria but may be needed for acute urticarial vasculitis



Urticaria



Angio-oedema

Some reported causes of non-physical urticaria

- Food allergies—fish, eggs, dairy products, chocolate, nuts, strawberries, pork, tomatoes
- Food additives—for example, tartrazine dyes, sodium benzoates
- Salicylates—both in medicines and foods
- Infection—bacterial, viral, and protozoal
- Systemic disorders—autoimmune and "collagen" diseases; reticuloses, carcinoma, and dysproteinaemias
- Contact urticaria—may occur from contact with meat, fish, vegetables, plants, and animals, among many other factors
- Papular urticaria—a term used for persistent itching papules at the site of insect bites; it is also sometimes applied to urticaria from other causes
- Inhalants—for example, house dust, animal danders

Dermatographism

Further reading

Berlit P, Moore P. *Vasculitis, rheumatic diseases and the nervous system.* Berlin: Springer-Verlag, 1992

Champion RH, Greaves MW, Black AK. *The urticarias.* Edinburgh: Churchill Livingstone, 1985

Czametzki BM. *Urticaria.* Berlin: Springer-Verlag, 1986

8 Blisters and pustules

Development, duration, and distribution

Several diseases may present with blisters or pustules. There is no common condition that can be used as a "reference point" with which less usual lesions can be compared in the same way as rashes can be compared with psoriasis. A different approach is needed for the assessment of blistering or pustular lesions, based on the history and appearance, and is summarised as the three Ds: development, duration, and distribution.

Development
Was there any preceding systemic illness—as in chickenpox, hand, foot, and mouth disease, and other viral infections? Was there a preceding area of erythema—as in herpes simplex or pemphigoid? Is the appearance of the lesions associated with itching—as in herpes simplex, dermatitis herpetiformis, and eczematous vesicles on the hands and feet?

Duration
Some acute blistering arises rapidly—for example, in allergic reactions, impetigo, erythema multiforme, and pemphigus. Other blisters have a more gradual onset and follow a chronic course—as in dermatitis herpetiformis, pityriasis lichenoides, and the bullae of porphyria cutanea tarda. The rare genetic disorder epidermolysis bullosa is present from, or soon after, birth.

Distribution
The distribution of blistering rashes helps considerably in making a clinical diagnosis. The most common patterns of those that have a fairly constant distribution are shown.

Itching is a very useful symptom. If all the accessible lesions are scratched and it is hard to find an intact blister it is probably an itching rash.

The differential diagnosis of blistering eruptions

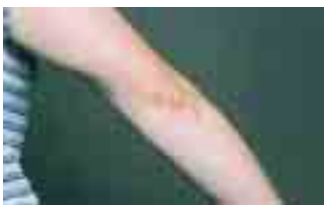
Widespread blisters
- Eczema—lichenification and crusting, itching
- Dermatitis herpetiformis—itching, extensor surface, persistent
- Chickenpox—crops of blisters, self limiting, prodromal illness
- Pityriasis lichenoides—pink papules, developing blisters
- Erythema multiforme—erythematous and "target" lesions, mucous membranes affected
- Pemphigoid—older patients, trunk, and flexures affected. Preceding erythematous lesions, deeply situated, tense blisters
- Pemphigus—adults, widespread superficial blisters, mucous membranes affected (erosions)
- Drug eruptions—history of drugs prescribed, overdose (barbiturates, tranquillisers)

Localised blisters
- Eczema—"pompholyx" blisters on hand and feet, itching
- Allergic reactions, including topical medication, insect bites
- Psoriasis—deep, sterile, non-itching blisters on palms and soles
- Impetigo—usually localised, staphylococci and streptococci isolated
- Herpes simplex—itching lesions developing turbid blisters

Diseases presenting with blisters and pustules

Itching	Non-itching
Eczema pompholyx on hands and feet	Erythema multiforme
Allergic reactions	Pemphigus vulgaris
Dermatitis herpetiformis	Bullous pemphigoid
Chickenpox	Bullous impetigo
Herpes simplex	Insect bite allergy
	Pustular psoriasis on hands and feet

Dermatitis herpetiformis



Dermatitis herpetiformis

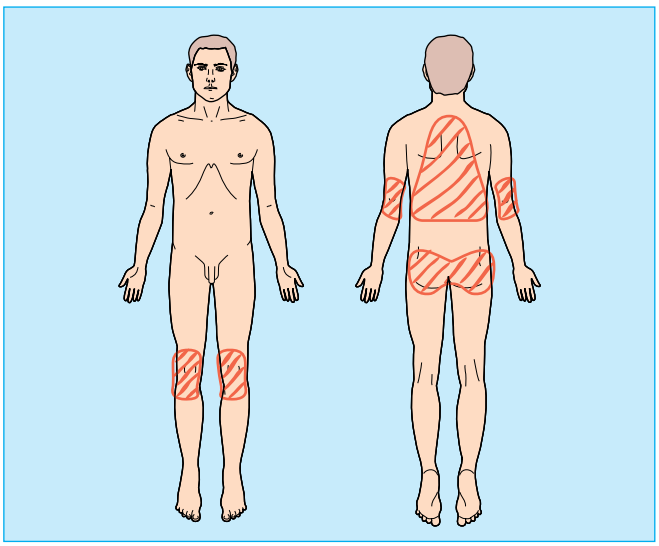
Clinical features: widespread blisters

Chickenpox
Chickenpox is so well known in general practice that it is rarely seen in hospital clinics and is sometimes not recognised. The prodromal illness lasts one to two days and is followed by erythematous lesions that rapidly develop vesicles, then pustules, followed by crusts in two to three days. Crops of lesions develop at the same sites—usually on the trunk, face, scalp, and limbs. The oral mucosa may be affected. The condition is usually benign.

Dermatitis herpetiformis
Dermatitis herpetiformis occurs in early and middle adult life and is characterised by symmetrical, intensely itching vesicles on the trunk and extensor surfaces. The vesicles are superficial. The onset is gradual, but may occur rapidly. The distribution is shown in the diagram.

Variants of dermatitis herpetiformis are larger blisters forming bullae and erythematous papules and vesicles.

Associated conditions
Coeliac disease with villous atrophy and gluten intolerance may occur in association with dermatitis herpetiformis. Linear IgA



Dermatitis herpetiformis—distribution pattern

disease is a more severe, widespread disease, in which there are "linear" deposits of IgA along the basement membrane of the epidermis and not only at the tips of the papillae as in dermatitis herpetiformis. Treatment is with dapsone or sulfapyridine together with a gluten free diet.

Erythema multiforme with blisters

Blisters can occur on the lesions of erythema multiforme to a variable degree; when severe, generalised, and affecting the mucous membranes it is known as Stevens–Johnson syndrome. The typical erythematous maculopapular changes develop over one to two days with a large blister (bulla) developing in the centre of the target lesions. In severe progressive cases there is extensive disease of the mouth, eyes, genitalia, and respiratory tract. The blisters are subepidermal, although some basement membrane remains on the floor of the blister.

Pityriasis lichenoides varioliformis acuta

As the name implies lichenified papules are the main feature of pityriasis lichenoides varioliformis acuta (or Mucha–Habermann disease), but vesicles occur in the acute form. Crops of pink papules develop centrally, with vesicles, necrosis, and scales—resembling those of chickenpox—hence the "varioliformis". There is considerable variation in the clinical picture, and a prodromal illness may occur. The condition may last from six weeks to six months. No infective agent has been isolated. The pathological changes parallel the clinical appearance with inflammation around the blood vessels and oedema within the dermis.

Pemphigoid

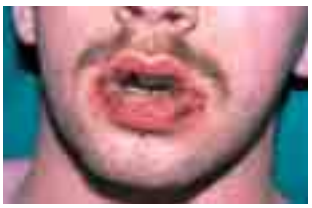
The bullous type of pemphigoid is a disease of the elderly in which tense bullae develop rapidly, often with a preceding erythematous rash, as well as on normal skin. It is mainly seen in the elderly and is slightly more common in women.The flexural aspects of the limbs and trunk and flexures are mainly affected. The bullae are subepidermal and persistent, with antibodies deposited at the dermo-epidermal junction. Unlike pemphigus there is a tendency for the condition to remit after many months.

Another type of pemphigoid occurs in which there is scarring of the oral mucous membrane and the conjunctiva. Occasionally localised lesions are seen on the legs with evidence of an immune reaction, but often the absence of circulating antibasement membrane antibodies. This is a relatively benign condition and often responds to topical steroids.

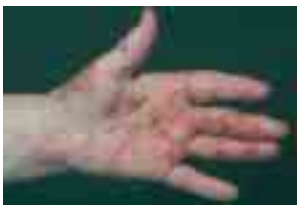
Treatment is with corticosteroids by mouth, 40–60 mg daily in most patients, although higher doses are required by some. Azathioprine aids remission, with reduced steroid requirements, but takes some weeks to produce an effect. It is essential to check the serum thiopurine methyl transferase (TPMT) level before starting treatment. Patients with low levels have impaired ability to metabolise azathioprine and are likely to suffer toxic effects. Topical steroids can be used on developing lesions.

Chronic scarring pemphigoid affects the mucous membranes with small bullae that break down, leading to erosions and adhesions in the conjunctivae, mouth, pharynx, and genitalia.

There is also a localised type of pemphigoid occurring on the legs of elderly women that runs a benign self-limiting course.



Erythema multiforme in Stevens–Johnson syndrome

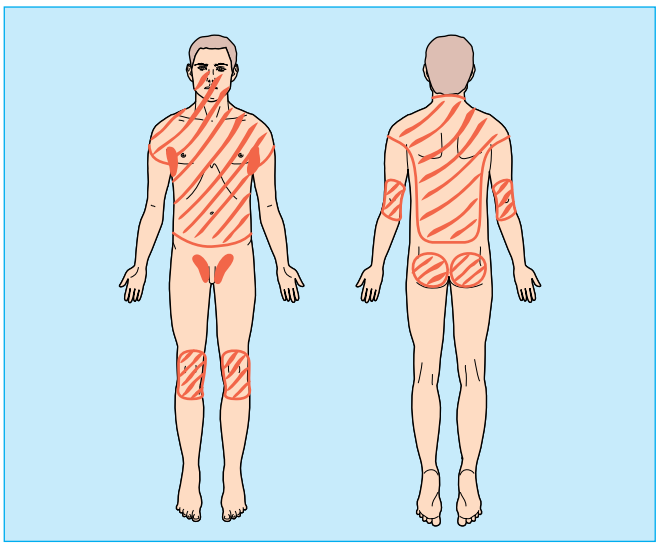


Erythema multiforme showing blisters

Pityriasis lichenoides

Bullous pemphigoid

Mucous membrane pemphigoid

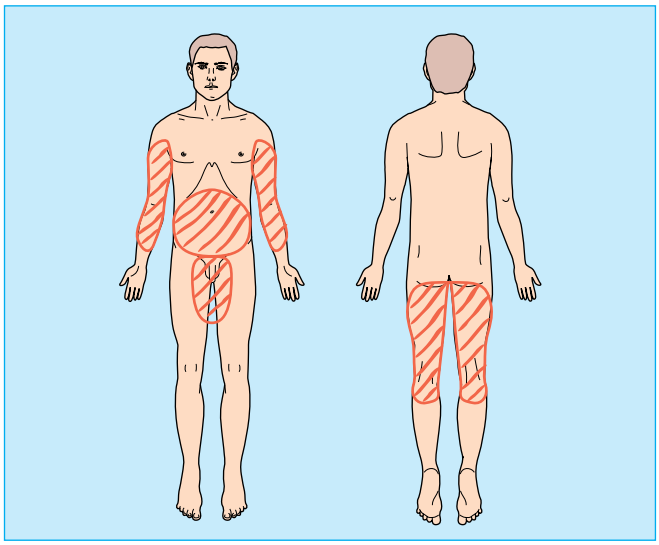


Bullous pemphigoid—distribution pattern

Pemphigus
The most common form of pemphigus vulgaris is a chronic progressive condition with widespread superficial bullae arising in normal skin. In about half of the cases this is preceded by blisters and erosions in the mouth. The bullae are easily broken, and even rubbing apparently normal skin causes the superficial epidermis to slough off (Nikolsky sign). These changes are associated with the deposition of immunoglobulin in the epidermal intercellular spaces. It is a serious condition with high morbidity, despite treatment with steroids and azathioprine. Pemphigus vegetans and pemphigus erythematosus are less common variants.

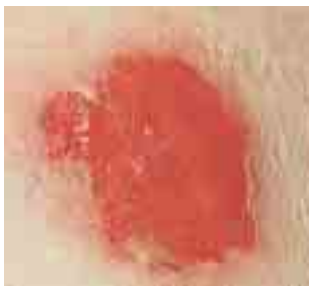
Differential diagnosis of ulcers in the mouth

- Trauma (dentures)
- Aphthous ulcers
- *Candida albicans* infection
- Herpes simplex
- Erythema multiforme (from drugs)
- Pemphigus
- Lichen planus
- Carcinoma



Pemphigus—distribution pattern

Pemphigus vulgaris



Nikolsky sign

Clinical features: localised blisters

Pompholyx, which means "a bubble", is characterised by persistent, itchy, clear blisters on the fingers, which may extend to the palms, with larger blisters. The feet may be affected. Secondary infection leads to turbid vesicle fluid. Pompholyx may be associated with a number of conditions—atopy, stress, fungal infection elsewhere, and allergic reactions. It may occur as a result of ingesting nickel in nickel sensitive patients and a similar reaction has been reported to neomycin.

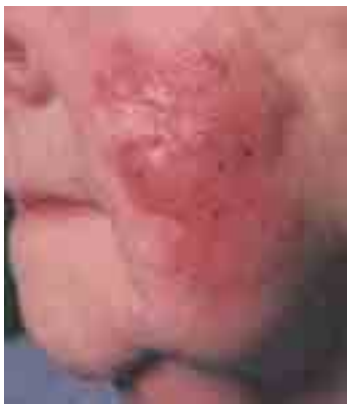
Pustular psoriasis is characterised by deep seated sterile blisters, often with no sign of psoriasis elsewhere—hence the term palmopustular pustulosis. Foci of sepsis have long been considered a causative factor and recent studies have shown a definite association with cigarette smoking. The pattern of HLA antigens indicates that this may be a separate condition from psoriasis.

Bullous impetigo is seen in children and adults. Staphylococci are usually isolated from the blister fluid. The blisters are commonly seen on the face and are more deeply situated than in the non-bullous variety.

Herpes simplex. Primary infection with type 1 virus occurs on the face, lips, and buccal mucosa in children and young adults. Type II viruses cause genital infection. Itching may be severe.

Pompholyx

Pustular psoriasis



Bullous impetigo

Herpes simplex—type I virus infection

Herpes simplex—type II virus infection

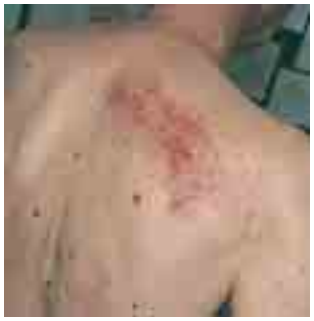
Herpes zoster is due to varicella virus producing groups of vesicles in a dermatome distribution, usually thoracic, trigeminal, or lumbosacral. It is more common after the fourth decade of life.

Insect bite allergy. Large blisters, which are usually not itching, can occur on the legs of susceptible individuals.

Bullous drug eruptions. Fixed drug eruptions can develop bullae, and some drugs can cause a generalised bullous eruption, particularly:

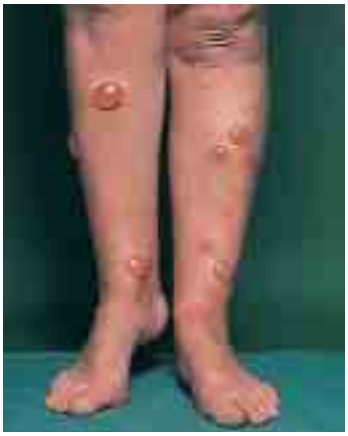
Barbiturates (particularly if taken in overdose)
Sulphonamides
Penicillins
Penicillamine, Captopril } pemphigus-like blisters
Frusemide (may be phototoxic)

Remember that there may be an associated erythematous eruption.

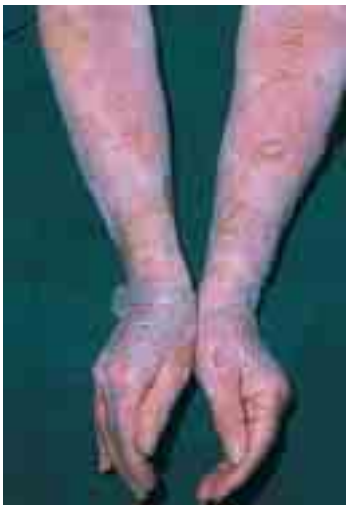


Herpes zoster

Herpes zoster



Insect bite allergy



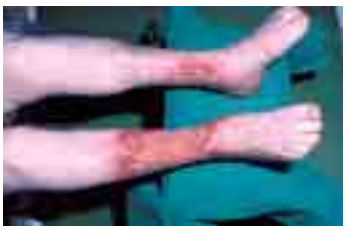
Drug reaction to sulfapyridine

9 Leg ulcers

Although the patient will not probably die of this disease, yet, without great care, it may render her miserable. The disease may be very much relieved by art, and it is one of very common occurrence.

Sir Benjamin Brodie (1846)

Despite the great increase in our understanding of the pathology of leg ulcers, their management is still largely "art". Consequently there are numerous treatments, each with their enthusiastic advocates. There are, however, basic concepts which are helpful in management. As about 95% of leg ulcers are of the "venous" or gravitational variety these will be considered first.



Venous leg ulcers

Pathology of venous ulcers

The skin

Ulcers arise because the skin dies from inadequate provision of nutrients and oxygen. This occurs as a consequence of (a) oedema in the subcutaneous tissues with poor lymphatic and capillary drainage and (b) the extravascular accumulation of fibrinous material that has leaked from the blood vessels. The result is a rigid cuff around the capillaries, preventing diffusion through the wall, and fibrosis of the surrounding tissues.

Varicose veins

Ulcers and fibrosis

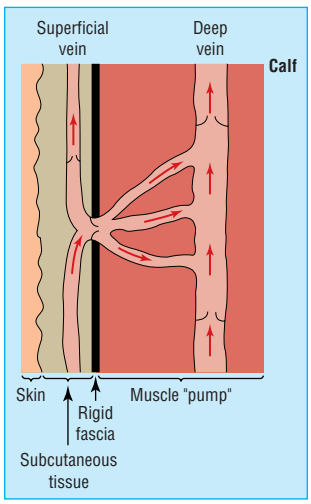
The blood vessels

Arterial perfusion of the leg is usually normal or increased, but stasis occurs in the venules. The lack of venous drainage is a consequence of incompetent valves between the superficial veins and the deeper large veins on which the calf muscle "pump" acts. In the normal leg there is a superficial low pressure venous system and deep high pressure veins. If the blood flow from superficial to deep veins is reversed then the pressure in the superficial veins may increase to a level that prevents venous drainage, Vancose veins with "back pressure" causing stasis and oedema.

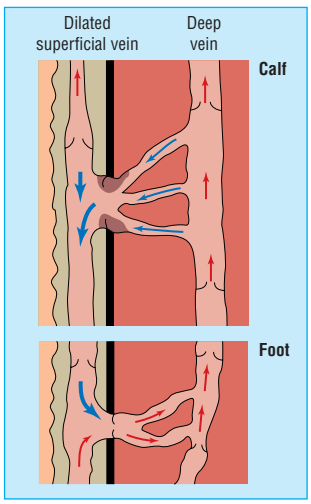
Incompetent valves

Incompetent valves leading to gravitational ulcers may be preceded by:

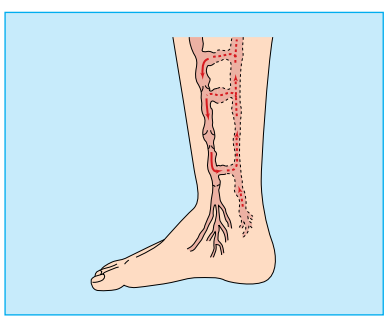
(1) deep vein thrombosis associated with pregnancy or, less commonly, leg injury, immobilisation, or infarctions in the past
(2) primary long saphenous vein insufficiency
(3) familial venous valve incompetence that presents at an earlier stage—there is a familial predisposition in half of all patients with leg ulcers—or
(4) deep venous obstruction.



Healthy valves in legs



Incompetent valves in legs



Incompetent flow in legs

Who gets ulcers?
Mainly women get ulcers—2% of those over 80 have venous ulcers as a long term consequence of the factors listed above. Leg ulcers are more likely to occur and are more severe in obese people.

Clinical changes

Oedema and fibrinous exudate often lead to fibrosis of the subcutaneous tissues, which may be associated with localised loss of pigment and dilated capillary loops, an appearance known as "atrophic blanche". This occurs around the ankle with oedema and dilated tortuous superficial veins proximally and can lead to "champagne bottle legs", the bottle, of course, being inverted. Ulceration often occurs for the first time after a trivial injury.

Lymphoedema results from obliteration of the superficial lymphatics, with associated fibrosis. There is often hypertrophy of the overlying epidermis with a "polypoid" appearance, also known as lipodermatosclerosis.

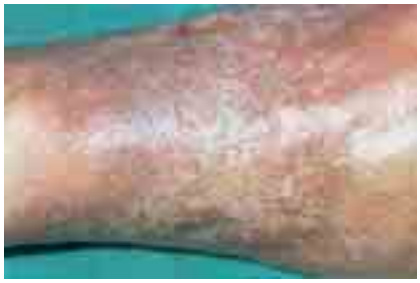
Venous ulcers occur around the ankles, commonly over the medial malleolus. The margin is usually well defined with a shelving edge, and a slough may be present. There may also be surrounding eczematous changes. Venous ulcers are not usually painful but arterial ulcers are.

It is important to check the pulses in the leg and foot as compression bandaging of a leg with impaired blood flow can cause ischaemia and necrosis.

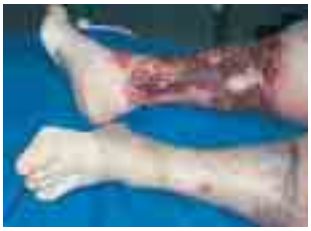
Treatment

When new epidermis can grow across an ulcer it will, and the aim is to produce an environment in which this can take place. To this end several measures can be taken:

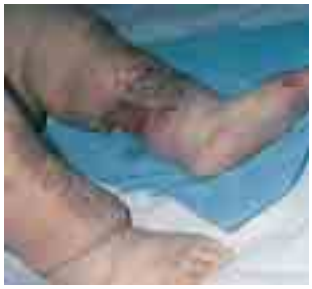
(1) Oedema may be reduced by means of: (a) diuretics; (b) keeping the legs elevated when sitting; (c) avoiding standing as far as possible; (d) raising the heels slightly from time to time helps venous return by the "calf muscle pump"; (e) applying compression bandages to create a pressure gradient towards the thigh.
(2) Exudate and slough should be removed. Lotions can be used to clean the ulcer and as compresses—0·9% saline solution, sodium hypochlorite solution, Eusol, or 5% hydrogen peroxide.



Atrophie blanche

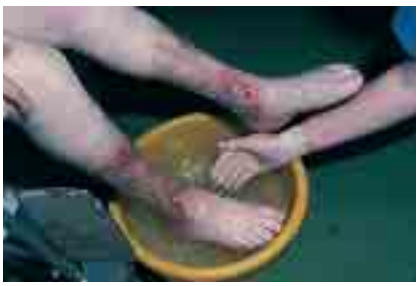


"Champagne bottle legs" with ulceration

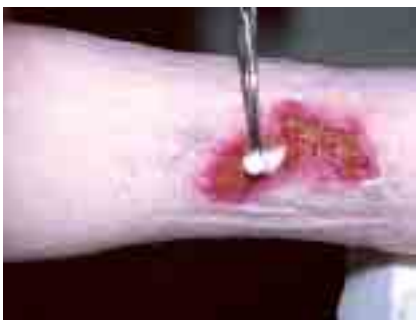


"Champagne bottle legs" with ulceration

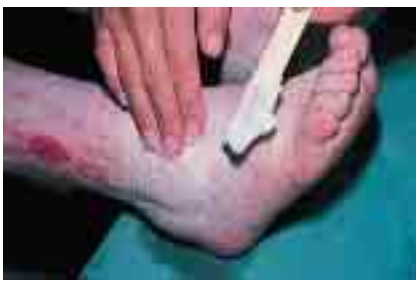
Bandaging



Cleaning the ulcers



Cleaning with saline solution



Applying antiseptic cream

There is some evidence that antiseptic solutions and chlorinated solutions (such as sodium hypochlorite and Eusol) delay collagen production and cause inflammation. Enzyme preparations may help by "digesting" the slough. To prevent the formation of granulation tissue use silver nitrate 0·25% compresses, a silver nitrate "stick" for more exuberant tissue, and curettage, if necessary.

(3) The dressings applied to the ulcer can consist of: (a) simple non-stick, paraffin gauze dressings—an allergy may develop to those with an antibiotic; (b) wet compresses with saline or silver nitrate solutions for exudative lesions; (c) silver sulfadiazine (Flamazine) or hydrogen peroxide creams (Hioxyl); and (d) absorbent dressings, consisting of hydrocolloid patches or powder, which are helpful for smaller ulcers.

(4) Paste bandages, impregnated with zinc oxide and antiseptics or ichthammol, help to keep dressings in place and provide protection. They may, however, traumatise the skin, and allergic reactions to their constituents are not uncommon.

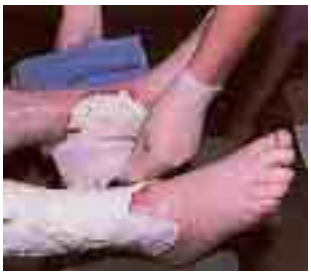
(5) Treatment of infection is less often necessary than is commonly supposed. All ulcers are colonised by bacteria to some extent, usually coincidental staphylococci. A purulent exudate is an indication for a broad spectrum antibiotic and a swab for bacteriology. Erythema, oedema, and tenderness around the ulcers suggest a β haemolytic streptococcal infection, which will require long term antibiotic treatment. Dyes can be painted on the edge of the ulcer, where they fix to the bacterial wall as well as the patient's skin. In Scotland bright red eosin is traditionally used, while in the south a blue dye, gentian violet, is favoured. Systemic antibiotics have little effect on ulcers but are indicated if there is surrounding cellulitis. A swab for culture and sensitivity helps to keep track of organisms colonising the area.

(6) Surrounding eczematous changes should be treated. Use topical steroids, not more than medium strength, avoiding the ulcer itself. Ichthammol 1% in 15% zinc oxide and white soft paraffin or Ichthopaste bandages can be used as a protective layer, and topical antibiotics can be used if necessary. It is important to remember that any of the commonly used topical preparations can cause an allergic reaction: neomycin, lanolin, formaldehyde, tars, Chinaform (the "C" of many proprietary steroids).

(7) Skin grafting can be very effective. There must be a healthy viable base for the graft, with an adequate blood supply; natural re-epithelialisation from the edges of the ulcer is a good indication that a graft will be supported. Pinch grafts or partial thickness grafts can be used. Any clinical infection, particularly with pseudomonal organisms, should be treated.

(8) Maintaining general health, with adequate nutrition and weight reduction, is important.

(9) Corrective surgery for associated venous abnormalities.



Applying ichthopaste bandages as a protective layer

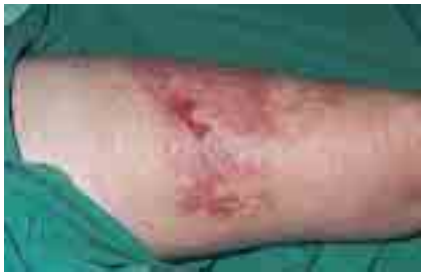
Bandaging ulcers

Treatment of venous leg ulcers

- Take measures to eliminate oedema and reduce weight—make sure the patient understands these
- Never apply steroid preparations to the ulcer itself or it will not heal. Make sure that both nurses and patients are aware of this
- Beware of allergy developing to topical agents—especially to antibiotics
- There is no need to submit the patient to a variety of antibiotics according to the differing bacteria isolated from leg ulcer slough, unless there is definite evidence of infection of adjacent tissue clinically
- A vascular "flare" around the ankle and heel with varicose veins, sclerosis, or oedema indicates a high risk of ulceration developing
- Make sure arterial pulses are present. A Doppler apparatus can be used

Arterial ulcers

Ulcers on the leg also occur as a result of: (a) atherosclerosis with poor peripheral circulation, particularly in older patients; (b) vasculitis affecting the larger subcutaneous arteries; and (c) aterial obstruction in macroglobulinaemia, cryoglobulinaemia, polycythaemia, and "collagen" disease, particularly rheumatoid arthritis.



Arterial ulcer

Arterial ulcers are sharply defined and accompanied by pain, which may be very severe, especially at night. The leg, especially the pretibial area, is affected rather than the ankle. In patients with hypertension a very tender ulcer can develop posteriorly (Martorelli's ulcer).

As mentioned above, compression bandaging will make arterial ulcers worse and may lead to ischaemia of the leg.

Diagnosis

The differing presentation of arterial and venous ulcers helps in distinguishing between them, but some degree of aterial insufficiency often complicates venous ulcers.

Phlebography and Doppler ultrasound may help in detecting venous incompetence and arterial obstruction, which can sometimes be treated surgically.

Ulcers on the leg may also occur secondary to other diseases, because of infection, in malignant disease, and after trauma.

Secondary ulcers

Ulcers occur in diabetes, in periarteritis nodosa, and in vasculitis. Pyoderma gangrenosum, a chronic necrotic ulcer with surrounding induration, may occur in association with ulcerative colitis or rheumatoid vasculitis.

Infections

Infections that cause ulcers include staphylococcal or streptococcal infections, tuberculosis (which is rare in the United Kingdom but may be seen in recent immigrants), and anthrax.

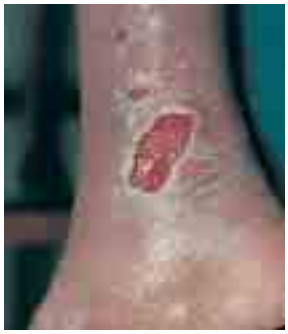
Malignant diseases

Squamous cell carcinoma may present as an ulcer or, rarely, develop in a pre-existing ulcer. Basal cell carcinoma and melanoma may develop into ulcers, as may Kaposi's sarcoma.

Trauma

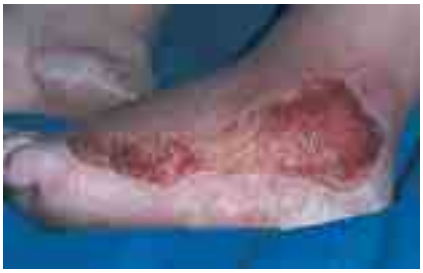
Patients with diabetic or other types of neuropathy are at risk of developing trophic ulcers. Rarely they may be self induced—"dermatitis artefacta".

Ulcer in diabetic foot



Tuberculous ulceration

Squamous cell carcinoma in venous ulcer



Dermatitis artefacta

Further reading

Browse NL, Burnand KG, Lea M. *Diseases of the veins: pathology, diagnosis and treatment.* London: Edward Arnold, 1988

Kappert A. *Diagnosis of peripheral vascular disease.* Berne: Hauber, 1971

Ryan TJ. *The management of leg ulcers,* 2nd ed. Oxford: Oxford University Press, 1987

10 Acne and rosacea

Acne goes with adolescence, a term derived from the Greek "acme" or prime of life. The young girl who is desperately aware of the smallest comedo and the young man, with his face or back a battlefield of acne cysts and scars, are familiar to us all. Both need treatment and help in coming to terms with their condition.

Acne cysts and scars

What is acne?

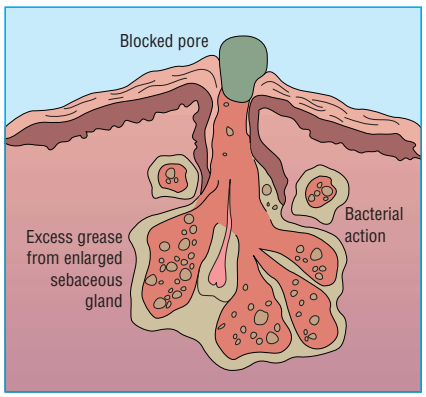
Acne lesions develop from the sebaceous glands associated with hair follicles—on the face, external auditory meatus, back, chest, and anogenital area. (Sebaceous glands are also found on the eyelids and mucosa, prepuce and cervix, where they are not associated with hair follicles.) The sebaceous gland contains holocrine cells that secrete triglycerides, fatty acids, wax esters, and sterols as "sebum". The main changes in acne are:

(1) an increase in sebum secretion;
(2) thickening of the keratin lining of the sebaceous duct, to produce blackheads or comedones—the colour of blackheads is due to melanin, not dirt;
(3) an increase in *Propionibacterium acnes* bacteria in the duct;
(4) an increase in free fatty acids;
(5) inflammation around the sebaceous gland, probably as a result of the release of bacterial enzymes.

Sebaceous gland—histology of acne

Underlying causes
There are various underlying causes of these changes.

Hormones
Androgenic hormones increase the size of sebaceous glands and the amount of sebum in both male and female adolescents. Oestrogens have the opposite effect in prepubertal boys and eunuchs. In some women with acne there is lowering of the concentration of sex hormone binding globulin and a consequent increase in free testosterone concentrations. There is probably also a variable increase in androgen sensitivity. Oral contraceptives containing more than 50 micrograms ethinyloestradiol can make acne worse and the combined type may lower sex hormone binding globulin concentrations, leading to increased free testosterone. Infantile acne occurs in the first few months of life and may last some years. Apart from rare causes, such as adrenal hyperplasia or virilising tumours, transplacental stimulation of the adrenal gland is thought to result in the release of adrenal androgens—but this does not explain why the lesions persist. It is more common in boys.



Sebaceous gland—pathology in acne

Fluid retention
The premenstrual exacerbation of acne is thought to be due to fluid retention leading to increased hydration of and swelling of the duct. Sweating also makes acne worse, possibly by the same mechanism.

Diet
In some patients acne is made worse by chocolate, nuts, and coffee or fizzy drinks.

Hormones—the cause of all the trouble

- Androgens
  - increase the size of sebaceous glands
  - increase sebum secretion
- Androgenic adrenocorticosteroids—have the same effect
- Oestrogens—have the opposite effect

Seasons
Acne often improves with natural sunlight and is worse in winter. The effect of artificial ultraviolet light is unpredictable.

External factors
Oils, whether vegetable oils in the case of cooks in hot kitchens or mineral oils in engineering, can cause "oil folliculitis", leading to acne-like lesions. Other acnegenic substances include coal tar, dicophane (DDT), cutting oils, and halogenated hydrocarbons (polychlorinated biphenols and related chemicals). Cosmetic acne is seen in adult women who have used cosmetics containing comedogenic oils over many years.

Iatrogenic factors
Corticosteroids, both topical and systemic, can cause increased keratinisation of the pilosebaceous duct. Androgens, gonadotrophins, and corticotrophin can induce acne in adolescence. Oral contraceptives of the combined type can induce acne, and antiepileptic drugs are reputed to cause acne.

Types of acne

Acne vulgaris
- Affects comedogenic areas
- Occurs mainly in puberty, in boys more than girls
- Familial tendency

Infantile acne
- Face only
- Clears spontaneously

Severe acne
- Acne conglobata
- Pyoderma faciale
- Gram negative folliculitis

Occupational acne
- Oils
- Coal and tar
- Chlorinated phenols
- DDT and weedkillers

Steroids
- Systemic or topical

Hormones
- Combined type of oral contraceptives and androgenic hormones

Types of acne

Acne vulgaris
Acne vulgaris, the common type of acne, occurs during puberty and affects the comedogenic areas of the face, back, and chest. There may be a familial tendency to acne. Acne vulgaris is slightly more common in boys, 30–40% of whom have acne between the ages of 18 and 19. In girls the peak incidence is between 16 and 18 years. Adult acne is a variant affecting 1% of men and 5% of women aged 40. Acne keloidalis is a type of scarring acne seen on the neck in men.

Patients with acne often complain of excessive greasiness of the skin, with "blackheads", "pimples", or "plukes" developing. These may be associated with inflammatory papules and pustules developing into larger cysts and nodules. Resolving lesions leave inflammatory macules and scarring. Scars may be atrophic, sometimes with "ice pick" lesions or keloid formation. Keloids consist of hypertrophic scar tissue and occur predominantly on the neck, upper back, and shoulders and over the sternum.

Acne excoreé
The changes of acne are often minimal but the patient, often a young girl, picks at the skin producing disfiguring erosions. It is often very difficult to help the patient break this habit.

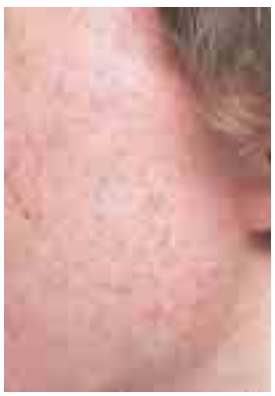
Infantile acne
Localised acne lesions occur on the face in the first few months of life. They clear spontaneously but may last for some years. There is said to be an associated increased tendency to severe adolescent acne.

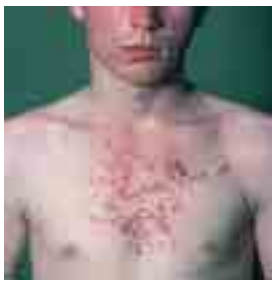
Acne conglobata
This is a severe form of acne, more common in boys and in tropical climates. It is extensive, affecting the trunk, face, and limbs. In "acne fulminans" there is associated systemic illness with malaise, fever, and joint pains. It appears to be associated with a hypersensitivity to *P. acnes*. Another variant is pyoderma faciale, which produces erythematous and necrotic lesions and occurs mainly in adult women.

Gram negative folliculitis occurs with a proliferation of organisms such as klebsiella, proteus, pseudomonas, and *Escherichia coli*.


Acne keloidalis


Acne with comedones


"Ice-pick" scars


Acne vulgaris


Acne fulminans


Pyoderma faciale


Acne conglobata


Gram negative folliculitis

Occupational
Acne-like lesions occur as a result of long term contact with oils or tar as mentioned above. This usually results from lubricating, cutting, or crude oil soaking through clothing. In chloracne there are prominent comedones on the face and neck. It is caused by exposure to polychlortriphenyl and related compounds and also to weedkiller and dicophane.

Treatment of acne

In most adolescents acne clears spontaneously with minimal scarring. Reassurance and explanation along the following lines helps greatly:

(1) The lesions can be expected to clear in time.
(2) It is not infectious.
(3) The less patients are self conscious and worry about their appearance the less other people will take any notice of their acne.

It helps to give a simple regimen to follow, enabling patients to take some positive steps to clear their skin and also an alternative to picking their spots.

Patients with acne should be advised to hold a hot wet flannel on the face (a much simpler alternative to the commercial "Facial saunas"), followed by gentle rubbing in of a plain soap. Savlon solution, diluted 10 times with water, is an excellent alternative for controlling greasy skin. There are many proprietary preparations, most of which act as keratolytics, dissolving the keratin plug of the comedone. They can also cause considerable dryness and scaling of the skin.

Benzoyl peroxide in concentrations of 1–10% is available as lotions, creams, gels, and washes. Resorcinol, sulphur, and salicylic acid preparations are also available.

Vitamin A acid as a cream or gel is helpful in some patients. A topical tretinoin gel has recently been introduced.

Ultraviolet light therapy is less effective than natural sunlight but is helpful for extensive acne. It is a helpful additional treatment in the winter months.

Oral treatment. The mainstay of treatment is oxytetracycline, which should be given for a week at 1 g daily then 500 mg (250 mg twice daily) on an empty stomach. Minocycline or doxycycline are alternatives that can be taken with food. Perseverance with treatment is important, and it may take some months to produce an appreciable improvement. Erythromycin is an alternative to tetracycline, and co-trimoxazole can be used for Gram negative folliculitis. Tetracycline might theoretically interfere with the absorption of progesterone types of birth control pill and should not be given in pregnancy.

Topical antibiotics. Erythromycin, the tetracyclines, and clindamycin have been used topically. There is the risk of producing colonies of resistant organisms.

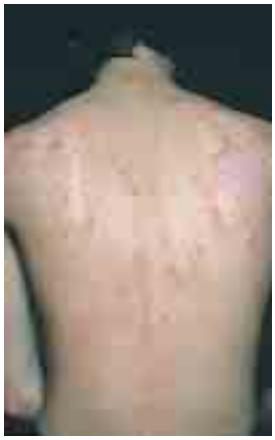
Antiandrogens. Cyproterone acetate combined with ethinyloestradiol is effective in some women; it is also a contraceptive.

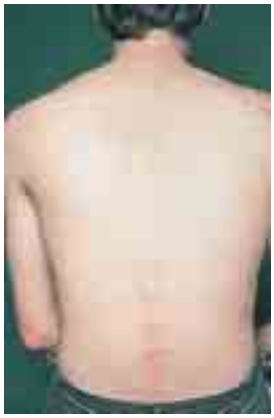
Synthetic retinoids. For severe cases resistant to other treatments these drugs, which can be prescribed only in hospital, are very effective and clear most cases in a few months. 13-cis-Retinoic acid (isotretinoin) is usually used for acne. They are teratogenic, so there must be no question of pregnancy, and can cause liver changes with raised serum lipid values. Regular blood tests are therefore essential. A three month course of treatment usually gives a long remission. Recently topical isotretinoin gel has been introduced.

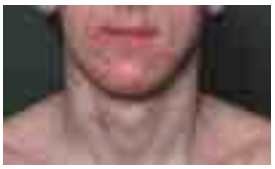
Residual lesions, keloid scars, cysts, and persistent nodules can be treated by injection with triamcinolone or freezing with

Treatment of acne

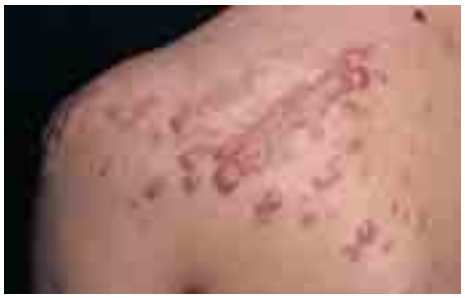
First line	Second line	Third line
Encourage positive attitudes	Topical vitamin A acid	Oral retinoids for 3–4 months (hospitals only)
Avoid environmental and occupational factors	Topical antibiotics	
Topical treatment Benzoyl peroxide Salicylic acid	Ultraviolet light	
A tetracycline by mouth for several months	Antiandrogens	

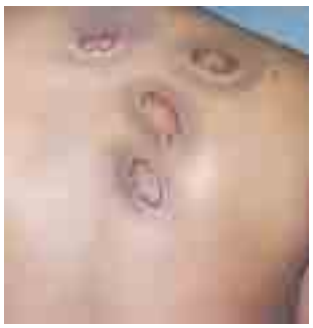

Acne before treatment


Acne after treatment



Severe cystic acne before (left) and after (right) treatment with tetracycline


Keloid scars


Keloid scars on dark skin

liquid nitrogen. For severe scarring dermabrasion can produce good cosmetic results. This is usually carried out in a plastic surgery unit.

Rosacea

Rosacea is a persistent eruption occurring on the forehead and cheeks. It is more common in women than men.

There is erythema with prominent blood vessels. Pustules, papules, and oedema occur. Rhinophyma, with thickened erythematous skin of the nose and enlarged follicles, is a variant. Conjunctivitis and blepharitis may be associated. It is usually made worse by sunlight.

Rosacea should be distinguished from:

- Acne, in which there are blackheads, a wider distribution, and improvement with sunlight. Acne, however, may coexist with rosacea—hence the older term "acne rosacea".
- Seborrhoeic eczema, in which there are no pustules and eczematous changes are present.
- Lupus erythematosus, which shows light sensitivity, erythema, and scarring but no pustules.
- Perioral dermatitis, which occurs in women with pustules and erythema around the mouth and on the chin. There is usually a premenstrual exacerbation. Treatment is with oral tetracyclines.

Treatment

The treatment of rosacea is with long term courses of oxytetracycline, which may need to be repeated. Topical treatment along the lines of that for acne is also helpful. Topical steroids should not be used as they have minimal effect and cause a severe rebound erythema, which is difficult to clear. Avoiding hot and spicy foods may help.

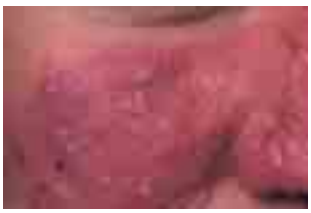
Recent reports indicate that synthetic retinoids are also effective.

Lupus erythematosus

Remember the following points

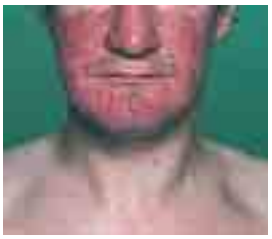
- Avoid topical steroids
- Persevere with one antibiotic not short courses of different types
- Do not prescribe a tetracycline for children and pregnant women
- Oxytetracycline must be taken on an empty stomach half an hour before meals

Blepharitis



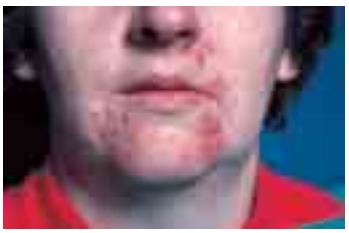
Rosacea

Rosacea



Rosacea

Rhinophyma



Perioral dermatitis on treatment with cyproterone acetate

Further reading

Cunliffe WJ, Cotterill JA. *The acnes: clinical features, pathogenesis and treatment.* St Louis: Mosby, 1989.

Plewig G, Kligman AM. *Acne and rosacea*, 2nd ed. Berlin: Springer-Verlag, 1992

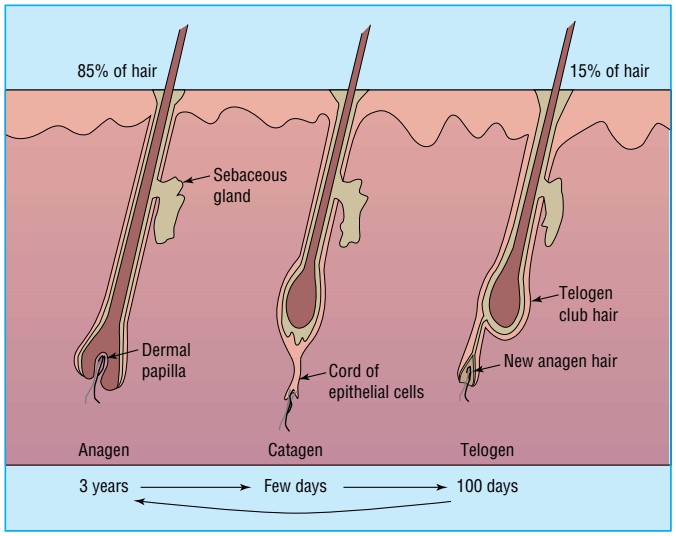
11 The hair and scalp

D Kemmett

Introduction

Hair, which has an essential physiological role in animals, is mainly of psychological importance in man. A good head of hair provides some degree of warmth for the human head and also protection from ultraviolet radiation, but its significance is otherwise in the eye of the beholder. In the form of wool, hair is of economic importance and considerable research has been carried out into the cycles of growth and the structure of wool keratin in sheep.

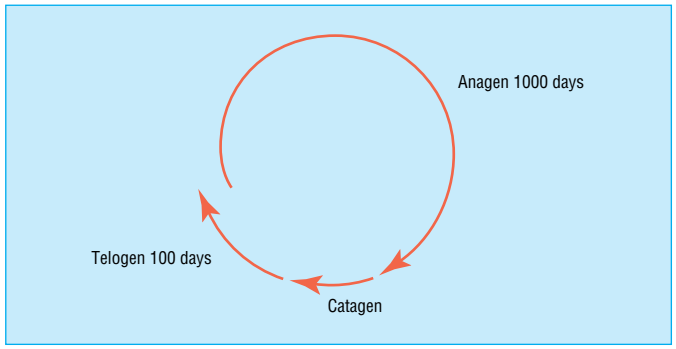
Too much hair, particularly on the face of women, is an embarrassment and cosmetic problem and loss of hair from the scalp is equally troublesome. Changes in hair growth are not only of cosmetic significance but can also be associated with underlying diseases. Diseases occurring in the skin of the scalp can damage hair follicles leading to loss of hair. This chapter covers: (a) the normal pattern of hair growth; (b) causes of hair loss; (c) skin diseases involving the scalp; (d) causes of excess hair growth; (e) abnormalities of the hair itself; and (f) treatment.



Diagrammatic cross-section of hair at various growth phases

The normal pattern of hair growth

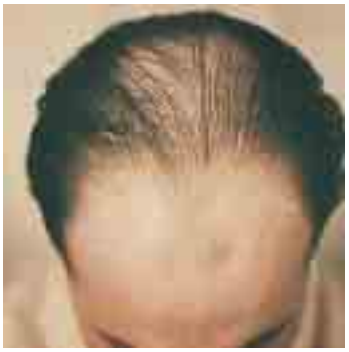
Unlike other epidermal structures which grow continuously, hair has a cyclical pattern of growth. The growing phase or anagen lasts an average of 1000 days on the scalp followed by an involutional phase known as catagen which is quite short, lasting only a few days. The hair then enters a resting phase, telogen, lasting about 100 days. In man, hair growth is normally asynchronous, with each individual hair following its own cycle independently of the others. The basal layer of the hair bulb from which the hair itself is produced is known as the matrix and contains melanocytes from which melanin pigment is incorporated into the hair. The type of melanin determines the colour and in grey or white hair, pigment production is reduced or absent.



Hair growth cycle

The body surface with the exception of the palms, soles, the lips, and the genitalia, is covered with fine vellus hairs that do not have a medulla and are not pigmented. These hairs develop into longer coarse, medullated, terminal hair on the scalp and eyebrows. At puberty a similar change occurs in the pubic area and the axillae, also on the face and trunk, in the male. These changes are androgen dependent, even in females, but testicular androgen is required to produce beard growth and balding in men.

Racial characteristics and the genetic make-up of the individual determine the type and colour of the hair. Straight black oriental hair is clearly different from the nordic blonde type.



Male pattern baldness

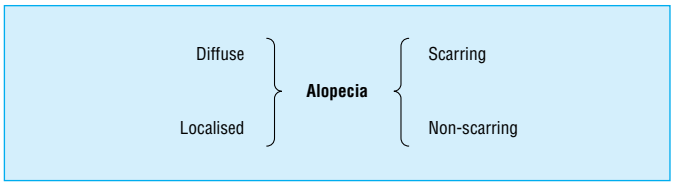
Hair loss

This is known as alopecia, said to be derived from the Latin "alopex", a fox, presumably because of the bald patches of mange seen in wild foxes.

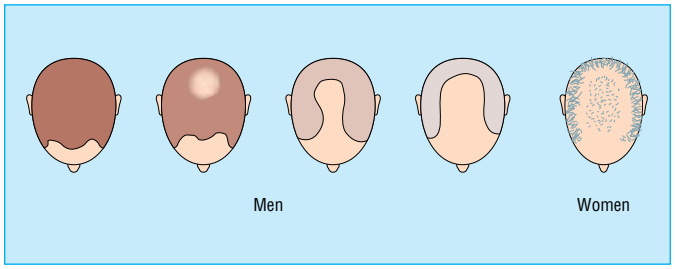
Adult male pattern alopecia is so common as to be considered normal. Circulating levels of testosterone are not

raised in bald men but there is evidence that availability of the hormone to the hair follicle is increased. In postmenopausal women there may be widespread thinning of the hair but loss of hair at the temples often occurs to some degree at an earlier age.

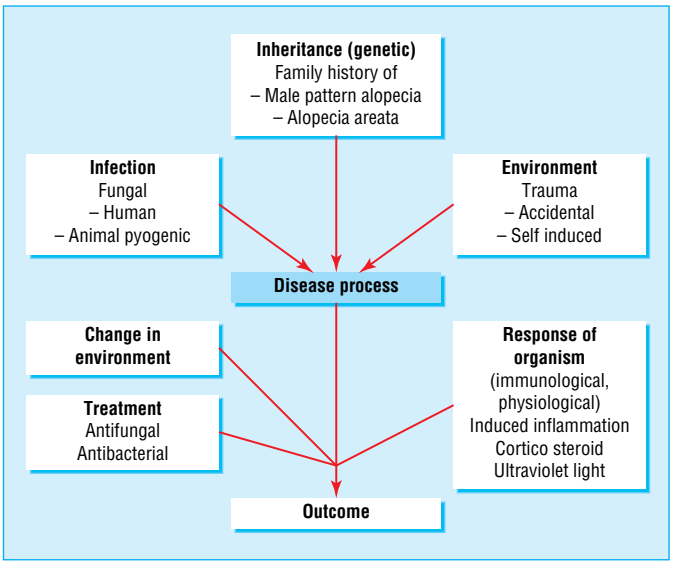
Alopecia may be diffuse or localised. If it is simply due to a physiological derangement of hair growth, the follicles remain intact, whereas inflammation may lead to scarring and loss of the hair follicles. Hence, hair loss can be classified into the categories shown in the illustration on the right.



Classification of alopecia



Adult pattern of alopecia: comparison between men and women



Factors leading to development of alopecia

Diffuse hair loss

An interruption of the normal hair cycle leads to generalised hair loss. This may be due to changes in circulating hormones, drugs, inflammatory skin disease, and "stress" of various types.

Telogen effluvium occurs if all the hairs enter into the resting phase together, most commonly after childbirth or severe illness. Two or three months later the new anagen hair displaces the resting telogen hair, resulting in a disconcerting, but temporary, hair loss from the scalp. Stress of any type, such as an acute illness or an operation, causes a similar type of hair loss.

Postfebrile alopecia occurs when a fever exceeds 39 °C, particularly with recurrent episodes. It has been reported in a wide range of infectious diseases, including glandular fever, influenza, malaria, and brucellosis. It also occurs in fever associated with inflammatory bowel disease.

Dietary factors such as iron deficiency and hypoproteinaemia may play a role, but are rarely the sole cause of diffuse alopecia.

Severe malnutrition with a protein deficiency results in dystrophic changes with a reduction in the rate of hair growth.

Congenital alopecia may occur in some hereditary syndromes.

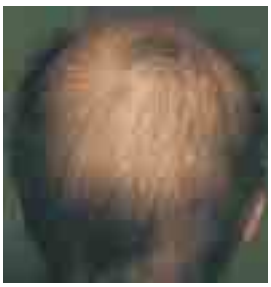
Anagen effluvium occurs when the normal development of hair and follicle is interfered with, resulting in inadequate growth. As a result, hairs are shed earlier than usual, while still in the anagen phase.

Endocrine causes of diffuse alopecia include both hypo- and hyperthyroidism, hypopituitarism, and diabetes mellitus. In hypothyroidism the hair is thinned and brittle, whereas in hypopituitarism the hair is finer and soft but does not grow adequately.

Systemic drugs—cytotoxic agents, anticoagulants, immunosuppressants, and some antithyroid drugs—may cause diffuse hair loss, usually an "anagen effluvium" as mentioned above.

Inflammatory skin disease, when widespread, can be associated with hair loss, for example in erythroderma due to psoriasis or severe eczema.

Deficiency states are a rare cause of alopecia. Patients who suffer from hair loss are often convinced that there is some deficiency in their diet and may sometimes produce the results of an "analysis" of their hairs which show deficiencies in specific trace elements. In fact it is very difficult to cause actual hair loss even in gross malnutrition and in those dying from starvation in refugee camps, the hair growth in the scalp is usually



Anagen effluvium

Diffuse alopecia caused by ciclosporin

Diffuse alopecia caused by erythrodermic psoriasis

present. In chronic malnutrition or kwashiorkor, the hair assumes a curious red/brown colour which may be due to iron deficiency.

Treatment
Wherever possible, the cause should be treated. This may be a matter of replacement therapy in hormonal deficiency. In alopecia due to stress, once this cause is removed hair growth may revert to normal. Treatment of inflammatory skin disease will result in some improvement of the hair loss.

Androgenic alopecia in men is best accepted, with assurance that it indicates normal virility.

Minoxidil causes hair growth and is commercially available as a lotion. This has to be applied continuously every day as the scalp reverts to a level of loss which would have occurred without treatment as soon as it is stopped. It is effective in about half the patients with male pattern alopecia.

Localised alopecia

Alopecia areata is a common form of hair loss. It is seen in 2% of patients attending the average dermatology clinic in the United Kingdom. There may be small patches of hair loss or the whole scalp may be affected. Resolution occurs in a few months or the condition may persist for years. There may be slight inflammation of the skin in the affected areas—in keeping with the possibility of an underlying immune reaction against the hair follicles. There is also an association with autoimmue disease and atopy.

In the affected areas the follicles are visible and empty. The hairs about to be lost have an "exclamation mark" appearance and in some areas that are resolving, fine vellus hairs are seen. Patches commonly occur on the scalp, face, or eyebrows. In *alopecia totalis*, the whole head is involved, and in *alopecia universalis* hair is lost from the whole of the body.

In many patients, particularly if it is a first episode, regrowth occurs within a few months with fine pale hairs appearing first, being replaced by normal adult hair. In older patients, non-pigmented hair may persist in previous patches of alopecia. Factors associated with a poor prognosis are:

(1) Repeated episodes of alopecia
(2) Very extensive or complete hair loss (alopecia totalis)
(3) Early onset before puberty
(4) In association with atopy

Differential diagnosis includes trauma from the habit of plucking hair (trichotillomania) in mentally disturbed patients and traction alopecia from tight hair rollers or hair styles that involve tension on the hair. In fungal infections (tinea capitis) there is scaling and broken hairs. Fungal spores or hyphae are visible in hair specimens on microscopy.

Inflammation is present with loss of hair follicles in lupus erythematosus and lichen planus.

Treatment
An initial limited area of alopecia areata in adult life can be expected to regrow and treatment is generally not needed. Treatments that are carried out include:

(1) Injection of triamcinolone diluted with local anaesthetic which usually stimulates localised regrowth of hair. Unfortunately it often falls out again and there is a risk of causing atrophy. Topical steroid lotion can be used but results are variable.

Causes of diffuse non-scarring alopecia

Androgenetic alopecia
- Male pattern
- Female pattern

Endocrine-thyroid disease (hypothyroidism and hyperthyroidism)
- Hypopituitarism
- Diabetes mellitus

Stress (telogen effluvium)
- Postpartum
- Postoperative
- Postfebrile

Drugs (anagen effluvium)
- Cytotoxics
- Anticoagulants
- Antithyroid agents
- Ciclosporin

Erythrodermic skin disease
- Psoriasis
- Eczema
- Inflammatory

Deficiency states
- Protein malnutrition
- Iron deficiency


Alopecia areata


Alopecia areata, showing exclamation mark hairs


Alopecia totalis


Trichotillomania


Traction alopecia

(2) Ultraviolet light or psoralen with ultraviolet A can give good, if transient, results in a few patients but it has little effect in the majority. It may act by suppressing an immune reaction around the hair root.
(3) Induced contact dermatitis and irritants are occasionally effective. Cantharadin and dithranol have been used for many years as irritants. Primula leaves or chemicals (for example, diphencyprone) can be applied to produce an acute contact dermatitis. The mechanism by which acute inflammation stimulates hair growth is not understood.

Scarring alopecia

The absence of hair follicles is an important physical sign as it indicates:

(1) The presence of an inflammatory process that requires further investigation.
(2) That there is unlikely to be any substantial recovery of hair growth.

The presence of inflammation does not necessarily produce marked erythema—in lichen planus and lupus erythematosus, the inflammatory changes are often chronic. Systemic lupus erythematosus produces areas of inflammation that extend, leaving residual scarring. In discoid lupus erythematosus there is more scaling with keratotic plugs in the follicle. Localised scleroderma (morphoea) also causes alopecia, often with a linear atrophic lesion—the *en coup de sabre* pattern.

More acute inflammatory changes are seen as a result of pyogenic infection or kerion in which there is a marked inflammatory reaction to fungal infection from cattle. In "folliculitis decalvans" there is florid folliculitis with deep seated pustules and scarring. Treatment is with prolonged antibiotics.

Tinea capitis can be associated with alopecia.

Trauma can also cause scarring with alopecia.

Aetiological factors in alopecia areata

Genetic
- Familial in about 20% of cases
- Associated with Down's syndrome

Immunological
- T lymphocytic infiltrate around hair follicles
- Associated with autoimmune disease

Stress
- May be associated in individual patients

Causes of scarring alopecia

Trauma
- For example, burns

Inflammation

Acute
- bacterial (pyogenic infection, syphilis)
- viral (herpes simplex, herpes zoster, varicella)
- fungal (keroin caused by animal ringworm)

Chronic
- lupus erythematosus
- lichen planus
- folliculitis decalvans
- morphoea

Rare
- pyoderma gangrenosum
- necrobiosis lipoidica
- sarcoidosis

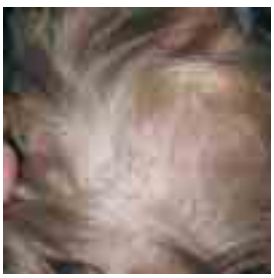

Scarring in lupus erythematosus

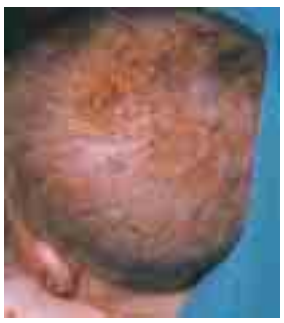

En coup de sabre pattern in alopecia

Skin disease involving the scalp

The scalp can be involved in any skin disease, but most commonly in psoriasis and seborrhoeic eczema. A mild degree of scaling from accumulation in skin scales is so common as to be normal (dandruff). Increased accumulation of scales is seen in seborrhoeic dermatitis in which pityrosporum organisms may play a part. Sometimes masses of thick adherent scales develop in *pityriasis amiantacea*, usually due to psoriasis. Eczema and contact dermatitis can also involve the scalp.


Folliculitis decalvans


Tinea capitis


Atopic eczema


Pityriasis amiantacea


Contact eczema, hair dye

Treatment
Scaling and inflammatory changes can be improved with the use of sulphur and salicylic ointment. It is effective but messy and best applied at night. Tar preparations, oil of cade, and coconut oil in various formulations are all effective. Topical steroids can also be used to suppress inflammation.

Hair shaft abnormalities

Congenital abnormalities of the hair shaft itself lead to weak, thin and broken hairs. In some cases there is a characteristic appearance, for example "spun glass" appearance of pili torti with a twisted hair. In monilethrix there are regular nodes in the hair shaft.

There are other abnormalities of the hair shaft which are not associated with increased fragility, such as the Willi hair syndrome, progressive kinking of the hair and uncombable hair, in which the hair grows in disorderly profusion, completely resistant to combing and brushing. In pili annulati there may be a spangled appearance due to bright bands in the hair shaft.

Excessive hair

Two types of overgrowth of hair occur:

Hirsuties
This is the growth of coarse terminal hair in a male distribution occurring in a woman. This is not always easy to assess since what is unacceptable and emotionally disturbing to one woman may be quite acceptable to another. The amount of hair growth may appear to be within normal limits in a patient complaining of excessive hair but it should never be dismissed as of no consequence if it is important to the woman concerned. Nevertheless, it is important to limit the number of investigations once it is clear that there is no underlying abnormality. Strong reassurance may then be the most helpful management. It is of course most apparent on the face but is often present on the thighs, abdomen and back as well.

Hirsuties occurs most commonly after the menopause and may be present to some degree in normal women as a result of familial or racial traits. It may arise without any underlying hormonal disorder or as a result of virilising hormones. These causes are listed in the box on the right. In addition to androgens, a number of drugs can cause hirsuties. It is important to remember that hirsuties may be part of a virilising syndrome or polycystic ovaries. It is useful to measure the serum testosterone and oestrogen level, as well as urinary 17 oxosteroid concentrations.

Hypertrichosis
This is an excessive growth of hair which may be generalised or localised. It may be due to metabolic disturbance or drugs, or simply a feature of a localised lesion such as a mole.

Hypertrichosis caused by minoxidil

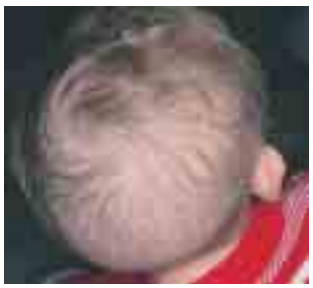
Cutaneous diseases of the scalp

Common
Inflammatory
- Psoriasis
- Seborrhoeic eczema
- Contact dermatitis

Infection
- Folliculitis—staphylococcal or streptococcal
- Fungal infection
  - Microsporum, with hair loss
  - Trichophyton, with scaling and inflammation
- Herpes—zoster and simplex

Infestation
- Pediculosis

Less common
- Lupus erythematosus
- Lichen planus
- Pemphigoid and pemphigus



Monilethrix

Monilethrix

Causes of hirsutism

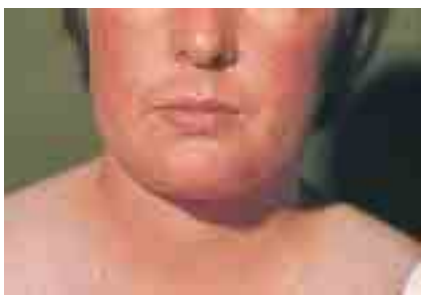
Hereditary, racial with wide variation of normal pattern

Endocrine
Adrenal
- Virilising tumours
- Cushing's syndrome
- Adrenal hyperplasia

Ovarian
- Virilising tumours
- Polycystic ovary syndrome

Pituitary
- Acromegaly
- Hyperprolactinaemia

Iatrogenic
- Anabolic steroids
- Androgens, corticosteroids
- Danazol, phenytoin
- Psoralens (in psoralen with ultraviolet A)



Hirsuties due to virilising tumour

Treatment

It is clearly important to treat any underlying cause and this will usually improve the condition. Otherwise treatment is symptomatic and includes:

- Removal of hair by shaving or hair removing creams.
- Electrolysis and diathermy, which gives permanent destruction of the hair follicle.
- Antiandrogen drugs such as cyproterone can be used under specialist supervision.

Causes of hypertrichosis

- Naevoid—in pigmented naevi and Becker's naevus
- Congenital (rare)
- Acquired—porphyria, hyperthyroidism, anorexia nervosa some developmental defects, for example, Hurler's syndrome, tumours (hypertrichosis lanuginosa), drugs (diazoxide, minoxidil, ciclosporin)

Further reading

Orfanos CE, Happle R. *Hair and hair disease*. Berlin: Springer-Verlag, 1990

Rook A, Dawber R. *Diseases of the hair and scalp*, 3rd ed. Oxford: Blackwell Scientific, 1997

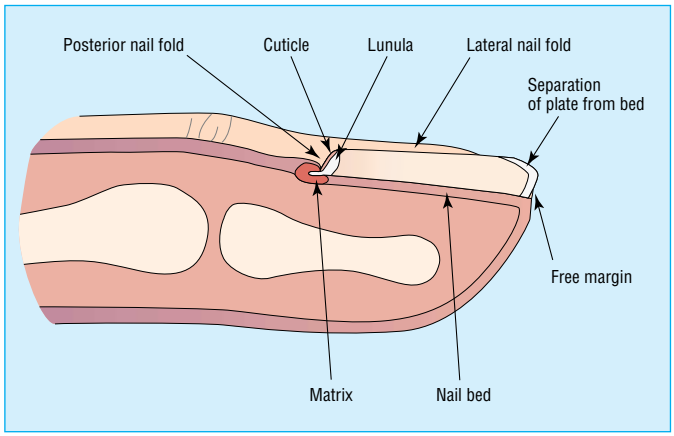
12 Diseases of the nails

AL Wright

In animals and birds claws are used for digging and grasping as well as for fighting. The human nail may be used as a weapon but its main function is to protect the distal soft tissues of the fingers and toes.

As an ectodermal derivative composed of keratin, the nail plate grows forward from a fold of epidermis over the nail bed which is continuous with the matrix proximally. The keratin composing the nail is derived mainly from the matrix with contributions from the dorsal surface of the nail fold, and the nail bed also adds keratin to the deep surface.

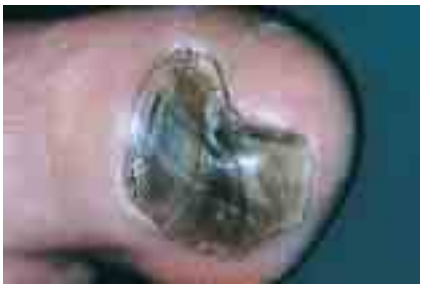
The nail grows slowly for the first day after birth then more rapidly until it slows in old age. The rate of nail growth is greater in the fingers than the toes, particularly on the dominant hand. It is slower in women but increases during pregnancy. Finger nails grow at approximately 0·8 mm per week and toe nails 0·25 mm per week.



Section through finger showing nail structure

Physical signs in the nails

The changes in the nail may be due to a local disease process, a manifestation of a skin disease, or a systemic disorder. Hereditary disorders may also affect the nails. It is therefore important to take both a general history and specifically enquire about skin diseases. It sometimes happens that the nail changes are the only sign of a dermatological disease, although the patient may have a previous history of lichen planus or psoriasis, for example. Localised infection or trauma will affect one or two nails. Skin disease, such as psoriasis, affects many or all nails, usually symmetrically, whereas systemic illness or drugs will affect all the nails.



Effect of trauma

Local nail changes

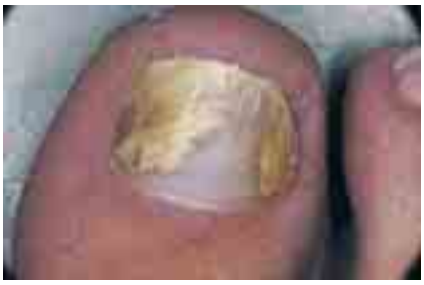
Trauma

- Acute trauma. This may remove the whole nail.
- Chronic trauma as a result of badly fitting footwear may cause thickening of the nail with deformed growth, onychogryphosis. Chronic trauma due to overenthusiastic manicuring or habitual picking at finger nails can result in deformity and impaired growth.
- Repeated trauma in occupations that involve repetitive action, such as assembling cardboard boxes, may cause detachment of the nail (*onycholysis*) or splitting of the nails.

Paronychia

Infection

- Infection of the tissues around the nail (*paronychia*) is often mixed with pyogenic organisms, including pseudomonas, as well as yeasts such as candida. This condition occurs most frequently in those employed in the food industry and occupations where there is repeated exposure to a moist environment and minor trauma. The index and middle fingers are most frequently involved.
- Infection of the nail plate itself occurs in fungal infections, which are commonly due to *Trichophyton* or *Epidermophyton* species.



Fungal infection

- Those living in the tropics may acquire infection with *Scopulariopsis brevicaulis*, which produces a black discoloration of the nail.

Skin diseases affecting the nails

Since the nail plate consists of specialised keratin produced by basal cells, it is not surprising that it is affected by skin diseases. Some conditions, such as psoriasis, may produce characteristic changes whereas in other conditions, such as eczema, the changes are much less specific.

Psoriasis causes an accumulation of keratin, as in lesions of the skin. This may result in the nail being both thickened and raised from the nail bed (*onycholysis*). There may be the changes of pustular psoriasis in the surrounding tissues, indistinguishable from acrodermatitis pustulosa. Loss of minute plugs of abnormal keratin results in "pitting".

Lichen planus produces atrophy of the nail plate which may completely disappear. The cuticle may be thickened and grow over the nail plate, known as *pterygium formation*.

Eczema may be associated with brittle nails that tend to split. Thickening and deformity of the nail occurs in eczema or contact dermatitis, sometimes with horizontal ridging.

Darier's disease results in dystrophy of the nail and longitudinal streaks which end in triangular-shaped nicks at the free edge. On the skin there may be the characteristic brownish scaling papules on the central part of the back, chest, and neck. These are made worse by sun exposure.

Alopecia areata is quite often associated with changes in the nails including ridges, leuconychia, and friable nails. It may be associated with "20 nail dystrophy".

Autoimmune conditions such as pemphigus and pemphigoid may be associated with a variety of changes including ridging, splitting of the nail plate, and atrophy in some or all of the nails.

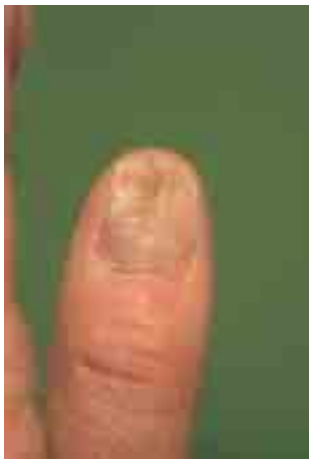
Discoloration of the nail and friability are associated with *lupus erythematosus*.

Pitting of nail

Lichen planus

Dystrophy due to lupus erythematosus

Pterygium formation due to lichen planus



Nail dystrophy

Beau's line

General diseases affecting the nails

Nail changes in systemic illness

Acute illness results in a transverse line of atrophy known as a Beau's line. Shedding of the nail, *onychomedesis*, may occur in severe illness.

Chronic diseases

Clubbing affects the soft tissues of the terminal phalanx with swelling and an increase in the angle between the nail plate and the nail fold. It is associated with chronic respiratory disease, cyanotic heart disease, and occasionally in chronic gastrointestinal conditions. It is occasionally hereditary and may be unilateral in association with vascular abnormalities.

Colour changes

All the nails may be *white* (leukonychia) due to hypoalbuminaemia in conditions such as cirrhosis of the liver. Brown discoloration is seen in renal failure and the "*yellow nail syndrome*", may be associated with abnormalities of the lymphatic drainage. The nail may have a yellow colour in jaundice. *Drugs* may cause changes in colour, for example tetracycline may produce yellow nails, antimalarials a blue

Clubbing

Leukonychia

Yellow nail syndrome

discolouration, and chlorpromazine a brown colour. Leuconychia or whiteness of the nails occurs in fungal infections. Small white spots on the nail are quite commonly seen and are thought to be due to trauma of the nail plate.

Longitudinal pigmented streaks result from increased melanin deposition in the nail plate.

Longitudinal brown streaks are frequently seen in individuals with racially pigmented skin, particularly after trauma. This is rare in caucasians but occurs as a result of a benign pigmented naevus at the base of the nail and in associated lentigo. The most important cause to remember is *subungual melanoma*, which may present with a longitudinal deep brown or black streak. Hutchinson's sign with pigmentation extends into the surrounding tissues, particularly the cuticle. Adrenal disease may rarely be associated with longitudinal streaks.

Pigmented streaks

Malignant melanoma
Normal in pigmented skin
- Melanocytic naevi
- Lentigo
- Addison's disease

Melanocystic naevus

Specific changes in the nail plate

Thickening
This may be due to:

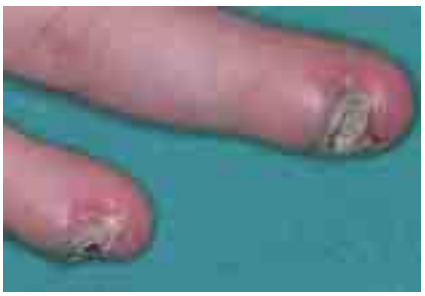
- Hyperkeratosis—psoriasis; fungal infection
- Hypertrophy—chronic trauma (onychogryphosis); pachyonychia congenita
- Atrophy lupus erythematosus—lichen planus; congenital dystrophy.

Thickening of the nail plate may be due to hyperkeratosis in psoriasis, in which case the changes will be symmetrical and there may well also be pitting of the nail and onycholysis. Similar changes are seen in fungal infection of the nail, which may be symmetrical on the toes. Nail clippings should be sent for microscopy and mycological culture.

Hypertrophy of the nail plate occurs as a result of chronic trauma, with only a few nails affected, and is usually seen in the feet.

Pachyonychia congenita is a rare congenital disorder characterised by hypertrophic nails.

Hyperkeratosis, due to the accumulation of keratin under the nail plate, is also seen in psoriasis. It occurs occasionally in association with chronic dermatitis.



Psoriasis

Detachment of the nail plate (onycholysis)
- Psoriasis
- Fungal infection
- Trauma
- Thyrotoxicosis.

Onycholysis is due to a detachment of the nail from the nail bed. If it is extensive, there may be complete loss of the nail plate. It is most commonly seen in psoriasis and occasionally in fungal infections of the nail. It may occur as a result of trauma or thyrotoxicosis

Onycholysis due to psoriasis

Pitting of the surface of the nail plate
- Psoriasis
- Alopecia

Pitting of the nail plate is due to punctate depressions on the surface of the nail plate. They are most often seen in psoriasis but may occur in alopecia areata.

Onycholysis due to psoriasis

Horizontal ridging
- Beau's lines may be seen after systemic illness and acute episodes of hand dermatitis.

Longitudinal ridging
- Single due to pressure from nail fold tumours
- Multiple due to lichen planus
- Alopecia areata
- Psoriasis
- Darier's disease.

Ridging represents a disturbance of nail growth. Inflammation as seen in acute paronychia or trauma can result in a single nail developing a horizontal ridge. After an acute illness there may be horizontal lines on all the nails. The lines may also occur with eczema.

A single longitudinal ridge can result from pressure due to benign or malignant tumours in the nail fold. A mucoid cyst can produce a longitudinal ridge. Multiple longitudinal lines are characteristic of lichen planus, psoriasis, alopecia areata, and Darier's disease.

Koilonychia is a concave deformity of the nail plate, generally occurring in the finger nails. It may be idiopathic or occur as a result of iron deficiency anaemia.

Darier's disease

Longitudinal ridge

Mucoid cyst

Lesions adjacent to the nail

Mucoid cysts develop subcutaneously over the distal interphalangeal joint and may be adjacent to the nail, producing abnormalities of growth. These cysts develop as an extension of the synovial membrane and are linked to the joint by a fine tract. Very careful excision is required for a cure.

Naevi may occur adjacent to the nail and a benign melanocytic naevus can produce a pigmented streak. Subungual melanoma may produce considerable pigmentation of the nail and often causes pigmentation of the cuticle, so called Hutchinson's sign. Sometimes subungual melanoma is amelanotic so there is no pigmentary changes and any rapidly growing soft tumour should raise suspicions of this condition.

Subungual exostosis can cause a painful lesion under the nail. It is confirmed by *x* ray examinations.

Glomus tumours arise as tender nodules.

Periungual fibrokeratomas also develop in patients with tuberous sclerosis.

Big toe exostosis

Nail dystrophy with alopecia areata

Treatment of nail conditions

It is clearly not possible to treat congenital abnormalities of the nail, but avoiding exposure to trauma may help. Nail changes associated with dermatological conditions may improve as the skin elsewhere is treated. Systemic treatment of associated dermatoses will of course tend to improve the nail as well, for example methotrexate or retinoids for psoriasis.

Infective lesions respond to antifungal or antibiotic treatment. In chronic paronychia there is often a mixed infection and a systemic antibiotic combined with topical nystatin may be required. It is also important to keep the hands as dry as possible.

The imidazole antifungal drugs are fairly effective but are fungistatic. Terbinaline is fungicidal and a short course is as effective as prolonged treatment with the older drug griseofulvin.

Nail dystrophy with eczema

Further reading

Baden HP. *Diseases of the hair and nails.* Chicago: Year Book Medical, 1986

Baran R. *Nail disorders: common presenting signs, differential diagnosis and treatment.* Edinburgh: Churchill Livingstone, 1991

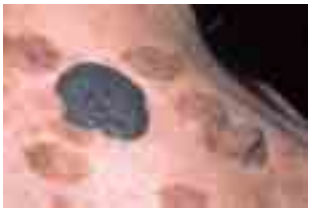
De Berker DA, Baran R, Dawber RP. *Handbook of diseases of the nails and their management.* Oxford: Blackwell Scientific, 1995

13 Lumps and bumps

The skin is a common site for neoplastic lesions, but most invade only locally and with treatment usually do not pose any threat to the life of the patient. The exception is malignant melanoma, which is dealt with in chapter 15. This is a rare tumour with a high mortality, and recent publicity campaigns have been aimed at preventing the tragedy of fatal metastases from a neglected melanoma.

As a result a large number of patients are being seen with pigmented skin lesions and nodules, only a very few of which are neoplastic. The question is how to distinguish the benign, the malignant, and the possibly malignant. The following guidelines may help in deciding whether the lesion can be safely left or should be treated.

A correlation of the clinical and pathological features is helpful in making a confident diagnosis of the more common tumours.



Seborrhoeic wart

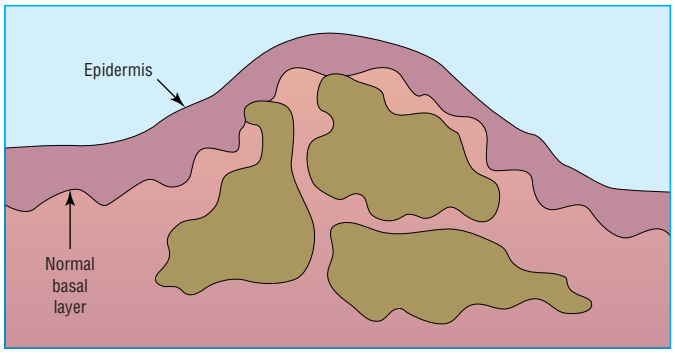
Seborrhoeic wart

Seborrhoeic warts

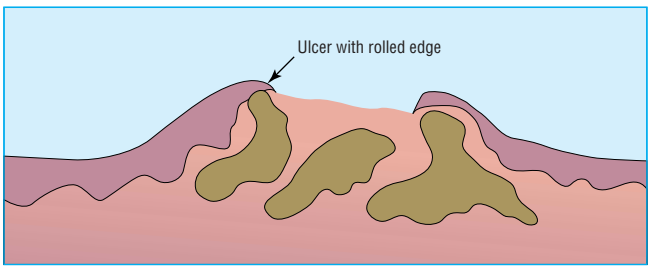
Seborrhoeic warts come in various shapes, sizes, and colours. When deeply pigmented, inflamed, or growing they may appear to have the features of a malignant lesion, but the following features are characteristic:

- Well defined edge.
- Warty, papillary surface—often with keratin plugs.
- Raised above surrounding skin to give a "stuck on" appearance.

Individual lesions vary considerably in size, but are usually 0·5–3·0 cm in diameter. Protuberant and pedunculated lesions occur. Solitary lesions are commonly seen on the face and neck but more numerous, large lesions tend to occur on the trunk. They become more common with increasing age.



Nodular basal cell carcinoma

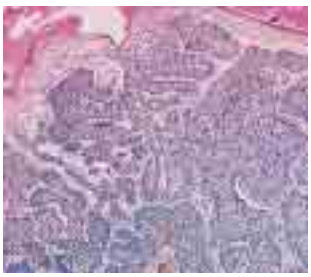


Ulcerated basal cell carcinoma

Basal cell carcinoma

By contrast, the early basal cell carcinoma—or rodent ulcer—presents as a firm nodule, clearly growing within the skin and below it, rather than on the surface. The colour varies from that of normal skin to dark brown or black, but there is commonly a "pearly" translucent quality. As its name implies, the tumour is composed of masses of dividing basal cells that have lost the capacity to differentiate any further. As a result no epidermis is formed over the tumour and the surface breaks down to form an ulcer, the residual edges of the nodule forming the characteristic "rolled edge". Once the basal cells have invaded the deeper tissues the rolled edge disappears.

Clumps of neoplastic basal cells



Neoplastic basal cells—histology

Variants

Variants of the usual pattern can cause problems in diagnosis. Cystic basal cell carcinomas occur, and those that show differentiation towards hair follicles or sweat glands may

Basal cell carcinoma (later stage of figure above)

Cystic basal cell carcinoma

have a less typical appearance. Pigmented lesions can resemble melanoma. The superficial spreading type may be confused with a patch of eczema. This usually occurs on the trunk, does not itch, and shows a gradual but inexorable increase in size. A firm "whipcord" edge may be present. The sclerosing type has scarring of the epidermis associated with basal cell carcinoma.

Treatment
Various methods of destroying tumour tissue are used and the results are similar for radiotherapy and surgical excision:

- Ulcerated lesions may invade tissue planes, blood vessels, and nerves more extensively than is clinically apparent.
- Although modern techniques of radiotherapy result in minimal scarring and atrophy these may cause problems near the eye.
- Basal cell carcinomas in skin creases, such as the nasolabial fold, tend to ulcerate and are hard to excise adequately.
- Surgical excision has the advantage that should the lesion recur, radiotherapy is available to treat it, whereas it is not desirable to treat recurrences after radiotherapy with further irradiation.


Superficial basal cell carcinoma


Pigmented basal cell carcinoma


Sclerosing basal cell carcinoma


Squamous cell carcinoma—initial changes

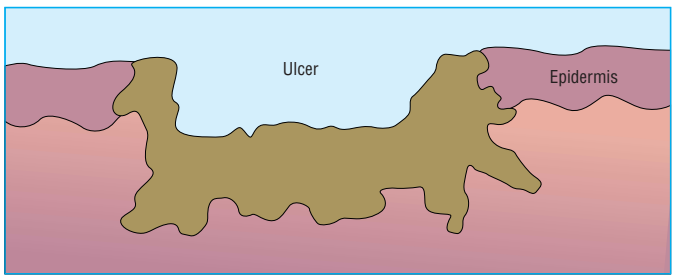

Squamous cell carcinoma—more differentiated tumour

Squamous cell carcinoma

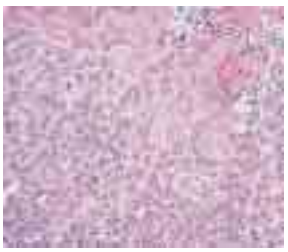
Squamous cell carcinoma represents proliferation of the epidermal keratinocytes in a deranged manner—with a visible degree of differentiation into epidermal cells that may show individual cell keratinisation and "pearls" of keratin. In other tumours bizarre cells with mitoses, cells with clear cytoplasm, or spindle cells may be seen.

This type of cancer often develops from a preceding solar keratosis or an area of Bowen's disease. They may also complicate a chronic ulcer due to stasis, as in venous ulcer of the ankle, or infection such as leprosy or tuberculosis. In addition to local spread, metastases can occur with involvement of other organs such as the liver, lung or brain, and lymphadenopathy. The first change clinically is a thickening of the skin with scaling or hyperkeratosis of the surfaces. The more differentiated tumours often have a warty, keratotic crust whereas others may be nodular. The edge is poorly defined. There may be associated dilated, telangiectatic blood vessels. The original hard, disc-like lesion becomes nodular and ulcerates with strands of tumour cells invading the deeper tissue. The thick warty crust, often found elsewhere, may be absent from the lesions on the lip, buccal mucosa, and penis.

These histological changes complement the clinical appearance and are clearly different from those of basal cell carcinoma.



Squamous neoplastic cells


Squamous cell carcinoma—histological appearance

Treatment
Small lesions should be excised as a rule, making sure that the palpable edge of the tumour is included, with a 3–5 mm margin. Radiotherapy is effective but fragile scars may be a disadvantage on the hand. Cryotherapy or topical fluorouracil can be used for histologically confirmed, superficial lesions and also for solar keratoses.

Solar keratoses

Solar keratoses occur on sites exposed to the sun and are more common in those who have worked out of doors or sunbathed excessively. The common sites are the face, back of the hands,


Solar keratoses

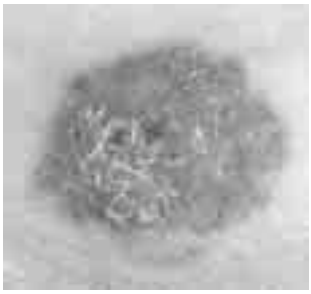
arms, and legs. They also develop on the scalp in bald men and on the lips, particularly in pipe smokers. They show alterations in keratinisation and have the potential to become dysplastic and eventually develop into squamous cell carcinoma, a change often preceded by inflammation. They can be regarded as squamous cell carcinoma grade 1/2.

The clinical appearance varies from a rough area of skin to a raised keratotic lesion. The edge is irregular and they are usually less than 1 cm in diameter. Inflammation and tenderness may be associated with progression to carcinoma.

Treatment

Treatment with cryotherapy, using liquid nitrogen or carbon dioxide, repeated if necessary, is usually effective.

5-Fluorouracil cream is useful for larger or multiple lesions. It is applied twice daily for two weeks, which produces inflammation and necrosis. Simple dressings are applied for the next two weeks. This process can be repeated if necessary. As it is a cytotoxic drug it should be handled with care and applied sparingly with a cotton bud while wearing gloves.



Bowen's disease

Paget's disease of the nipple

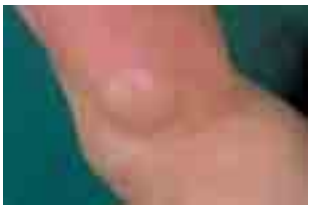
Other conditions

Bowen's disease is characterised by a well defined, erythematous macule with little induration and slight crusting. It is a condition of the middle aged and elderly, occurring commonly on the trunk and limbs. It is an intraepidermal carcinoma, which has been reported to follow the ingestion of arsenic in "tonics" taken in years gone by or exposure to sheep dip, weedkiller, or industrial processes. After many years florid carcinoma may develop with invasion of deeper tissues. It may be confused with a patch of eczema or superficial basal cell carcinoma. Lesions on covered areas may be associated with underlying malignancy. Erythroplasia of Queyrat is a similar process occurring on the glans penis or prepuce.

Paget's disease of the nipple presents with unilateral non-specific erythematous changes on the aureola and nipple, spreading to the surrounding skin. The cause is an underlying adenocarcinoma of the ducts. It should be considered in any patient with eczematous changes of one breast that fail to respond to simple treatment. Extramammary lesions occur.

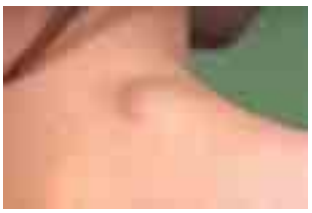
Keratoacanthoma is a rapidly growing fleshy nodule that develops a hard keratotic centre. Healing occurs with some scarring. Although benign, it may recur after being removed with curette and cautery, particularly from the face, and is best excised.

Keratoacanthoma



Dermatofibroma

Skin tags



Lipoma

Benign tumours

Dermatofibroma is a simple, discrete firm nodule, arising in the dermis at the site of an insect bite or other trivial injury. Often there is a brown or red vascular lesion initially, which then becomes fibrotic—a sclerosing haemangioma. The histiocytoma is similar but composed of histiocytes.

Skin tags may be pigmented but rarely cause any diagnostic problems unless inflamed. Some are in fact pedunculated seborrhoeic warts and others simple papillomas (fibroepithelial polyps).

Lipomas are slow growing benign subcutaneous tumours.

Other benign tumours

A wide variety of tumours may develop from the hair follicle and sebaceous, exocrine (sweat), and apocrine glands. The more common include *syringomas*—slowly growing, small, multiple nodules on the face of eccrine gland origin.

Syringoma

Naevus sebaceous is warty, well defined, varying in size from a small nodule to one several centimetres in diameter. Lesions occur in the scalp of children, may be present at birth, and gradually increase in size. They may proliferate or develop into a basal cell carcinoma in adult life and they are therefore best removed.

Verrucous epidermal naevi are probably a variant, found on the trunk and limbs.

Cysts

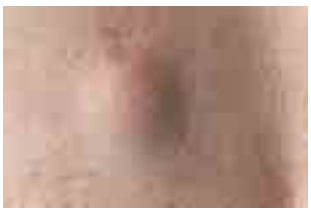
The familiar *epidermoid cyst*—also known as sebaceous cyst or wen—occurs as a soft, well defined, mobile swelling usually on the face, neck, shoulder, or chest. It is not derived from sebaceous glands but contains keratin produced by the lining wall.

Pilar cysts on the scalp are similar lesions derived from hair follicles.

Milia are small keratin cysts consisting of small white papules found on the cheek and eyelids.

Naevus sebaceous

Verrucous epidermal naevus



Epidermoid cyst

Milia

Vascular lesions

The more common vascular naevi are described.

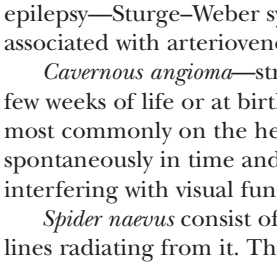
The *port wine stain*, or naevus flammeus, presents at birth as a flat red lesion, usually on the face, neck, or upper trunk. There is usually a sharp midline border on the more common unilateral lesions. In time the affected area becomes raised and thickened because of proliferation of vascular and connective tissue. If the area supplied by the ophthalmic or maxillary divisions of the trigeminal nerve is affected there may be associated angiomas of the underlying meninges with epilepsy—Sturge–Weber syndrome. Lesions of the limb may be associated with arteriovenous fistulae.

Cavernous angioma—strawberry naevi—appear in the first few weeks of life or at birth. A soft vascular swelling is found, most commonly on the head and neck. The lesions resolve spontaneously in time and do not require treatment unless interfering with visual function.

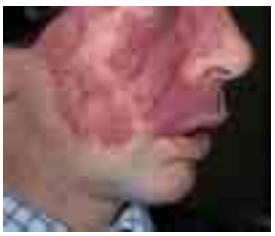
Spider naevus consist of a central vascular papule with fine lines radiating from it. They are more common in children and women. Large numbers in a man raise the possibility of liver disease.

Campbell de Morgan spots are discrete red papules 1–5 mm in diameter. They are more common on the trunk.

Pyogenic granuloma is a lesion that contains no pus but is in fact vascular and grows rapidly. It may arise at the site of trauma. Distinction from amelanotic melanoma is important.



Naevus flammeus



Sturge–Weber syndrome

Cavernous angioma in a five month old baby

Patient five years later without treatment

Spider naevi

Pyogenic granuloma

Further reading

Enzinger F, Weiss S. *Soft tissue tumours*, 2nd ed. St Louis: Mosby, 1988

Mackie RM. *Skin cancer: an illustrated guide to the etiology, clinical features, pathology and management of benign and malignant cutaneous tumours*, 2nd ed. London: Martin Dunitz, 1996

14 The sun and the skin

R StC Barnetson

People with darkly pigmented skin very rarely get skin cancer. Those of a Celtic constitution, when exposed to strong sunlight in countries such as Australia, get skin cancer very readily. Australia has the highest incidence of skin cancer in the world, with 140 000 new cases per year, and 1200 deaths per year, mainly from melanoma.

It is therefore important to understand that there is a variation in skin sensitivity to sunlight. This is rated from one to six (Fitzpatrick classification). Skin type one subjects have red hair and do not tan, burn very easily in the sun and develop skin cancer readily, whereas skin type six subjects have black skin (with an inbuilt sun protection factor of 10) and very rarely develop skin cancer. This is a useful guide in assessing the risk of sun damage and in determining the dose of ultraviolet B in treatment.

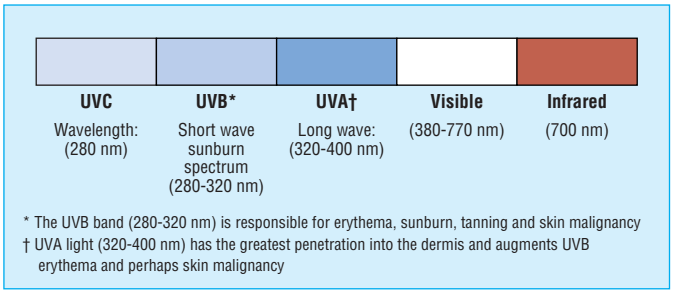
Skin types and sun

- Type 1—Never tans, freckles, red hair, blue eyes
- Type 2—Tans with difficulty, less freckled
- Type 3—Tans easily, dark hair, brown eyes
- Type 4—Always tans, Mediterranean skin
- Type 5—Brown skin (for example, Indian)
- Type 6—Black skin (for example, African)

Aborigines do not get skin cancer

Ultraviolet radiation

There are three types of ultraviolet radiation—the short wavelength ultraviolet C (100–280 nm), ultraviolet B (290–320 nm), and long wavelength ultraviolet A (320–400 nm). Beyond this is visible light then infrared, and radiowaves. ultraviolet C does not penetrate beyond the stratosphere as it is absorbed by the ozone layer. Ultraviolet B is very important in both sunburn and the development of skin cancer. Ultraviolet A is thought to be of increasing importance in the development of skin cancer, and causes tanning but not sunburn. It is also important in people with photosensitivity. The effects of ultraviolet radiation may be classified as short term (sunburn, photosensitivity) or long term (skin cancer, wrinkling, solar elastosis, solar keratoses, seborrhoeic warts).



Light spectrum (UVC=ultraviolet C, UVB=ultraviolet B, UVA=ultraviolet A)

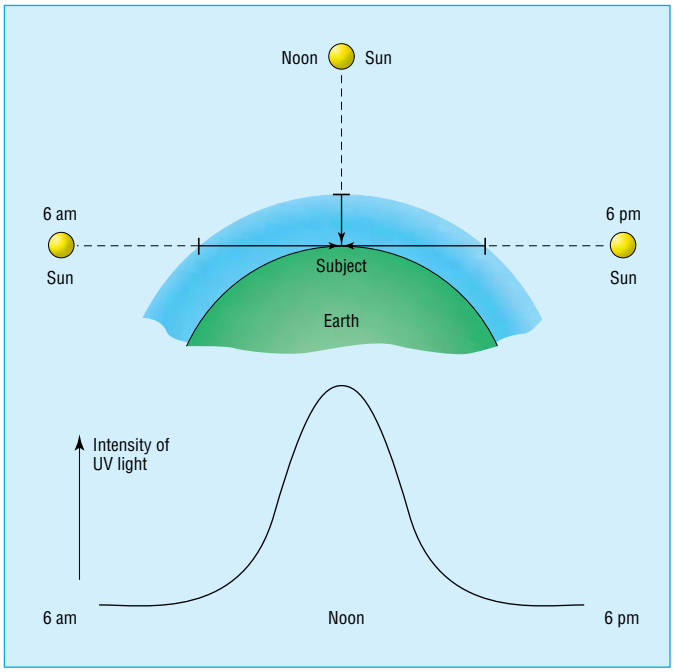
There is general awareness that the sun causes cancer in the skin, with some people becoming obsessively fearful of any exposure to sun. A sensible approach with emphasis on reasonable precautions is called for. Useful points are:

- Most moles are entirely harmless.
- Detecting the changes in moles or early melanoma enables the diagnosis to be made at an early stage with a good chance of curative treatment.
- The non-melanotic, epidermal cancers—basal cell and squamous cell carcinomas—grow slowly and are generally not life threatening. But squamous cell carinoma arising at sites of trauma, on the extremities, or in ulcers may metastasise. Exposure to sun has usually occurred many years previously.

Prevention of sun damage and skin cancer

Prevention of sun damage and skin cancer will depend on reducing exposure to ultraviolet radiation. This can be achieved in a number of ways:

- Covering the skin with clothes. It must be remembered however that light clothes such as shirts or blouses may only have a sun protection factor of four. A wide-brimmed hat is essential to protect the face and neck.
- Sunscreens will greatly reduce sun exposure for exposed parts such as the face and hands. Sunscreens are much more



Diurnal variation in UV intensity of light from sun

efficient than previously, particularly those with a sun protection factor greater than 30; they are now water resistant, and most have a broad spectrum, protecting against ultraviolet B and ultraviolet A. This is important because there is now increasing evidence that ultraviolet A is important in the development of skin cancer.
- Exposure to midday sun, particularly in tropical or subtropical latitudes, should be avoided. At this time of the day the sunlight passes vertically through the atmosphere and there is less filtering of dangerous ultraviolet light. So remember the adage: "Between eleven and three, stay under a tree" in the summer months.

Effects of sun

Short term
- Sunburn
- Photosensitivity

Long term
- Skin wrinkling
- Telangiectasia
- Hyper and hypopigmentation
- Solar elastosis
- Actinic keratosis
- Seborrhoeic warts
- Skin cancers

Development of skin cancers

Sun-damaged skin

A number of different features characterise sun-damaged skin, which is often seen in the elderly particularly if they have lived in a sunny climate such as Australia. The skin has many fine wrinkles and often has a sallow yellowish discoloration particularly on the face and other exposed parts of the body. Hyperpigmentation occurs as result of recent sun exposure, which may be diffuse or localised in the form of solar lentigo. In some areas there may be hypopigmentation, particularly where solar keratoses have been treated with liquid nitrogen (cryotherapy). There may be marked telangiectasia and numerous blood vessels are seen. In some, there may be thickening and a yellow hue of the skin, particularly of the neck, due to elastin deposition in the upper dermis; this is known as solar elastosis.

Sun damaged skin

Solar elastosis

Forms of skin cancer

There are three common forms of skin cancer caused by ultraviolet light: basal cell carcinoma, squamous cell carcinoma, and malignant melanoma. Whereas there seems to be a direct relationship with the amount of ultraviolet exposure and basal cell carcinoma and squamous cell carcinoma, the relationship with ultraviolet exposure and melanoma is more complex and it seems likely that intermittent exposure to ultraviolet light is the main factor (for example, exposure to sunlight on holidays). These different types of neoplastic change that occur in the skin are discussed in chapters 13 and 15.

Photosensitivity

Exposure to sun in non-pigmented races causes inflammation in the skin, depending on the skin type and amount of exposure. In some individuals there is an abnormal sensitivity to sunlight. This may arise because of an idiopathic reaction to sunlight or allergic reaction that is activated by sunlight. Some chemicals seen to induce photosensitivity without causing an allergy. Other causes are metabolic diseases and inflammatory conditions that are made worse by sun exposure.

Causes of photosensitivity
- Idiopathic—for example, polymorphic light eruption, actinic prurigo, solar urticaria
- Photoaggravated dermatoses—for example, lupus erythematosus, eczema
- Metabolic—porphyria—for example, erythropoietic, hepatic
- Drug induced—for example, sulphonamides, phenothiazines
- Chemical induced (topical)—for example, tar, anthracene

Polymorphic light eruption

This is the most common of the idiopathic photosensitive rashes and occurs predominantly in women. It is due to both the shorter (ultraviolet B) and longer (ultraviolet A) wavelength types of sunlight. The eruption occurs from hours to days after exposure and varies in severity from a few inflamed papules to extensive inflamed oedematous lesions. There may be only a few trivial lesions initially, but increasingly severe reactions can develop restricting the patients ability to venture outside. A useful measure of seventy is to ask the

Photosensitivity caused by drugs

patient if they cross to the shady side of the street to avoid the sun. Treatment includes topical or systemic steroids for the acute rash and prevention by using sunscreens. Desensitisation by narrow waveband phototherapy before exposure is effective.

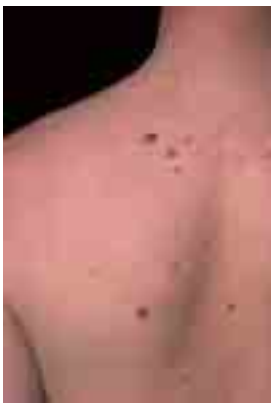
Solar urticaria

This is a much less common condition and may be induced by longer wavelength (ultraviolet A) and visible radiation as well as ultraviolet B. It is characterised by rapidly developing irritation and in the exposed skin is followed by urticarial wheals. It can occur as part of a photoallergic reaction, in which case avoidance of the relevant allergen will prevent the condition.

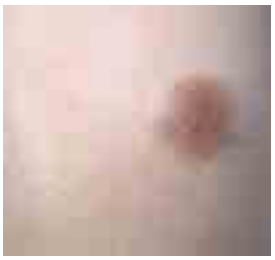
Treatment is with antihistamines and sunscreens. In some cases phototherapy with ultraviolet B, narrow waveband or psoralen with ultraviolet A (PUVA), is helpful.

15 Black spots in the skin

There has been a great increase in public awareness of melanoma, and any dark lesions of the skin are sometimes regarded with the same dread as Long John Silver's "black spot" in *Treasure Island*—a sign of imminent demise. However, the vast majority of pigmented lesions are simply moles or harmless pigmented naevi. The most important thing is to know which moles can be safely ignored and which should be removed. Benign moles are described first, then malignant melanoma, followed by a discussion of the differences between these two.



Benign moles



Benign mole

Benign moles

Benign moles are naevi with a proliferation of melanocytes and a variable number of dermal naevus cells. Some moles are congenital and are present from birth, but most develop in early childhood and adolescence. The number of moles remains constant during adult life with a gradual decrease from the sixth decade onwards.

There is often an increase in both the number of moles and the degree of pigmentation during pregnancy.

Benign pigmented naevus

Acquired melanocytic naevi

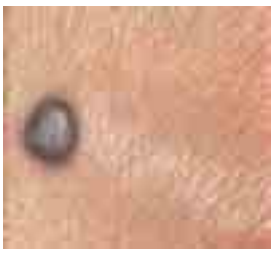
Acquired melanocytic naevi are the familiar moles and present in a number of different ways depending on the type of cells and the depth in the skin.

Junctional naevi are flat macules with melanocytes proliferating along the dermo-epidermal border.

Compound naevi have pigmented naevus cells at the dermo-epidermal border and in the dermis, producing a raised brown lesion. The dermal melanocytes may accumulate around the skin appendages and blood vessels and form a band of cells without melanin or more deeply penetrating strands of spindle cells. Proliferating naevus cells may throw the overlying epidermis into folds, giving a papillary appearance.

In a purely *intradermal naevus* the junctional element is lost, with the deeper cells showing characteristics of neural tissue. Other types of acquired pigmented naevi include the following.

Blue naevus is a collection of deeply pigmented melanocytes situated deep in the dermis, which accounts for the deep slate-blue colour.

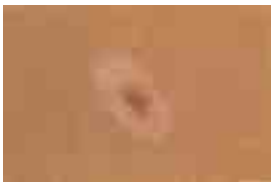


Blue naevus

Spitz naevus

Spitz naevus presents as a fleshy pink papule in children. It is composed of large spindle cells and epitheloid cells with occasional giant cells, arranged in "nests". It is benign and the old name of juvenile melanoma should be abandoned.

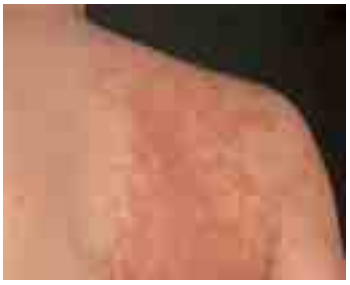
Halo naevus consists of a melanocytic naevus with a surrounding halo of depigmentation associated with the presence of antibodies against melanocytes in some cases. The whole naevus gradually fades in time.



Halo naevus

Becker's naevus is an area of increased pigmentation, often associated with increased hair growth, which is usually seen on the upper trunk or shoulders. It is benign.

Freckles or ephelides are small pigmented macules, less than 0·5 cm in diameter, that occur in areas exposed to the sun in fair skinned people. These macules fade during the winter months.



Becker's naevus

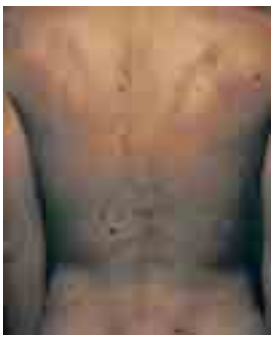
Congenital pigmented naevi

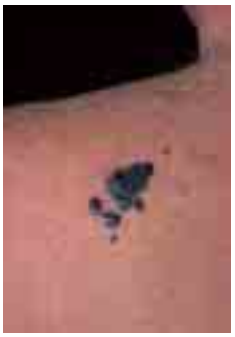
Congenital pigmented naevi are present at birth, generally over 1 cm in diameter, and vary from pale brown to black in colour. They often become hairy and more protuberant, possibly with

an increased risk of malignant change. Larger lesions can cover a considerable area of the trunk and buttocks, such as the bathing trunk naevi, and their removal may present a considerable problem.

Dysplastic naevi
These show very early malignant change and may progress to malignant melanoma. They are deeply pigmented often with an irregular margin.

In *dysplastic naevus syndrome* multiple pigmented naevi that occur predominantly on the trunk, becoming numerous during adolescence. They vary in size—many being over 0·5 cm—and tend to develop into malignant melanoma, particularly if there is a family history of this condition.


Congenital hairy naevus


Dysplastic naevus syndrome

Melanoma

Melanoma is an invasive malignant tumour of melanocytes. Most cases occur in white adults over the age of 30, with a predominance in women.

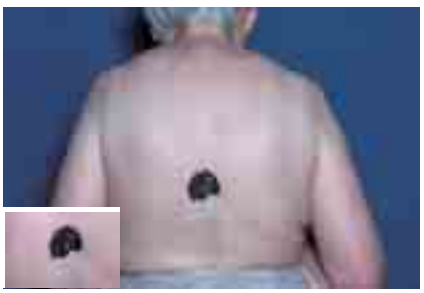
Incidence
The incidence of melanoma has doubled over the past 10 years in Australia (currently 40/100 000 population) and shown a similar increase in other countries. In Europe twice as many women as men develop melanoma—about 12/100 000 women and 6/100 000 men.

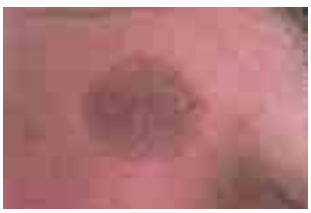
Prognosis
The prognosis is related to the thickness of the lesion, measured histologically in millimetres from the granular layer to the deepest level of invasion. Lesions less than 0·76 mm thick have a 100% survival at five years, 0·76–1·5 mm thick an 80% survival at five years, and lesions over 3·5 mm less than 40% survival. These figures are based on patients in whom the original lesion had been completely excised. A recent study in Scotland has shown an overall five year survival of 71·6–77·6% for women and 58·7% for men.

Sun exposure
The highest incidence of melanoma occurs in countries with the most sunshine throughout the year. However, skin type and the regularity of exposure to sun are also important. The incidence is much greater in fair skinned people from higher latitudes who have concentrated exposure to sun during holidays than in those with darker complexions who have more regular exposure throughout the year. Severe sunburn may also predispose to melanoma.

Genetic factors
Since melanin protects the skin from ultraviolet light it is not surprising that melanoma occurs most commonly in fair skinned people who show little tanning on exposure to sun, particularly those of Celtic origin. Members of families with the dysplastic naevus syndrome are more likely to develop melanoma in their moles. These patients have multiple naevi from a young age.

Pre-existing moles
It is rare for ordinary moles to become malignant but congenital naevi and multiple dysplastic naevi are more likely to develop into malignant melanoma.


Melanoma


Melanoma


Nodular melanoma


Superficial melanoma with nodules


Lentigo maligna


Nodule developing in superficial spreading melanoma

Types of melanoma

There are four main types of melanoma.

Superficial spreading melanoma is the more common variety. It is common on the back in men and on the legs in women. As the name implies the melanoma cells spread superficially in the epidermis, becoming invasive after months or years. The margin and the surface are irregular, with pigmentation varying from brown to black. There may be surrounding inflammation and there is often clearing of the central portion. The invasive phase is associated with the appearance of nodules and increased pigmentation. The prognosis is correspondingly poor.

Lentigo maligna melanoma occurs characteristically in areas exposed to sun in elderly people. Initially there is a slowly growing, irregular pigmented macule that is present for many years before a melanoma develops.

Nodular melanoma presents as a dark nodule from the start without a preceding in situ epidermal phase. It is more common in men than women and is usually seen in people in their fifties and sixties. Because it is a vertical invasive growth phase from the beginning there is a poor prognosis.

Acral melanoma occurs on the palm and soles and near or under the nails. Benign pigmented naevi may also occur in these sites and it is important to recognise early dysplastic change by using the criteria set out below. A very important indication that discoloration of the nail is due to melanoma is Hutchinson's sign—pigmentation of the nail fold adjacent to the nail. It is important to distinguish talon noir, in which a black area appears on the sole or heel. It is the result of trauma—for example sustained while playing squash—causing haemorrhage into the dermal papillae. Paring the skin gently with a scalpel will reveal distinct blood filled papillae, to the relief of doctor and patient alike.

Other types of melanoma

As the melanoma cells become more dysplastic and less well differentiated they lose the capacity to produce melanin and form an amelanonitic melanoma. Such non-pigmented nodules may be regarded as harmless but are in fact extremely dangerous.

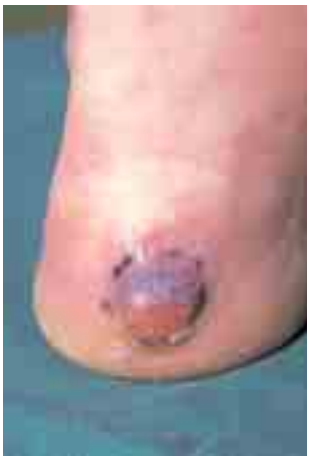
Superficial spreading melanoma

Nodular melanoma in a lentigo

Benign lentigo

Acral melanoma

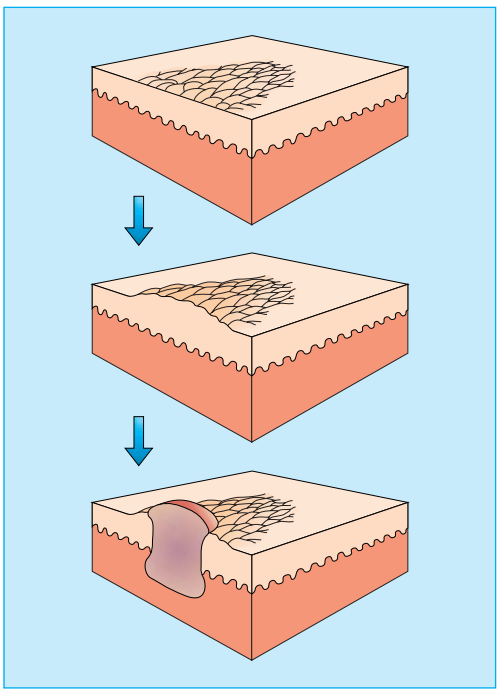
Talon noir of left heel



Dysplastic melanoma

Amelanotic melanoma

Malignant melanoma in a black person; note the surrounding "halo"



Progressive growth in depth of malignant melanoma

Prognosis

This depends on the depth to which the melanoma has penetrated below the base of the epidermis—lesions confined to the epidermis having better prognosis than those penetrating into the dermis. The Clark classification describes the depth of penetration as follows:

Level I—within the epidermis
Level II—few melanoma cells within the dermal papillae
Level III—many melanoma cells in the papillary dermis
Level IV—invasion of the reticular dermis
Level V—invasion of the subcutaneous tissues

The Breslow classification is based on measurements of tumour thickness from the granular layer overlying epidermis. A depth of less than 1·5 mm is associated with a 90% five year survival, 1·5–3·5 mm with a 75% five year survival, and greater than 3·5 mm with only a 50% five year survival.

In deeper tumours "sentinel lymph node" biopsy may be carried out to assess whether lymphatic spread has occurred.

How to tell the difference

Benign moles show little change and remain static for years. Any change may indicate that a mole is in fact a melanoma or that a mole is becoming active. Size, shape, and colour are the main features and it is change in them that is most important. Patients with moles should have these changes explained to them, in particular that they indicate activity of the cells, not necessarily malignant change.

Criteria for suspecting malignant changes in pigmented lesions

- Growth—Benign pigmented naevi continue to appear in adolescents and young adults. Any mole increasing in size in an adult over the age of 30 may be a melanoma
- Shape—Moles usually have a symmetrical, even outline, any indentations being quite regular; melanomas usually have an irregular edge with one part advancing more than the others
- Colour—Variation in colour of benign moles is even but a melanoma may be intensely black or show irregular coloration varying from white to slate blue, with all shades of black and brown. Inflammation may give a red colour as well. The amelanotic melanoma shows little or no pigmentation
- Size—Apart from congenital pigmentation naevi most benign moles are less than 1 cm in diameter. Any lesion growing to over 0·5 cm should be carefully checked
- Itching—Normally a mole does not itch but a melanoma may. Irritated seborrhoeic warts also itch
- Bleeding and *crusting* occur in an actively growing melanoma

If more than two of these features are present refer the patient for specialist opinion

A simple summary:
A—Asymmetry of the lesion
B—Irregularity of the border
C—Variations in colour
D—Diameter larger than 0·5 cm

Further reading

Ackerman AB. *Malignant melanoma and other melanocytic neoplasms.* Baltimore: Williams and Wilkins,1984
Mackie R (ed). *Primary and secondary prevention of malignant melanoma.* Basle: Karger
Roses DF. *Diagnosis and management of cutaneous malignant melanoma.* Philadelphia: Saunders, 1983
Seigler HE. *Clinical management of melanoma.* The Hague: Nijhoff, 1982

16 The skin and systemic disease—Genetics and skin disease (JA Savin)

When a man has on the skin of his body a swelling or an eruption or a spot… and the disease appears to be deeper than the skin it is a leprous disease.

Leviticus 13: 2–3

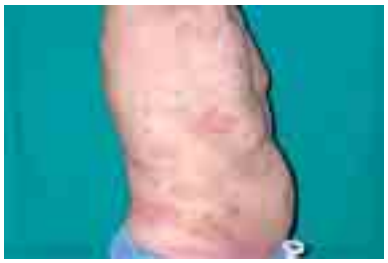
In ancient times changes in the skin were taken to indicate that the whole body was diseased and although arguments continue about what the Old Testament writers understood by "leprous", there was clearly an appreciation of the connection between the skin and systemic illness. Clinical signs in the skin may give valuable diagnostic clues to underlying disease. The cutaneous signs of systemic disease is a very large subject; and what follows is only an outline of the more common skin changes that may be associated with systemic illness.

A disease affecting internal organs may produce the same changes in both the skin and other organs—as in the connective tissue diseases. However, underlying conditions may be associated with skin changes brought about by quite different processes, as in acanthosis nigricans or dermatomyositis in which there is an underlying neoplasm with characteristic skin signs. Sometimes severe skin disease itself may be the cause of generalised illness.

The skin is also a common site for allergic reactions to drugs, with a rash being the first clinical sign. The florid skin lesions of AIDS illustrate the results of infections when the immune response is impaired.

When to suspect an underlying systemic disease

- An unusual rash which does not have the features of one of the common primary inflammatory skin conditions
- Evidence of systemic illness—weight loss, and other symptoms such as breathlessness, altered bowel function or painful joints
- Erythema of the skin due to inflammation around the blood vessels, usually without epidermal changes—reactive erythema. Vasculitis, in which there are palpable erythematous lesions which may be painful or nodular
- Unusual changes in pigmentation or texture of the skin
- Palpable dermal lesions that may be due to granuloma, metastases, lymphoma, or deposits of fat or minerals



Rash from penicillin

Conditions affecting both the skin and the internal organs

Immune reactions

Allergic reactions to drugs such as penicillin can occur. In this case the penicillin molecule attaches to serum protein. This compound acts as an antigen and may form a complex with IgG antibody. It is this complex which attaches to blood vessel walls to produce an inflammatory reaction. This presents as a rash developing a few days to two weeks after treatment on the skin, but if it occurs in the kidneys the resulting tissue damage can have serious consequences. This is an example of Type III allergy with antigen–antibody complexes being deposited in the small blood vessels. Sometimes a much more acute anaphylactic reaction develops A fixed drug eruption is characterised by a localised patch of erythema that flares up whenever the drug is taken. Erythema multiforme can occur in drug reactions.

Erythema multiforme

Connective tissue diseases involve complex immunological processes that affect both internal organs and the skin. This means that it is particularly important to realise the significance of any associated skin changes.

Lupus erythematosus

This condition has been described as "a disease with a thousand faces" because of the wide range of organs involved

and the numerous ways in which it can present. In three quarters of the patients the skin is involved. There are four main types, with numerous variations.

In *systemic lupus erythematosus* (SLE) the commonest skin change is an acute erythematous eruption occurring bilaterally on the malar area of the face in a "butterfly" distribution. There may also be photosensitivity, hair loss, and areas of vasculitis in the skin. There is often intolerance of sunlight. It is more common in females with a female:male ratio of 8:1.

The systemic changes include fever, arthritis and renal involvement, but there may be involvement of a wide range of organs. The criteria for diagnosing the condition include at least four of the features in the box on the right.

Subacute lupus erythematosus is a variant in about 10% of patients with lupus erythematosus that presents with non-scarring erythematosus plaques mainly on the face, hands and arms. Papulo squamous lesions also occur. They may be annular. Systemic involvement is less common and severe than in SLE. It is associated with a high incidence of neonatal lupus erythematosus in children born to mothers with the condition. The antinuclear factor test is positive in 60% and anticytoplasmic antibodies are present in 80% of patients.

Discoid lupus erythematosus (DLE) is a condition in which circulating antinuclear antibodies are very rare. There are quite well defined photosensitive inflammatory lesions, with some degree of atrophy and hyperkeratosis of the follicles, giving a "nutmeg grater" feel. It occurs predominantly on the face or areas exposed to the sun, becoming worse in the summer months. Scarring is common causing hair loss in lesions on the scalp.

Treatment of SLE with the threatened or actual involvement of other organs is important. Prednisolone is usually required and sometimes immunosuppressant drugs such as azathioprine as well. Treatment of DLE is generally with topical steroids. Hydroxychloroquine by mouth is also used, generally in a dose of 200 mg daily. This drug can diminish visual acuity in higher doses and this should be checked every few months. A simple chart, the Amsler Chart, is available for patients to use, consisting of a central dot with a grid which becomes blurred when held at arm's length when there is any impairment of acutity.

Clinical variants of lupus erythematosus

- Systemic
- Subacute cutaneous
- Discoid (neonatal)
- Systemic sclerosis

Criteria for diagnosing systemic lupus erythematosus

- Malar rash
- Discoid plaques
- Photosensitivity
- Arthritis
- Mouth ulcers
- Renal changes
- Serositis
- Neurological involvement
- Haematological changes
- Immunological changes
- Antinuclear antibodies

Systemic lupus erythematosus

Dermatomyositis

This condition is associated in adults with underlying carcinoma—commonly of the breasts, lung, ovary, or gastrointestinal tract. It is characterised by localised erythema with a purple hue (heliotrope), predominantly on the eyelids, cheeks, and forehead. There may be similar changes on the dorsal surface of the fingers, often with dilated nail fold capillaries. These changes may precede the discovery of an underlying tumour and may also fade away once it is removed. There is a variable association with muscle discomfort and weakness, mainly in the upper limb girdle. The finding of muscle weakness together with specific electromyographic changes and an inflammatory infiltrate in the muscle means there is almost certainly an underlying malignancy, so suitable investigation is indicated.

Dermatomyositis

Systemic sclerosis

As the name implies, there is extensive sclerosis of the connective tissue of the lungs, gastrointestinal tract, kidneys, and heart. Endothelial cell damage in the capillaries results in fibrosis and sclerosis of the organs concerned. The skin becomes tethered to the subcutaneous tissues and immobile, leading to fixed claw like hands, constricted mouth with furrowed lips, and beak-like nose. There are vascular changes producing Raynaud's phenomenon and telangiectasia around

Persistent dermatomyositis

the mouth and on the fingers. There are also flat "mat-like" telangiectasia on the face.

Workers manufacturing polyvinyl chloride can develop skin changes similar to systemic sclerosis with erosions of the bones, hepatic and pulmonary lesions. Pesticides and epoxy resin can also produce scleroderma-like changes.

It is associated with antinuclear antibodies (speckled or nucleolar), and in about 50% of cases, circulating immune complexes may be present.

A variant is the *CREST syndrome*. In this type of scleroderma there is **C**alcinosis with calcium deposits below the skin on the fingers and toes, **R**aynaud's phenomenon with poor peripheral circulation, immobility of the o**E**sophagus, dermal **S**clerosis of the fingers and toes, and **T**elangiectasia of the face and lips and adjacent to the toe and finger nails. It has a better prognosis than systemic sclerosis. Antinuclear antibodies at the centromere are frequently present.

Morphoea is a benign form of localised systemic sclerosis in which there is localised sclerosis with very slight inflammation. There is atrophy of the overlying epidermis. The early changes often consist of a dusky appearance to the skin.

CREST syndrome

- C—Calcinosis cutis
- R—Raynaud's phenomenon
- E—Esophagus
- S—Scleroderma
- T—Telangiectasia

CREST syndrome

Lichen sclerosus

The full name is lichen sclerosis et atrophicus—or LSA. This is a relatively uncommon condition seen mainly in women in whom well defined patches of superficial atrophy of the epidermis occur with a white colour. There is fibrosis of the underlying tissues. It frequently occurs in the vulva and perineum and may also appear on the penis as balanitis xerotica obliterans. Extragenital lesions may occur anywhere on the skin. It may occur in a more acute form in children where it tends to resolve, but in adults it is a very chronic condition. There is an increased incidence of squamous cell carcinoma. Treatment is with topical steroids and excision of any areas that appear to be developing tumours.

The cause of the hyalinized collagen and epidermal atrophy is unknown, but in early lesions there is an infiltrate of lymphocytes with CD3, CD4, CD8, and CD68 markers. There is also an increase in Langerhans cells, so there may well be an immunological basis for these changes.

Calcinosis cutis

Vascular changes

Vascular lesions are associated with a wide range of conditions including infections, neoplasia, and allergic reactions. *Hormones*, particularly oestrogen, may affect the small blood vessels of the skin to produce telangiectasia and small angiomas, such as spider naevi.

Vasculitis and purpura, described in chapter 7, may be associated with disease of the kidneys and other organs. "Splinter haemorrhages" under the nails are usually the result of minor trauma but may be associated with a wide range of conditions, including subacute bacterial endocarditis and rheumatoid arthritis.

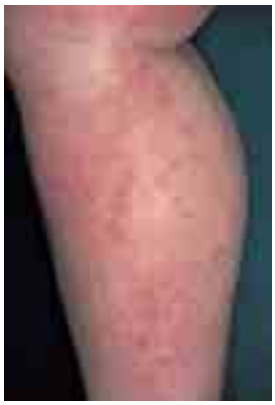
Livedo reticularis is a cyanotic, net-like discoloration of the skin over the legs. It may be idiopathic or associated with arteritis or changes in blood viscosity.

Erythrocyanosis is a dusky, red, cyanotic change in the skin over the legs and thighs, where there is a deep layer of underlying fat. The condition becomes worse in the winter months. It is most common in young women and usually resolves over the years. Lupus erythematosus, sarcoidosis, and tuberculous infection may localise in affected areas.

Telangiectasia and clubbing may be features of scleroderma in the CREST syndrome described above.

In *carcinoid* and *phaeochromocytoma* vasoactive substances cause episodes of flushing and telangiectasia.

Vasculitis



Livedo reticulosis

In *hereditary haemorrhagic telangiectasia* thin walled ectatic blood vessels develop in the mucous membranes and the skin—generally on the upper half of the body and the nail beds.

Erythemas

Erythema is macular redness of the skin due to congestion in the capillaries. It occurs as part of immunological reactions in the skin as in drug allergies and specific patterns of viral infections, such as measles. There are other types that show a specific pattern but are associated with a wide range of underlying conditions, such as erythema multiforme.

Erythema multiforme is associated with herpes and other viral infections or streptococcal and various bacterial infections, but also with many other conditions, particularly connective tissue disease, sarcoidosis, and reactions to drugs such as sulphonamides.

The lesions consist of erythematous macules becoming raised and typically developing into "target lesions" in which there is a dusky red or purpuric centre with a pale indurated zone surrounded by an outer ring of erythema. The lesions may be few or multiple and diffuse, often involving the hands, feet, elbows, and knees. Blisters may develop.

In the more severe forms there may be dermal changes and blister formation with involvement of the mucous membranes (*Stevens–Johnson syndrome*). There is often pyrexia with gastrointestinal and renal lesions. It can progress to *toxic epidermal necrolysis*, which some consider a form of erythema multiforme.

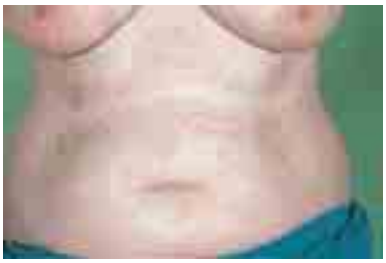
Erythema annulare is a specific pattern in the skin with a large number of reported associations, ranging from fungal and viral infections to sarcoidosis and carcinoma. It consists of a small erythematous macule that enlarges to form an expanding ring, usually on the trunk.

Erythema chronicum migrans is associated with *Borrelia* infection and Lyme disease—it is described on page 106.

There are many other types of "figurate erythemas". *Erythema gyratum repens* is associated with underlying carcinoma and *erythema marginatum*, which is now rare, with rheumatic fever.

Erythemas associated with systemic conditions

- Erythema multiforme
- Erythema annulare
- Erythema chronicum migrans
- Erythema gyratum repens
- Erythema marginatum



Figurate erythema

Angiomas

Spider naevi, which show a central blood vessel with radiating branches, are frequently seen in women (especially during pregnancy) and children. If they occur in large numbers, particularly in men, they may indicate liver failure. Palmar erythema and yellow nails may also be present.

Congenital angiomas

Eruptive angiomas may be associated with systemic angiomas of the liver, lung, and brain. Port wine stain due to abnormality of the dermal capillaries commonly develops on the head and neck. It may be associated with congenital vascular abnormalities of the meningies and epilepsy. Vascular abnormalities of the eye, and also glaucoma, occur with lesions on the face.

Erythema of the nailbeds

This may be associated with connective tissue disease, such as lupus erythematosus, scleroderma, and dermatomyositis.

Erythema of nailbeds and clubbing

Changes in pigmentation

Hypopigmentation

Hormonal

A widespread partial loss of melanocyte functions with loss of skin colour is seen in hypopituitarism and is caused by an absence of melanocyte stimulating hormone.

Genetic

In albinism, an autorecessive condition, there is little or no production of melanin with loss of pigment from the skin, hair, and eyes. Other genetic conditions with loss of skin pigment include piebaldism, phenylketonuria, and tuberous sclerosis.

Localised depigmentation is most commonly seen in vitiligo, in which a family history of the condition is found in one third of the patients. In the sharply demarcated, symmetrical macular lesions there is loss of melanocytes and melanin. There is an increased incidence of organ specific antibodies and their associated diseases.

Other causes of hypopigmented macules include: postinflammatory conditions after psoriasis, eczema, lichen planus, and lupus erythematosus; infections, for example, tinea versicolor and leprosy; chemicals, such as hydroquinones, hydroxychloroquine, and arsenicals, reactions to pigmented naevi, seen in halo naevus; and genetic diseases, such as tuberous sclerosis ("ash leaf" macules).

Hyperpigmentation

There is wide variation in the pattern of normal pigmentation as a result of heredity and exposure to the sun. Darkening of the skin may be due to an increase in the normal pigment melanin or to the deposition of bile salts in liver disease, iron salts (haemochromatosis), drugs, or metallic salts from ingestion. In agyria ingested silver salts are deposited in the skin.

Causes of hyperpigmentation include the following factors.

Hormonal

An increase in circulating hormones that have melanocyte stimulating activity occurs in hyperthyroidism, Addison's disease, and acromegaly. In women who are pregnant or taking oral contraceptives there may be an increase in melanocytic pigmentation of the face. This is known as melasma (or chloasma) and occurs mainly on the forehead and cheeks. It may fade slowly. Sometimes a premenstrual darkening of the face occurs.

Increased deposition of haemosiderin is generalised in haemochromatosis. Localised red-brown discoloration of the legs is seen with longstanding varicose veins. It also occurs in a specific localised pattern in Schamberg's disease, when there is a "cayenne pepper" appearance of the legs and thighs.

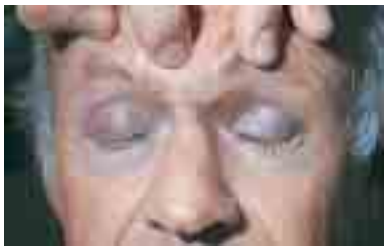
Neoplasia

Lymphomas may be associated with increased pigmentation. Acanthosis nigricans, characterised by darkening and thickening of the skin of the axillae, neck, nipples, and umbilicus, occurs with internal cancers, usually adenocarcinoma of the stomach. It is also seen in acromegaly. There is a benign juvenile type. Pseudoacanthosis nigricans is much more common, consisting of simple darkening of the skin in the flexures of obese individuals; it is not associated with malignancy.

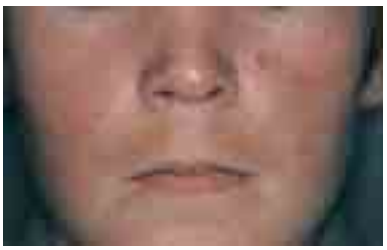
Autoimmune associations with vitiligo

- Thyroid disease
- Pernicious anaemia
- Hypoparathyroidism
- Addison's disease
- Diabetes
- Myasthenia gravis
- Alopecia areata
- Halo naevus
- Morphoea and lichen sclerosus

Vitiligo

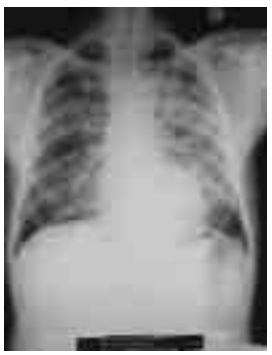


Agyria (silver salts in skin)



Melasma

Acanthosis nigricans



Carcinoma left upper zone

Pseudoacanthosis nigricans

Acanthosis nigricans

Drugs
Chlorpromazine, other phenothiazines, and minocycline may cause an increased pigmentation in areas exposed to the sun. Phenytoin can cause local hyperpigmentation of the face and neck.

Inflammatory reactions
Postinflammatory pigmentation is common, often after acute eczema, fixed drug eruptions, or lichen planus. Areas of lichenification from rubbing the skin are usually darkened.

Malabsorption and deficiency states
In malabsorption syndromes, pellagra, and scurvy there is commonly increased skin pigmentation.

Congenital conditions
There is clearly a marked variation in pigmentation and in the number of freckles in normal individuals. There may be localised well defined pigmented areas in neurofibromatosis with "cafe au lait" patches. Increased pigmentation with a blue tinge occurs over the lumbosacral region in the condition known as Mongolian blue spot.

Peutz–Jeghers syndrome is described under the section "The gut and the skin", below. There are pigmented macules associated with intestinal polyposis in the oral mucosa, lips, and face.


Neurofibromatosis


Neurofibromatosis


Increased pigmentation in malabsorption syndrome


Peutz–Jeghers syndrome

Malignant lesions

Malignant lesions may cause skin changes such as acanthosis nigricans and dermatomyositis or produce secondary deposits. Lymphomas can arise in or invade the skin. Pruritus may be associated with Hodgkin's disease.

Mycosis fungoides is a T cell lymphoma of cutaneous origin. Initially well demarcated erythematous plaques develop on covered areas with intense itching. In many cases there is a gradual progression to infiltrated lesions, nodules, and ulceration. In others the tumour may occur de novo or be preceded by generalised erythema.

Poikiloderma, in which there is telangiectasia, reticulate pigmentation, atrophy, and loss of pigment, may precede mycosis fungoides, but it is also seen after radiotherapy and in connective tissue diseases.

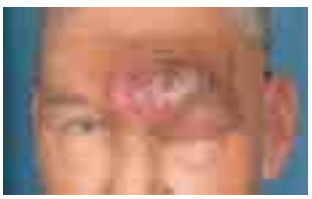
*Parapsoriasis*is a term used for well defined maculopapular erythematous lesions that occur in middle and old age. Some cases undoubtedly develop into mycosis fungoides and a biopsy specimen should be taken of any such fixed plaques that do not clear with topical steroids.

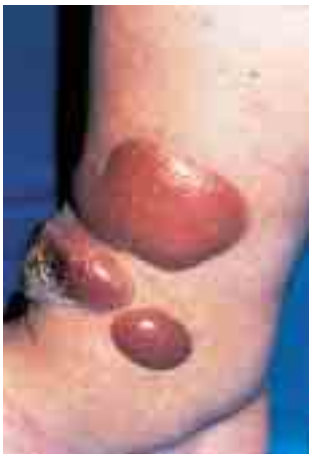
Skin markers of internal malignancy

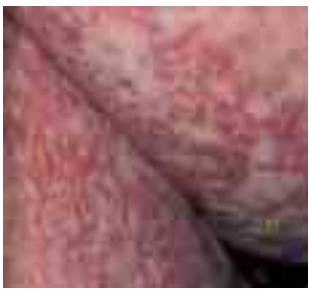
- Acanthosis nigricans—usually intra-abdominal lesions
- Erythematous rashes, "figurate erythema"
- Pruritus—usually lymphoma
- Dermatomyositis in the middle aged and elderly
- Acquired ichthyosis

Non-specific skin changes associated with malignant disease

- Secondary deposits
- Secondary hormonal effects
  - Acne (adrenal tumours)
  - Flushing (carcinoid)
  - Pigmentation
- Generalised pruritus (particularly lymphoma)
- Figurate erythema
- Superficial thrombophlebitis
- Various eruptive skin lesions seen in Gardener's and Bazex syndromes


Lymphoma


B cell lymphoma


Poikiloderma


Parapsoriasis

The gut and the skin

Vasculitis of various kinds, periarteritis nodosa, connective tissue diseases such as scleroderma, and many metabolic diseases produce both cutaneous and gastrointestinal lesions. There are, however, some specific associations.

Dry skin, asteatosis, and *itching*, with superficial eczematous changes and a "crazy paving" pattern, occur in malabsorption and cachectic states. Increased pigmentation, brittle hair and nails may also be associated.

Pyoderma gangrenosum gives rise to an area of non-specific inflammation and pustules break down to form a necrotic ulcer with hypertrophic margins. There is an underlying vasculitis. There is a strong association with ulcerative colitis and also with Crohn's disease, rheumatoid arthritis, abnormal gamma globulins, and leukaemia.

Dermatitis herpetiformis, which has already been discussed, is an intensely itching, chronic disorder with erythematous and blistering lesions on the trunk and limbs. It is more common in men than women. Most patients have a gluten sensitive enteropathy with some degree of villus atrophy. There is an associated risk of small bowel lymphoma.

Peutz–Jeghers syndrome is inherited as an autosomal dominant characterised by the appearance in infancy of pigmented macules of the oral mucosal membranes, lips, and face. Benign intestinal polyps, mainly in the ileum and jejunum, which rarely become malignant, are associated with the condition.

Other conditions include congenital disorders with connective tissue and vascular abnormalities that affect the gut, such as Ehlers–Danlos syndrome and pseudoxanthoma elasticum (arterial gastrointestinal bleeding), purpuric vasculitis (bleeding from gastrointestinal lesions), and neurofibromatosis (intestinal neurofibromas).

In *Crohn's disease* (regional ileitis) perianal lesions and sinus formation in the abdominal wall often occur. Glossitis and thickening of the lips and oral mucosa and vasculitis may be associated.

Liver disease may affect the skin, hair, and nails to a variable degree. Obstructive jaundice is often associated with itching which is thought to be due to the deposition of bile salts in the skin. Evidence of this is the fact that drugs which combine with bile salts such as cholestyramine improve pruritus in some patients. Jaundice is the physical manifestation of bile salts in the skin.

Liver failure is characterised by a number of skin signs, particularly vascular changes causing multiple spider naevi and palmar erythema due to diffuse telangiectasia. It is not unusual to see spider naevi on the trunk in women but large numbers in men should raise suspicion of underlying hepatic disease.

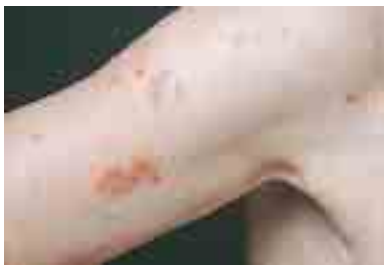
Porphyria cutanea tarda as a result of chronic liver disease produces bullae, scarring, and hyperpigmentation in sun exposed areas of the skin. *Xanthomas* may be associated with primary biliary cirrhosis and in chronic liver disease asteotosis, with dry skin producing a "crazy paving" pattern.

Diseases that pyoderma gangrenosum may occur with

- Ulcerative colitis
- Crohn's disease
- Rheumatoid arthritis
- Monoclonal gammopathy
- Leukaemia

Early pyoderma gangrenosum

Pyoderma gangrenosum



Dermatitis herpetiformis

Liver disease and the skin

Obstructive
- Jaundice
- Pruritus

Liver failure
- Multiple spider naevi (in men)
- Palmar erythema
- White nails—hypoalbuminaemia
- Porphyria cutanea tarda

Cirrhosis
- Xanthomas (primary biliary cirrhosis)
- Asteatosis

Diabetes and the skin

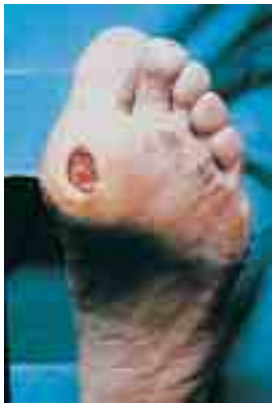
In diabetes the disturbances of carbohydrate–lipid metabolism, small blood vessel lesions, and neural involvement may be associated with skin lesions. The more common of these include the following.

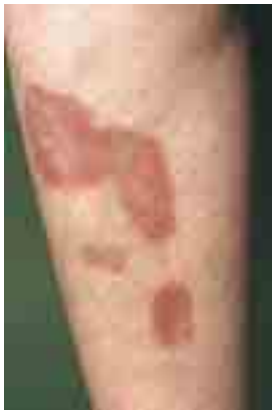
Infection
Diabetic patients have an increased susceptibility to staphylococcal, coliform, and pseudomonal infection. *Candida albicans* infection is also more common in diabetics.

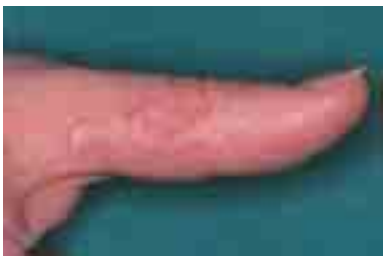
Vascular lesions
"Diabetic dermopathy", due to a microangiopathy, consists of erythematous papules which slowly resolve to leave a scaling macule on the limbs. Atherosclerosis with impaired peripheral circulation is often associated with diabetes. Ulceration due to neuropathy (trophic ulcers) or impaired blood supply may occur, particularly on the feet.

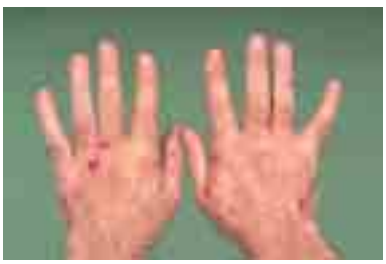
Specific skin lesions
Necrobiosis lipoidica
Between 40% and 60% of patients with this condition may develop diabetes, but it is not very common in diabetic patients (0·3%). As the name indicates, there is necrosis of the connective tissue with lymphocytic and granulomatous infiltrate. There is replacement of degenerating collagen fibres with lipid material. It usually occurs over the shin but may appear at any site.

Granuloma annulare
This usually presents with localised papular lesions on the hands and feet but may occur elsewhere. The lesions may be partly or wholly annular and may be single or multiple. There is some degree of necrobiosis, with histiocytes forming "palisades" as well as giant cells and lymphocytes. It is seen more commonly in women, usually those aged under 30. There is an association with insulin dependent diabetes. In itself it is a harmless and self limiting condition that slowly clears but may recur.


Diabetic ulcer


Necrobiosis lipoidica


Granuloma annulare


Porphyria

Other diseases

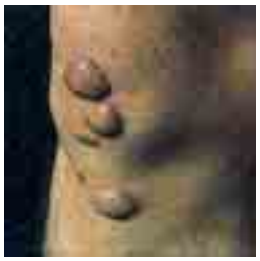
Porphyrias are due to the accumulation of intermediate metabolites in the metabolic pathway of haem synthesis. There are several types. In hepatic porphyrias there is skin fragility leading to blisters from exposure to the sun or minor trauma. In erythropoietic and erythrohepatic photoporphyrias there is intense photosensitivity. They are sometimes associated with sensitivity to long wavelength ultraviolet light that penetrates window glass.

Porphyria cutanea tarda usually occurs in men, with a genetic predisposition, who have liver damage as a result of an excessive intake of alcohol. There is impaired porphyria metabolism leading to skin fragility and photosensitivity, with blisters and erosions, photosensitivity on the face and the dorsal surface of the hands.

Xanthomas are due to the deposition of fat in connective tissue cells. They are commonly associated with hyperlipidaemia—either primary or secondary to diabetes, the nephrotic syndrome, hypothyroidism, or primary biliary cirrhosis. Four of the primary types are associated with an increased risk of atherosclerosis; type I is not. Diabetes may be associated with the eruptive type.

Necrotising fasciitis is an area of cellulitis that develops vesicles. Necrosis of the skin may indicate much more extensive, life threatening necrosis of the deeper tissues. Urgent surgical debridement is indicated.

Amyloid deposits in the skin occur in primary systemic amyloidosis and myeloma.


Xanthomas

Common types of xanthoma

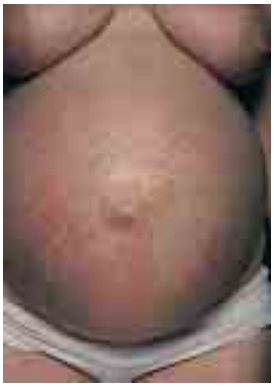
Clinical type	Association with hyperlipidaemia Primary	Secondary
Xanthelasma of the eyelids—yellow plaques	II (may be normal)	
Tuberous nodules on elbows and knees	II, III	+
Eruptive—small yellow papules on buttocks and shoulders	I, III, IV, V	+
Plane—yellow macules, palmar creases involved	I, III	+
Generalised—widespread macules		myeloma
Tendons—swelling on fingers or ankles	II, III	+

Pregnancy

Pregnancy may be associated with pruritus, in which the skin appears normal in 15–20% of women (prunigo gestationis). It is generally more severe in the first trimester.

Polymorphic eruptions also present with pruritis with urticaria papules and plaques (the PUPP syndrome). It usually occurs on the abdomen in the third trimester and then becomes widespread. There may be a postpartum flare up. It can be a distressing condition for the mother but the baby is not affected, and it rarely recurs in subsequent pregnancies. Topical steroids can be used, but systemic steroids should be avoided.

Pemphigoid gestationis is a rare disorder that may resemble PUPP initially but develops pemphigoid-like vesicles, spreading over the abdomen and thighs: autoantibodies to the basement membrane are present.



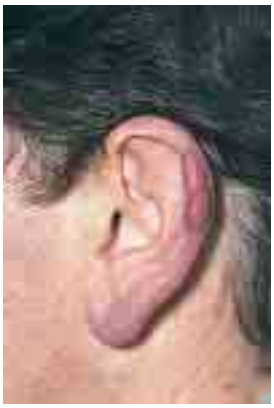
Polymorphic eruption

Pemphigoid gestationis

Sarcoidosis

Pulmonary and other systemic manifestation of sarcoidosis may occur without involvement of the skin. The most common changes are:

- Erythema nodosum, which is often a feature of early pulmonary disease.
- Papules, nodules, and plaques are associated with acute and subacute forms of the disease.
- Scar sarcoidosis, with papules occurring in scars.
- Lupus pernio is characterised by dusky red infiltrated lesions on the nose and fingers.



Nodule in sarcoidosis

Thyroid disease

Thyroid disease is associated with changes in the skin, which may sometimes be the first clinical signs. There may be evidence of the effect of altered concentrations of thyroxine on the skin, with changes in texture and hair growth. Associated increases in thyroid stimulating hormone concentration may lead to pretibial myxoedema. In autoimmune thyroid disease vitiligo and other autoimmune conditions may be present.

Sarcoid granuloma

Genetics and skin disease
by JA Savin

Though many genetic disorders of the skin are inherited in a classically Mendelian way (single gene disorders), others are genetically more complex. As a general rule, the common skin disorders that run in families, such as psoriasis, atopic eczema, and acne, tend to belong to the latter group.

Single gene disorders

Recent advances in genetic technology have been relatively easy to apply to these, usually rather uncommon, disorders, most of which had already been classified accurately on clinical grounds. Several things followed from this:

(1) The next step has often been an improvement in current clinical classifications, which can now be based logically on the underlying molecular abnormalities of the disorders in question. A good example of this is the modern classification of the inherited mechano-bullous disorders (also known as epidermolysis bullosa).
(2) New skin constituents were quickly recognized, their function being understood after studying the disorders in which they

Clinical signs of thyroid disease

Hypothyroidism	Hyperthyroidism
Dry skin	Soft, thickened skin
Oedema of eyelids and hands	Pretibial myxoedema
Absence of sweating	Increased sweating (palms and soles)
Coarse, thin hair—loss of pubic, axillary, and eyebrow hair	Thinning of scalp hair
Pale "ivory" skin	Diffuse pigmentation
Brittle poorly growing nails	Rapidly growing nails
Purpura, bruising, and telangiectasia	Palmar erythema Facial flushing

are abnormal. Soon it was realized that the same molecules could be the targets both for genetic abnormalities and for acquired skin diseases. One example of this is the way in which autoantibodies directed against one of the constituents of hemidesmosomes (BP180) cause pemphigoid, whereas mutations in the gene responsible for BP180 are the basis of the junctional type of epidermolysis bullosa.

(3) Advances have been made too in our understanding of the structure and function of normal skin and its appendages—for example, the finding that melanocortin-1 receptor gene variants are associated with fair skin, red hair, and skin tumours.

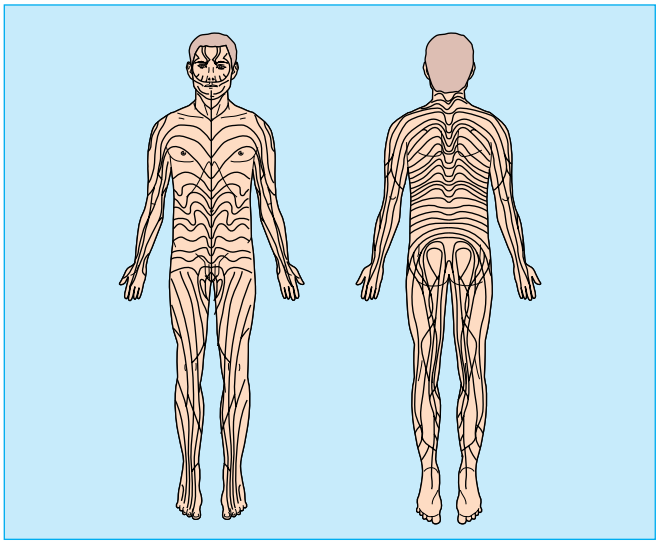
(4) Mosaics were soon recognized. Clinically these are linear abnormalities in the skin, usually present at birth, which often contain cells with the same genetic abnormalities as those of known generalised genodermatoses. A good example of this is the way the same abnormalities in the genes controlling the production of keratins 1 and 10 can be responsible both for a generalised skin condition (epidermolytic hyperkeratosis) and for warty linear naevi. The mosaic areas follow Blashko's lines, a bizarre pattern of lines and whorls, which are not the same as dermatomes.

(5) The prenatal diagnosis of severe genodermatoses has become more accurate, though gene therapy has not yet fulfilled its early promise.

Genetically complex disorders

Psoriasis is a good example. It clusters in some families but does not follow a classical Mendelian pattern of inheritance. Environmental triggers are important, as well as genetic factors. Over the last few years, several wide scans of the genome have been undertaken with the aim of identifying the location of the genes that determine susceptibility to psoriasis. Five have been confirmed, all on different chromosomes, and now designated as Psors1 to Psors5. A further six loci may have similar effects, but the evidence for them is less strong. Psors1, on chromosome 6p21.3, is an especially important gene for psoriasis susceptibility in many populations and lies within the area of the major histocompatibility complex (MHC). However it is not itself an HLA class 1 gene, and may belong to the newly described MHC class 1 chain-related (MIC) gene family. The possession of one allele (A5.1) of this gene seems to lead to a type of psoriasis that starts especially early, and is more common in familial than in sporadic cases.

In *atopic eczema*, matters are equally complicated. Environmental factors may well be responsible for the recent rise in its prevalence as the gene pool within the population is not likely to have changed greatly, but a genetic component is obvious too, even though affected children can be born to clinically normal parents. Within each family, atopic disorders tend to run true to type, so that, in some, most affected members will have eczema, in others, respiratory allergy predominates. The inheritance of atopic eczema probably involves genes that predispose to the state of atopy, and others that determine whether it is asthma, eczema, or hay fever that develops. One plausible gene for the inheritance of atopy encodes for the β subunit of the high affinity IgE receptor, and lies on chromosome 11q13. However several groups have failed to confirm earlier reports of this linkage, and a gene linked to atopic eczema has recently been found on chromosome 3q21.



Blashko's lines

The abnormality underlying some inherited skin disorders

Skin disorder	Abnormality in
Ehlers–Danlos syndrome	Collagen and the extracellular matrix
Dystrophic epidermolysis bullosa	Type VII collagen
Pseudoxanthoma elasticum	Elastic tissue
Xeroderma pigmentosum	DNA repair
Simple epidermolysis bullosa	Keratins 5 and 14
Epidermolytic hyperkeratosis	Keratins 1 and 10
Palmoplantar keratoderma	Keratins 9 and 16
Junctional epidermolysis bullosa	Laminins
X-linked recessive ichthyosis	Steroid sulphatase
Darier's disease	Epidermal cell adhesion
Albinism (tyrosinase negative type)	Tyrosinase

Further reading

Fine JD, Eady RA, Bauer EA, et al. Revised classification system for inherited epidermolysis bullosa. *J Am Acad Dermatol* 2000;42: 1051–66

Harper J. *Inherited skin disorders*. Oxford: Butterworth–Heinemann 1999

Moss C, Savin JA. *Dermatology and the new genetics*. Oxford: Blackwell Science, 1995

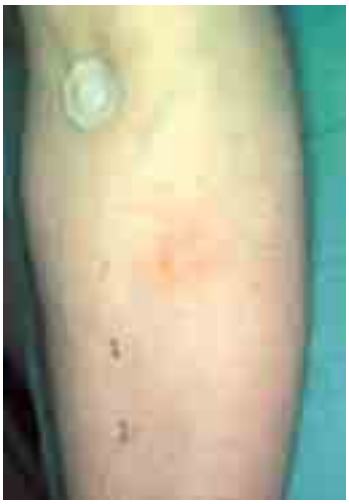
17 Cutaneous immunology—Autoimmune disease and the skin (DJ Gawkrodger)

Types of allergic reaction

Allergic and other immune reactions may occur in the skin—the "immunological battleground of the body"—rather than involving internal organs. An acute vasculitis occurring in the skin is unpleasant and requires treatment but the same reaction occurring in the kidneys can be life threatening. The pattern of skin changes can indicate the type of immune process involved and also whether there is likely to be systemic involvement. The immune response of the skin is also used clinically in the tuberculin skin test to detect the level of immunity to tuberculosis. It is also the means of immunisation when an injection of inactivated organisms induces an immune response that protects the entire body.

The different types of immune reaction are all manifested in the skin as part of a normal response to pathogens or as an allergic reaction. The difference is expressed by the word "allergy", first used by Von Pirquet in 1906, derived from the Greek (αγοξ ενγον), meaning literally "other work". In other words it is a response that is appropriate for pathogenic organisms such as a tubercle bacillus but is misdirected against a harmless substance such as a rubber glove or the metal of a watch strap buckle.

Immunological reactions are of four types—five if autoimmunity is counted—of responses mediated by antibodies known as the humoral response and one by the lymphocytes known as the cell mediated response.



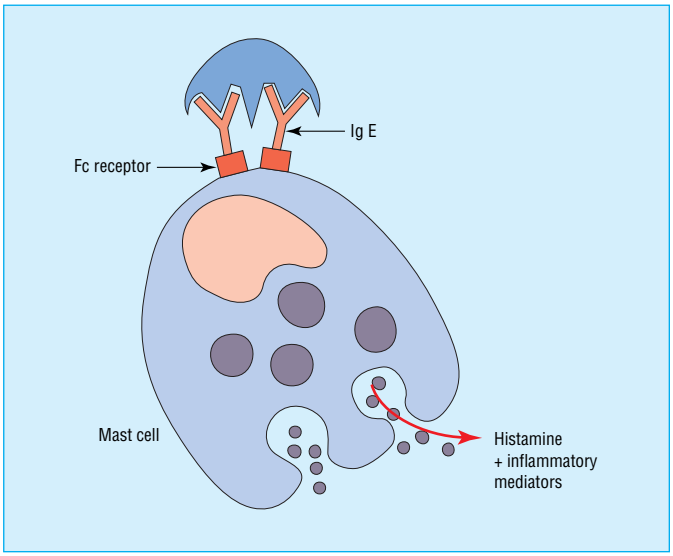
Reaction to fish protein

Immediate hypersensitivity

This type of reaction is caused by "reagin" antibodies, which consist mainly of lgE, that react with allergens such as housedust mite, animal dander, or grass pollens. These reactions may occur in both the skin or the lung to produce asthma. Allergic reactions to insect stings can cause severe systemic effects—"anaphylaxis", which literally means "without protection". Food proteins can also cause an immediate type of hypersensitivity reaction. The IgE molecule is attached to specific receptors on the surface of mast cells and when activated by linkage to specific allergen inflammatory mediators are released. This is an acute process, hence the name "immediate hypersensitivity".

The initial response occurring within five minutes is due to by the release of histamine, heparin, tryptophan. This is followed by inflammatory mediators—released in five to 30 minutes—leukotrien, prostaglandin. The later response, occurring after some hours, is caused by cytokines—predominantly tumour necrosis factor α (TNF-α) and interleukin 4 (IL4).

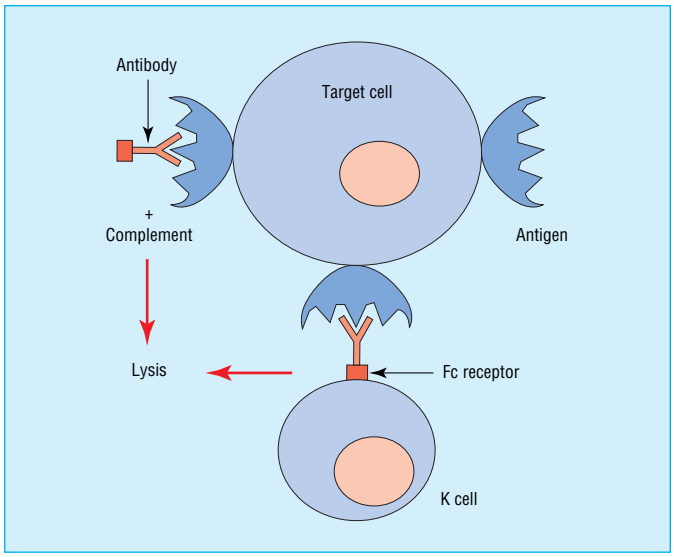
Severe reactions cause shock that is made worse by stress and exercise, as in the case of a young woman, allergic to wasp stings, who had a wasp sting when picnicking by a lake. She then plunged into the cold water, swimming vigorously, leading to a fatal anaphylactic reaction. Acute anaphylactic reactions to peanuts may be life threatening.



Type I—immediate hypersensitivity

Cytotoxic reactions

In this case cells become the target of attack by circulating antibodies. There are a number of causes, such as drugs or proteins attached to the cell surface that act as haptens so they



Type II—cytotoxic

become antigenic. This occurs in drug induced haemolysis from drugs. Alternatively, immune complexes are attached to the surface of the cell with the incorporation of complement leading to lysis. In haemolytic anaemia and incompatible blood transfusions antibodies are formed against erythrocytes. They may also be destroyed by killer cells. A typical example is haemolytic anaemia.

This immune response is the means of destroying cells that become antigenic as a result of being infected with virus.

In autoimmune diseases antibodies are directed against specific structures.

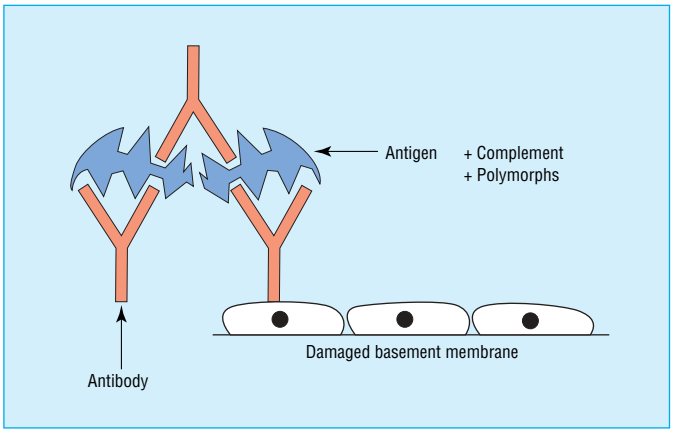
Antigen–antibody complex reactions

As a result of antibody production to antigens in the circulation, complexes form in the blood and these may be deposited in capillaries resulting in inflammatory changes. Similar changes may occur in the lung. This involves the activation of complement and the release of mediators of inflammation, producing vasodilatation and the accumulation of polymorphs.

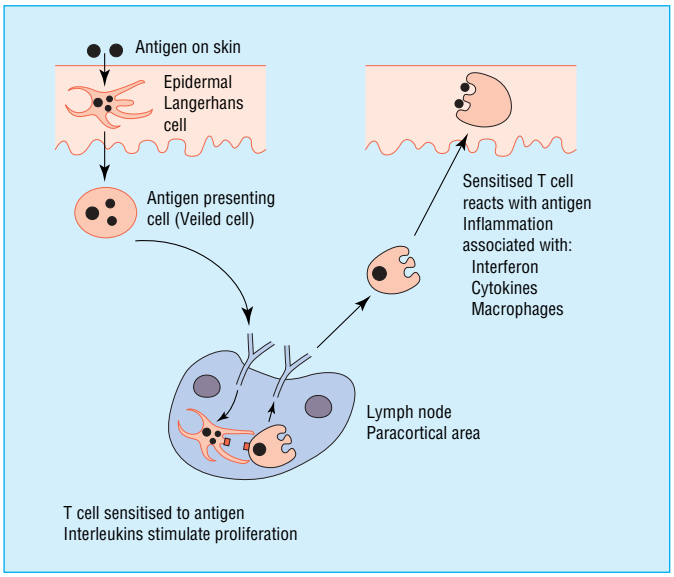
Delayed hypersensitivity

This type of reaction results from lymphocytes known as T cells, because of their derivation from the thymus, which react with antigen in the skin. The reaction is initiated by antigen attached to Langerhan's cells in the epidermis being transported to the paracortical area of the regional lymph node with the production of lymphocytes sensitised specifically for that antigen. There is also the production of interleukin which has a feedback effect in stimulating the production of more sensitised lymphocytes.

The reaction of the T lymphocytes in the epidermis results in the accumulation of macrophages and the release of inflammatory mediators.



Type III—circulating immune complexes



Type IV—delayed hypersensitivity

Autoimmune disease and the skin
by DJ Grawkrodger

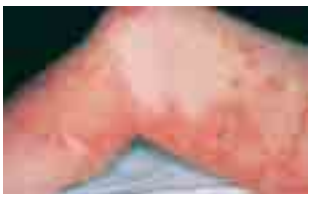
There is always the risk that the well developed human immune system may react against the body's own tissues, with a failure to distinguish between "self" and "non-self". An immune response develops which may be specific for a particular organ, such as the thyroid gland, or react against a number of different organs, as in connective tissue diseases. The skin can manifest both types of autoimumme response. The results of such reactions can be destruction of the cells concerned and the production of inflammation. There is an inherited tendency to autoimmune disease, marked by specific HLA (human lymphocyte antigen) in some cases.

The most common types of skin disease in which this autoimmune mechanism occurs are the blistering disorders, pemphigoid and pemphigus, as well as dermatitis herpetiformis.

Pemphigoid

In this condition large, tense blisters develop in which there are antibodies attached to the upper layer of the basement membrane at the dermo-epidermal junction, with an underlying inflammatory reaction producing a split above the basement membrane. Lysosomal enzymes are released damaging the basement membrane, resulting in separation of the epidermis and blister formation. The presence of

Reaction to metal

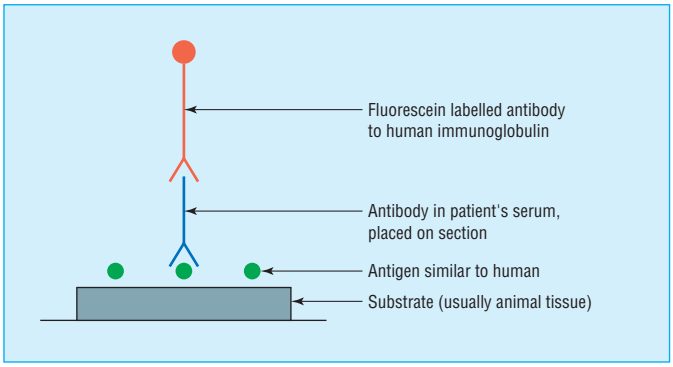


Blistering disorder as a result of an autoimmune response

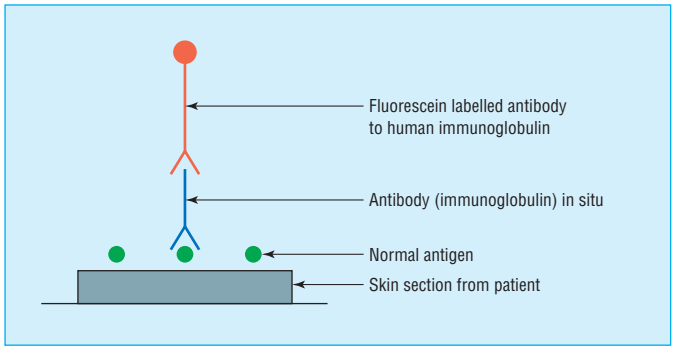
Split at dermo-epidermal junction

antibodies, usually IgG, can be shown by an antihuman IgG antibody labelled with fluorescein. When viewed under the microscope with ultraviolet light illumination, the presence of the IgG antibody is shown by fluorescence. The presence of circulating antibasement membrane antibodies in the serum can be shown either by direct immunofluorescence using a specimen of the patient's skin or by incubation by attachment to skin which has been incubated in serum from the patient.

The clinical features are described in chapter 8. The blisters develop, frequently with an erythematous background, on the limbs, trunk, and flexures. It is mainly seen in the elderly and is slightly more common in women.



Indirect immunofluorescence



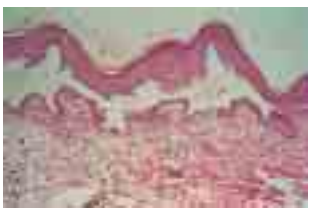
Direct immunofluorescence

Indirect immunofluorescence

Pemphigus

Pemphigus

In this condition, antibodies are found to have developed against the epidermis above the basement membrane. The main antibody is IgG, but IgM and IgA may also be found. As a result of this reaction, there is separation of the epidermal cells with the formation of a superficial blister. A row of basal cells remains attached to the basement membrane. Direct immunofluorescence of the skin from affected patients shows that antibodies are deposited on the intercellular substance of the epidermis. Circulating antibodies are often present. Oral lesions are much more common than in pemphigoid.



Intraepidermal split

Direct immunofluorescence

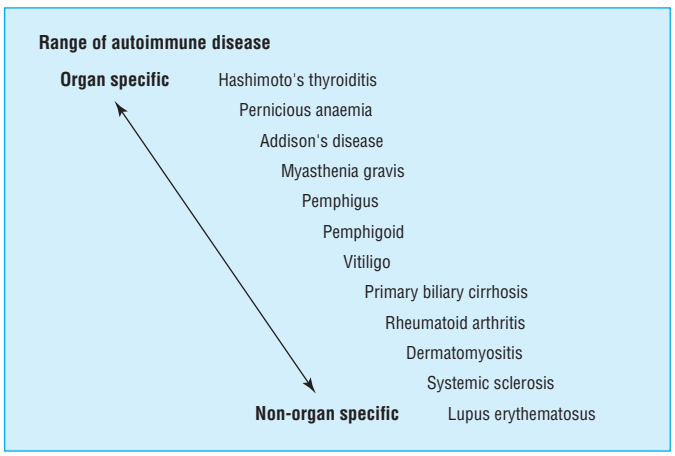
Other organ-specific autoimmune diseases of the skin

Vitiligo

In this condition there is a loss of pigment as a result of antibodies developing against melanocytes in the skin in a limited area. However, the areas affected tend to gradually increase. There may be other autoimmune diseases in the same patient, causing, for example, pernicious anaemia, and thyroid disease.

Alopecia areata

There is evidence that this condition may be associated with an immune reaction against the hair follicle. The increased incidence of antibodies to the thyroid gland and gastric parietal cells in patients with alopecia areata provides circumstantial support for an autoimmune aetiology.



Range of autoimmune disease

Non-organ-specific skin autoimmune disease

Systemic lupus erythematosus (SLE)

The hallmark of this condition is the presence of antibodies against various components of the cell nucleus. Although a wide range of organs may be affected, in three quarters of the patients the skin is involved, generally with an erythematous eruption occurring bilaterally on the face in a "butterfly" distribution. There may also be photosensitivity, hair loss, and areas of vasculitis in the skin. There is often intolerance of

Clinical variants of lupus erythematosus

- Systemic
- Subacute cutaneous
- Discoid
- (Neonatal)

sunlight. Subacute lupus erythmatosus is a variant that presents with an erythematous eruption in the skin and anticytoplasmic RNA molecules.

Discoid lupus erythematosus (DLE)
This is a condition in which circulating antinuclear antibodies are very rare. There are quite well defined inflammatory lesions, with some degree of atrophy occurring on the face and occasionally on the arms as well.

Treatment of SLE with the threatened or actual involvement of organs is important. Prednisolone is usually required and sometimes immunosuppressant drugs such as azathioprine as well. Treatment of DLE is generally with topical steroids. Hydroxychloroquine by mouth is also used, generally in a dose of 200 mg daily. This drug can diminish visual acuity and this should be checked every few months. A simple chart, the Amsler Chart, is available for patients to use, consisting of a central dot with a grid which becomes blurred when held at arm's length when there is any impairment of acuity.

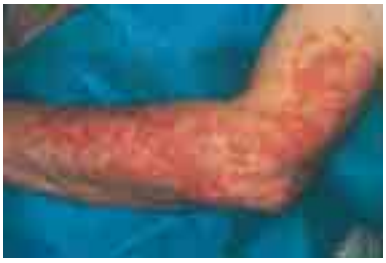
Systemic sclerosis
This is a condition in which there is extensive sclerosis of the subcutaneous tissues in the fingers and toes as well as around the mouth (scleroderma), with similar changes affecting the internal organs, particularly the lung and kidneys. There are vascular changes producing Raynaud's phenomenon and telangiectasia around the mouth and fingers. It is associated with antinuclear antibodies (speckled or nucleolar), and in about 50% of cases circulating immune complexes may be present. Endothelial cell damage in the capillaries results in fibrosis and sclerosis of the organs concerned. There is considerable tethering of the skin on the fingers and toes, which becomes very tight with a waxy appearance and considerable limitation of movement. A variant is the CREST syndrome.

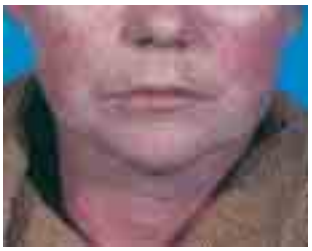
Morphoea is a benign form of localised systemic sclerosis in which there is localised sclerosis with very slight inflammation. There is atrophy of the overlying epidermis. The early changes often consisit of a dusky appearance to the skin.

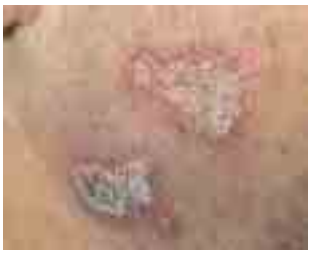
The clinical features are described in chapter 16.

Dermatomyositis
This condition is described in chapter 16, but the main immunological features are deposition of IgG, IgM, and C3 at the dermo-epidermal junction in about half the cases in the early stages, as well as a lymphocytic infiltrate with CD4+ cells and macrophages. There are reports of autoantibodies in some patients. Dermatomyositis may represent an immune reaction to an underlying mechanism or derangement of the normal immune response.

Lichen sclerosus
This condition is also described in chapter 16 and is characterised by atrophic patches of skin. It occurs mainly in females and predominantly involves the genitals and perineum. The cause is unknown but in early lesions there is a band of lymphocytes, mainly CD3, CD4, and CD8. Immunoglobulins and complement accumulate in the affected areas. There is an association with vitiligo, morphoea, alopecia, and pernicious anaemia, suggesting an autoimmune association.

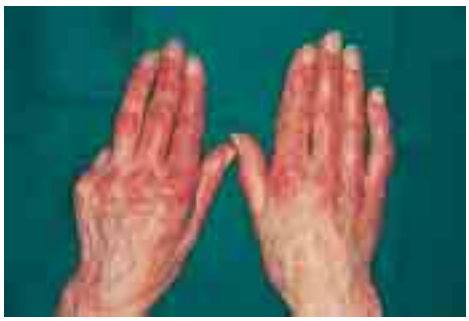

Subacute lupus erythematosus


Subacute lupus erythematosus


Discoid lupus erythematosus


Systemic sclerosis

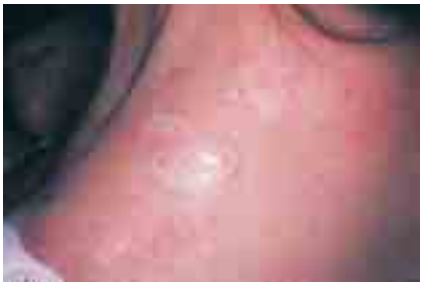

Morphoea


Dermatomyositis


Lichen sclerosus

Graft versus host disease

This reaction occurs following bone marrow transplantation in immunosuppressed patients. T lymphocytes produced by the graft react against the body's own tissues, producing a skin eruption which may resemble measles. There is lysis of the basement membrane with shedding of the skin and sometimes lichen planus-like eruption. In the more chronic form, localised lesions develop, with immunoglobulins deposited in the walls of blood vessels with the activation of complement.



Graft versus host disease

Further reading

Roitt IM, Brostoff J, Male D. *Immunology*, 6th ed. St Louis: Mosby, 2001

18 Bacterial infection

RJ Hay

The process of infection involves the interaction between two organisms—the host and the invader. The clinical changes result from mechanisms involved in this process, notably the micro-organism, its virulence, and the patient's immune defenses. The lesions produced often have a well defined appearance, such as impetigo or tinea cruris, but the changes may be less specific.

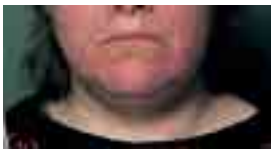
Several features enable us to recognise that infection is a possible cause of the patient's condition. Acute bacterial infections generally produce some or all of the classical characteristics of acute inflammation.

Cardinal signs of infection

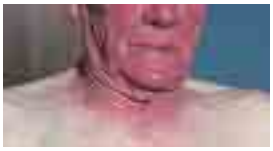
- Erythema
- Swelling and oedema
- Heat or warmth
- Pain and discomfort

Erythema of the face

	Usually unilateral	Usually bilateral	Photosensitive
Acute			
(1) Allergic reactions			
Cosmetics		+	+ or −
Plants	+	or +	+ or −
Drugs		+	+ or −
(2) Urticaria		+	−
Reactions to light			
(3) Photodermatitis		+	+
Solar urticaria		+	+
(4) Infection			
Erysipelas		+	−
Fifth disease ("slapped cheek")		+	−
(5) Rosacea		+	+ or −
Chronic—recurrent			
(6) Lupus erythematosus			
Systemic		+	+
Discoid	+	or +	+
(7) Seborrhoeic dermatitis		+	−
(8) Acne		+	+
(9) Perioral dermatitis		+	−
(10) Vascular naevus	+	−	−



Allergic reactions to cosmetics

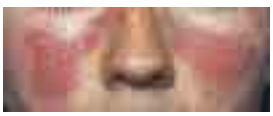


Photodermatitis

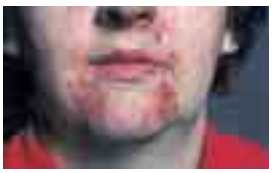
Erysipelas

Rosacea

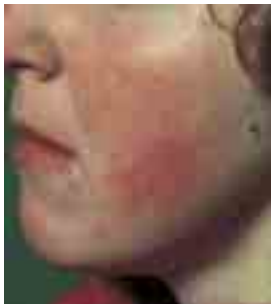
Fifth disease ("slapped cheek")



Systemic lupus erythematosus

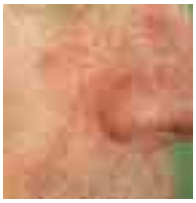


Perioral dermatitis



Vascular naevus

Discoid lupus erythematosus



Seborrhoeic dermatitis

Clinical presentation

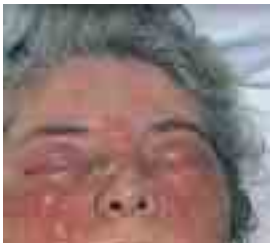
The woman shown in the photographs had acute *erysipelas* due to streptococcal infection, and all four features of inflammation were present. She was referred to the clinic with a diagnosis of an acute allergic response, which, from the appearance alone, was understandable. However, malaise and fever were also present and the lesions were warm and tender. The condition responded well to antibiotic treatment. The point of entry in such cases is thought to be a small erosion on the face. Erysipelas of the leg or foot may follow the development of a small fissure between the toes, but often there is no discernible portal of entry.

Erysipelas is the local manifestation of a Group A streptococcal infection, in the case illustrated the infection is confined to deep dermis as a form of cellulitis. However the same organism at distal sites, through the production of toxins or superantigens, can cause other skin lesions such as: (a) the rash of scarlet fever; (b) erythema nodosum; (c) guttate psoriasis; and (d) an acute generalised vasculitis.

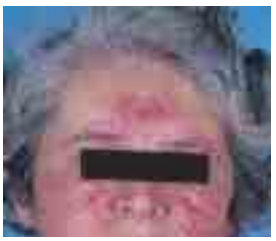
Other forms of local bacterial infection include *impetigo*, *folliculitis*, and *furuncles* (boils). These conditions are caused by *Staphylococcus aureus* and in the case of folliculitis or boils the infection is associated with a local abscess. *Staph. aureus* colonises the anterior nares or perineum of normal people; it also commonly colonises eczema and may cause an acute exacerbation of atopic dermatitis.

Impetigo is a superficial infection of the skin of which there are two forms. In the non-bullous form the affected skin is covered with crusts. Both staphylococci and streptococci are responsible. However the bullous form which presents with blisters is due to staphylococci. Folliculitis, an inflammation of the hair follicle, is commonly caused by *Staph. aureus*. Infection of the scalp or beard hair (sycosis barbae) is uncommon but may become chronic. Abscess formation around the hair follicles may result in furuncles or boils; where several furuncles coalesce the lesion is known as a carbuncle.

Ecthyma, which is most common on the leg, is due to bacterial infection penetrating through the epidermis to the dermis causing a necrotic lesion with a superficial crust and surrounding inflammation. Both streptococci and staphylococci are responsible.



Acute erysipelas: presentation



Acute erysipelas: patient shown in same patient two weeks later

Clinical presentation—points to note

- In any patient with a localised area of acute erythema, swelling, and fever—consider infection
- Remember that a generalised erythematous rash may be the manifestation of a localised infection. Scarlet fever arises from streptococcal sore throat, and herpes simplex of the lip may be associated with erythema multiforme
- The common pathogens are also commensals—recent studies showed that 69% of individuals are nasal carriers of *Staphylococcus aureus* and some may carry Group A streptococci in the throat

Mycobacterial disease

The clinical presentation of infections due to mycobacteria, a specific group of organisms that includes the causes of tuberculosis and leprosy, reflects the success of the host's response in eradicating organisms. There are clear differences, for instance, between disseminated miliary tuberculosis and lupus vulgaris or, for example, tuberculoid and lepromatous leprosy. These are discussed in chapter 23. As these infections are not common only lupus vulgaris and non-tuberculous or "atypical" mycobacterial infection are described.

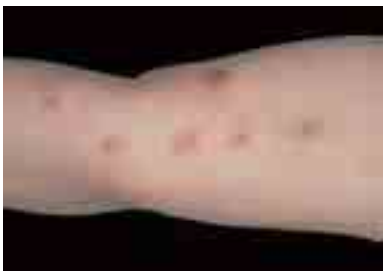
Tuberculous mycobacterial infections

Lupus vulgaris presents as a very slowly growing indolent plaque. It usually represents a localised skin infection disseminated from a deep focus of infection. Squamous carcinomas may develop in long standing cases.

Lupus vulgaris

Mycobacterial disease—histology

Non-tuberculous mycobacterial infections
The most common is "fishtank" or "swimming pool" granuloma, acquired from tropical fish or rarely swimming pools, respectively, and caused by *Mycobacterium marinum*. Nodular lesions develop slowly with ulceration and may spread along local lymphatics to give a chain of nodules (sporotrichoid spread). Injection abscesses may be caused by mycobacteria such as *M. chelonei*. Buruli ulcer, an extensive ulcerating condition due to *M. ulcerans*, is confined to the tropics.


Swimming pool granuloma

Other infections

Rochalimea infections include bacillary angiomatosis, which presents in AIDS patients with small haemangioma-like papules, and cat scratch disease where crusted nodules appear at the site of the scratch associated with the development of regional lymphadenopathy one or two months later. A maculopapular eruption on the face and limbs or erythema multiforme may occur.

Psittacosis and ornithosis may be associated with a rash.

Rickettsial infections, including typhus, Rocky Mountain spotted fever, and rickettsial pox, are all associated with rashes, often purpuric.


Secondary syphilis

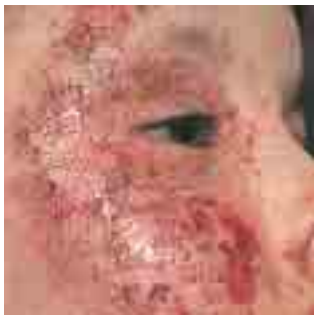
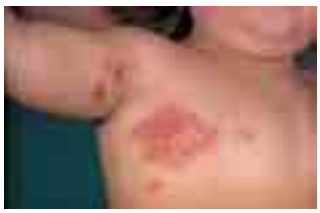
Syphilis
There is a disseminated erythematous rash in secondary syphilis that is followed by a papulosquamous eruption, which affects the trunk, limbs, and mucous membranes. The palms and soles may be involved. There are also small clustered mouth ulcers. In patients with AIDS the rash of secondary syphilis is florid and often crusted or scaly.

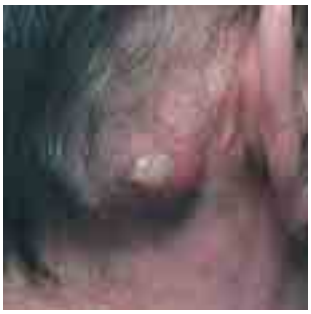
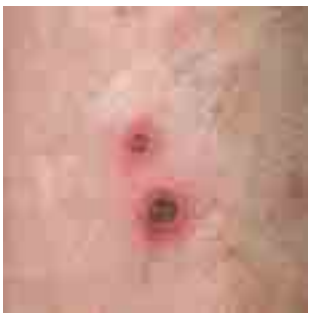
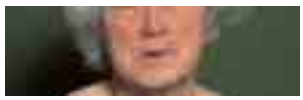
Further reading

Carnwres O, Harman RM. *Clinical tropical dermatology*. 2nd ed. Oxford: Blackwell Scientific, 1992

Noble WC. *The skin: microflora and microbial skin disease*. Cambridge: Cambridge University Press

Common patterns of cutaneous bacterial infection

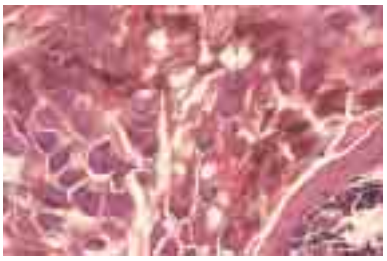
	Infected eczema	Impetigo (nonbullous)	Impetigo (bullous)
			
Appearance	Exudate Crusts Inflammation	Transient vesicles Exuding lesions with yellow crusts Erythema Affects mainly face and limbs, commonly in children	Erythema and bullae which rupture leaving superficial crusts Affects mainly face and limbs in children and adults
Cause	Persistent scratching Topical steroids	Toxic reaction between organisms and epidermis resulting in superficial epidermal split in bullous lesions	
Organisms	*Staphylococcus aureus* *Streptococcus pyogenes*	*Staphylococcus aureus* *Streptococcus pyogenes* in some outbreaks	*Staphylococcus aureus*
Treatment	(1) Weaker topical steroids (for the eczema) with topical antibiotics (2) Systemic antibiotics if necessary (3) Soaks with potassium permanganate	Topical antibiotics Systemic antibiotics against both streptococcal and staphylococcal infection	Topical and systemic antibiotics
Notes	• Avoid prolonged use of topical antibiotics • Return to using weaker steroid • It is wise to send a specimen for bacteriology: when infection has healed • Even without clinical evidence of infection most lesions of atopic eczema are colonised by *Staphylococcus aureus*	• Staphylococcal infection can cause generalised superficial shedding of the epidermis—"scalded skin syndrome" (Lyell's disease) • It is wise to send specimens for culture in an outbreak to identify presence of Group A streptococci potentially implicated in glomerulonephritis	

Boils (furuncles)	Folliculitis	Ecthyma	Erysipelas
			
Inflammatory nodule affecting the hair follicles develops into a pustule Tender induration with severe inflammation, followed by necrosis Heals with scarring Affects all ages Several boils may coalesce to form a carbuncle	Various forms: (1) *Scalp* *Children*—"Follicular impetigo" *Adults*— (a) Keloidal folliculitis Back of neck (b) Acne necrotica Forehead/hairline (2) *Face*—"Sycosis barbae" (3) *Legs*—Chronic folliculitis	Small bullae may be present initially An adherent crust is followed by a purulent ulcerated lesion with surrounding erythema and induration, which slowly heals Usually on legs	Well defined areas of erythema—very tender, not oedematous Vesicles may form Common sites—abdominal wall in infants; in adults the lower leg and face An area of broken skin, forming a portal of entry, may be found
Underlying disease—for example, atopy Mechanical damage from clothing, occlusion	(1) Underlying disease—for example, eczema (2) Infection may be by mechanical precipitated injury, greasy emollients, and occlusive dressings	Hot climate, occlusion More common in debilitated individuals May follow secondary infection of chicken pox	Lymphoedema and severe inflammation due to bacterial toxins
Staphylococcus aureus, usually of same strain as in nose and perineum	*Staphylococcus aureus* *Propionibacterium acnes* *Malassezia* spp. *Pseudomonas* spp. and other Gram negative organisms	Both *Streptococci* and *Staph. aureus*	*Strep. pyogenes* (group A, but may be B, C, or G) *Staphylococcus aureus* *Klebsiella pneumoniae* *Haemophilus influenzae*
(1) Antibiotic (penicillinase resistant) systemically (2) Cleaning of skin with weak chlorhexedine solution or a similar preparation	Topical and long term systemic antibiotics—for example, erythromycin Topical antifungal for *Malassezia* infection	Improve nutrition Use antibiotic effective against both staphylococci and streptococci	Penicillin or erythromycin
• Nasal and perineal swabs should be taken to identify carriers • Remember unusual causes—a bricklayer presented with a boil on the arm with necrosis due to anthrax (malignant pustule) acquired from the packing straw used for the bricks	• Gram negative folliculitis occurs on the face—a complication of long term treatment for acne • Gram negative folliculitis on the body is associated with exposure to contaminated baths or whirlpools	• Check for debilitating diseases, reticuloses, diabetes	• Cellulitis affects the deeper tissues and has more diverse causes, being essentially inflammation of the connective tissue • *Streptococcus, Staphylococcus, Haemophilus*, and other organisms may be found • One of the complications of erysipelas of the face is thrombosis of the cavernous sinus

19 Viral infections

Like the pyogenic bacteria, viruses produce local lesions and may also cause a widespread reaction to the infection such as erythema multiforme. However, the clinical manifestations of common viral infections of the skin are easily recognised.

Local infective lesions caused by DNA viruses which can be isolated from the lesions themselves include the herpes and pox virus groups. In patients with AIDS chronic and widespread viral infections of the skin occur.



Molluscum inclusion bodies (a pox virus)

Herpes

Herpes simplex

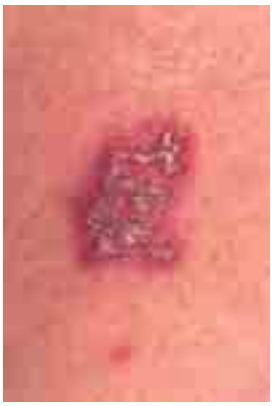
The herpes simplex virus consists of two viral subtypes. Type I is associated with lesions on the face and fingers and sometimes genital lesions. Type II is associated almost entirely with genital infections. Recurrent episodes of infection are common, with both due to latent infection of sensory nerve ganglia.

Primary herpes simplex (type I) infection usually occurs in or around the mouth, with variable involvement of the face. Lesions are small vesicles which crust over and heal but there may be considerable malaise. Type II infection affects the external genitalia and waist area.

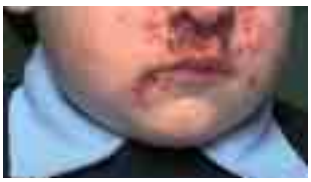
Recurrent infections are shorter lived (three to five days), occur in the distribution of a sensory nerve on the face or genitalia, and may be triggered by a variety of stimuli from sunlight to febrile illness.

Herpes of lips

Inoculation herpes



Herpes simplex



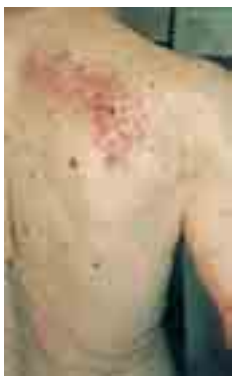
Eczema herpeticum

Herpes simplex—points to note

- The initial vesicular stage may not be seen in genital lesions, which present as painful ulcers or erosions
- There is usually a history of preceding itching and tenderness
- The most rapid methods of detecting virus from scrapings from the base of the ulcer are electronmicroscopy, immunofluorescence, or PCR
- Genital herpes in a pregnant woman carries a great risk of ophthalmic infection of the infant. Caesarean section may be indicated
- "Eczema herpeticum" or "Kaposi's varicelliform eruption" are terms applied to severe cutaneous and, less commonly, systemic, infection with herpes virus in patients with atopic eczema and some other skin conditions. Treatment is with oral or parenteral aciclovir

Herpes zoster

Varicella zoster virus (VZV) causes both chickenpox, the primary illness, and herpes zoster, which follows reactivation of the virus in the nerve ganglia. In zoster, pain, fever, and malaise may occur before erythematous papules develop in the area of the affected dermatome—most commonly in the thoracic area. Vesicles develop over several days, crusting over as they resolve. Secondary bacterial infection is common. Some patients develop episodes of pain in the affected area—postherpetic neuralgia after clearance of the rash. Skin lesions and nasopharyngeal secretions can transmit chickenpox.

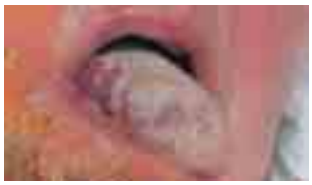


Herpes zoster

Herpes zoster

Herpes zoster—points to note

- Trigeminal zoster may affect:
  the ophthalmic nerve (causing severe conjunctivitis)
  the maxillary nerve (causing vesicles on the uvula or tonsils)
  the mandibular nerve (causing vesicles on the floor of the mouth and on the tongue)
- Disseminated zoster is a severe illness presenting with widespread lesions. Visceral lesions may present with pleuritic or abdominal pain
- Extensive and haemorrhagic vesicles may develop in patients with AIDS



Mandibular zoster

Treatment

Localised lesions of herpes simplex have been treated with a variety of medications from zinc sulphate to iodoxuridine. Topical acyclovir—a drug that inhibits herpes virus DNA polymerase—is effective but only shortens the duration of illness by a day or so. It is useful in primary infection but should be used as soon as the patient is aware of symptoms.

Severe, recurrent, herpes simplex, or herpes zoster can be treated with oral or intravenous aciclovir as early in the course of the illness as possible. Ganciclovir is an alternative.

Secondary infection may require antiseptic soaks, such as 1/1000 potassium permanganate, or topical or systemic antibiotics.

Steroids (prednisolone 40–60 mg/day) given during the acute stage of herpes zoster may diminish pain and postherpetic neuralgia.

Rest and analgesics are recommended treatment for extensive herpes simplex or herpes zoster infections.

Ophthalmic zoster

Pox viruses

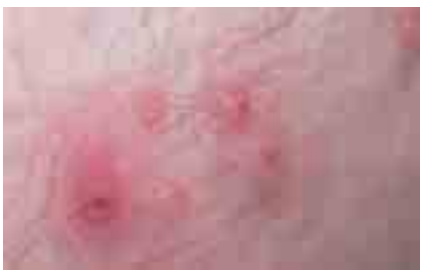
Steroids may cause disseminated infection in immunodeficient patients

The pox viruses are large DNA viruses, with a predilection for the epidermis. Variola (smallpox), once a disease with high mortality, has been eliminated by vaccination with modified vaccinia (cowpox) virus.

Molluscum contagiosum

The commonest skin infection due to a pox virus is molluscum contagiosum, a skin infection seen particularly in children. Despite its name it is not very contagious, but can occur in families.

In adults florid molluscum contagiosum may be an indication of underlying immunodeficiency, as in AIDS patients.



Molluscum contagiosum

Clinical features

The white, umbilicated papules of molluscum contagiosum are characteristic. Large solitary lesions may cause confusion as can secondarily infected, excoriated lesions. These lesions often itch, particularly in patients with atopy. Resolving lesions may be surrounded by a small patch of eczema.

Diagnosis

Diagnosis is usually based on clinical appearances or microscopy of the contents of papules. Sometimes there is confusion with viral warts.

Treatment

Most treatments result in discomfort and may not be tolerated by young children. An antibiotic–hydrocortisone ointment can be used for excoriated lesions. Treatment with liquid

nitrogen is probably the simplest treatment. Other methods include superficial curettage and carefully rotating a sharpened orange stick moistened with phenol in the centre of each lesion.

Other pox virus infections

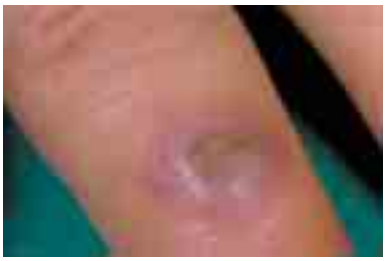
The other pox infections are of incidental interest.

Cowpox only sporadically infects cows from its natural reservoir, probably small mammals, and may affect humans. Papules on the hands enlarge and develop necrosis and crusting.

Milkers' nodules are due to a virus that causes superficial ulcers in cows' udders and calves' mouths. In humans papules form on the hands and develop into grey nodules with a necrotic centre, surrounding inflammation, and lymphangitis. A more generalised papular eruption can occur.

Orf is often recognised in rural areas. It is seen mainly in early spring as a result of contact with lambs. A single papule or group of lesions develops on the fingers or hands with purple papules developing into bulla. This ruptures to leave an annular lesion 1–3 cm in diameter with a necrotic centre. There is surrounding inflammation. The incubation period is a few days and the lesions last two to three weeks with spontaneous healing. Associated erythema multiforme and widespread rashes are occasionally seen.

Cowpox, early stage



Milker's nodule

Wart viruses

A growing recognition that there is an association between human papilloma viruses (HPV), which cause warts, and cancer has led to a renewed interest in these infections. The wart is one of the few tumours in which a virus can be seen to proliferate in the cell nucleus. The different clinical forms of wart are caused by range of HPV, currently divided into over 80 major types. These viruses are also responsible for cervical cancer and have been associated with squamous carcinomas in the immunosuppressed. Warts are classed as cutaneous or mucocutaneous. Epidermodysplasia verruciformis is a rare condition associated with a defect of specific immunity to wart virus. The following aspects should be remembered:

- Genital warts (due to HPV) very rarely undergo malignant change but HPV infection of the cervix, caused by type 16, frequently leads to dysplasia or in some cases malignant changes. Cervical smears must be taken.
- Very extensive proliferation of warts occurs in patients receiving immunosuppressive therapy, such as renal transplant recipients—in whom wart-like lesions can develop into squamous carcinomas—and in patients with AIDS.
- There is an association between HPV infections of the skin in immunosuppressed patients and the subsequent development of atypical-looking squamous carcinomas.
- Epidermodysplasia verruciformis, an unusual widespread eruption of flat erythematous warty plaques, can also develop into carcinoma.

Orf

Common warts

Treatment

Warts commonly occur in children and resolve spontaneously without treatment or with very simple measures. These include paints or lotions containing salicylic and lactic acids in various proportions, which should be applied daily. Salicylic acid (40%) plasters are useful for plantar warts; they are cut to shape and held in place with sticking plaster for two or three days. Glutaraldehyde solution is also used.

For large or painful warts other measures can be used:

- Liquid nitrogen is effective but has to be stored in special containers and replaced frequently. It can be applied with cotton wool or discharged from a special spray with a focused nozzle. Freezing is continued until there is a rim of frozen tissue around the wart but not for more than 30 seconds. Subsequent blistering may occur. Scarring is unusual. Carbon dioxide is more readily available and can be transported in cylinders that produce solid carbon dioxide "snow". The temperature (about −64 °C) is not as low as liquid N_2 (−196 °C)
- Heat cautery causes more scarring and requires local anaesthesia. The diathermy loop is effective for perianal warts.
- Curettage and cautery together are effective but leave scars and the warts may recur.
- Podophyllin, 15–25% in tincture of benzoin compound or alcoholic solution, is effective for genital warts when applied each week. It is, however, toxic when ingested or absorbed, may cause burns, and must never be used in pregnancy.

Treatment of warts with liquid nitrogen

Other treatments include laser therapy, immune enhancement (for example interferon β), and bleomycin injections. However, relapse is common whatever the remedy.

Virus diseases with rashes

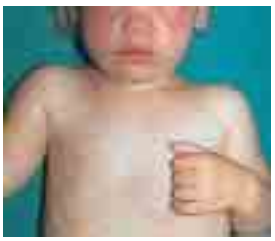
- Measles
- Rubella
- Infectious mononucleosis
- Erythema infectiosum
- Roseola infantum
- Gianotti–Crosti syndrome
- Hand, foot, and mouth disease
- Primary HIV infection

Virus diseases with rashes

Measles and rubella are much less common than previously as a result of widespread immunisation. However, measles is probably the best known example of an exanthem (a fever characterised by a skin eruption. In an enanthem the mucous surfaces are affected.) Other common clinical patterns can then be compared with it. All exanthems, except fifth disease (erythema infectiosum), due to RNA viruses.

Measles

- *Age.* Measles usually affects children, particularly those aged over five years.
- *Incubation* lasts seven to 14 days. Prodromal symptoms include: fever, malaise, upper respiratory symptoms; conjunctivitis; and photophobia.
- *Initial rash.* Early on Koplik's spots (white spots with surrounding erythema) appear on the oral mucosa. After two days a macular rash appears on the face, trunk, and limbs. Look behind the ears for early lesions.
- *Development and resolution.* The rash becomes papular, with coalescence. There may be haemorrhagic lesions and bullae which fade to leave brown patches.
- *Complications* are encephalitis, otitis media, and bronchopneumonia.
- *Diagnosis.* Specific antibodies may be detected; they are at their maximum at two to four weeks.



Measles

Rubella

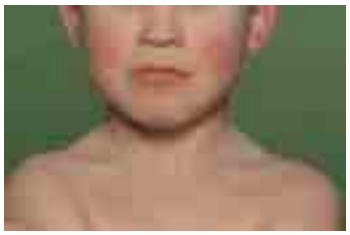
- *Age.* Rubella affects children and young adults.
- *Incubation* lasts 14–21 days.
- *Prodromal symptoms.* There are none in young children. Otherwise fever, malaise, and upper respiratory symptoms occur.
- *Initial rash.* Initially some patients develop erythema of the soft palate and lymphadenopathy. Later pink macules appear on the face, spreading to trunk and limbs over one to two days.

Rubella

- *Development and resolution.* The rash then clears over the next two days, and sometimes no rash develops at all.
- *Complications.* The most important complications are congenital defects in babies of women infected during pregnancy. The risk is greatest in the first month of pregnancy.
- *Diagnosis.* The diagnosis is made from the clinical signs above. Serum should be taken for antibodies and the test repeated at seven to 10 days.
- *Prophylaxis.* Active immunisation is routinely available for all schoolgirls.

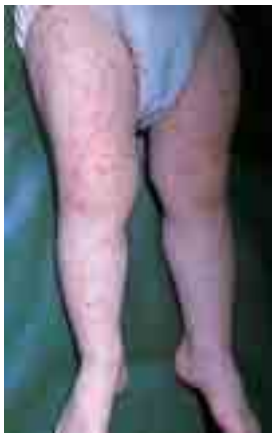
Erythema infectiosum (fifth disease)

- *Age.* Erythema infectiosum affects children aged two to 10 years, mainly girls.
- *Incubation* lasts five to 20 days.
- *Prodromal symptoms.* There are usually none, but there may be a slight fever with initial rash.
- *Initial rash.* The initial rash is a hot, erythematous eruption on the cheeks—hence the "slapped cheek syndrome". Over two to four days a maculopapular eruption develops on the arms, legs, and trunk.
- *Development and resolution.* The rash extends to affect hands, feet, and mucous membranes, then fades over one to two weeks.
- *Diagnosis* is made by finding a specific IgM antibody to parvovirus B19.
- *Complications.* There are no reported dermatological complications but haematological disorders such as thrombocytopenia, arthropathy, and fetal abnormalities may be associated.


Erythema infectiosum

Roseola infantum

- *Age.* Roseola infantum affects infants aged under two years.
- *Incubation* lasts 10–15 days.
- *Prodromal symptoms.* There is fever for a few days.
- *Initial rash.* A rose pink maculopapular eruption appears on the neck and trunk.
- *Development and resolution.* The rash may affect the face and limbs before clearing over one to two days.
- *Diagnosis.* The condition is diagnosed from its clinical features.
- *Complications* include febrile convulsions.


Gianotti–Crosti syndrome

Gianotti–Crosti syndrome

- *Age.* The Gianotti–Crosti syndrome affects children, usually those aged under 14 years.
- *Incubation* period is unknown.
- *Prodromal symptoms.* Lymphadenopathy and malaise accompany the eruption.
- *Initial rash.* Red papules rapidly develop on the face, neck, limbs, buttocks, palms, and soles.
- *Development and resolution.* Over two to three weeks the lesions become purpuric then slowly fade.
- *Diagnosis.* The syndrome may be due to a number of virus infections such as hepatitis B.
- *Complications.* Lymphadenopathy and hepatomegaly always occur and may persist for many months.


Hand, foot, and mouth disease

Hand, foot, and mouth disease

- *Age.* Hand, foot, and mouth disease (Coxsackie virus A) affects both children and adults.
- *Incubation* period is unknown.
- *Prodromal symptoms.* Fever, headache, and malaise may accompany the rash.

- *Initial rash.* Initially there may be intense erythema surrounding yellow-grey vesicles 1–1·5 mm in diameter. These are mainly distributed on the palms and soles and in the mouth. Sometimes a more generalised eruption may develop.
- *Development and resolution.* Over three to five days the rash fades.
- *Diagnosis.* Coxsackie A (usually A16) virus is isolated from lesions and stools. A specific antibody may be found in the serum.
- *Complications* are rare but include widespread vesicular rashes and erythema multiforme.

Other infections

Infectious mononucleosis

As well as the erythematous lesions on the palate a maculopapular rash affecting the face and limbs can occur.

Further reading

Joklik WK. *Virology*, 3rd ed. Norwalk: Appleton and Lange, 1988
Mandell GL, Douglas RG, Bennett JF. *Principles and practice of infectious diseases*, 3rd ed. New York: Churchill Livingstone, 1990
Timbury MC. *Notes on medical virology*, 9th ed. Edinburgh: Churchill Livingstone, 1991
White DO, Fenner FJ. *Medical virology*, 4th ed. New York: Academic Press, 1994

20 AIDS and the skin

MA Waugh

AIDS was first described in 1981 and since then 22 million people have died of the disease. The World Health Organisation estimates that in December 2000, 36 million people were infected with HIV. Of these, 1·5 million were children under the age of fifteen. It is estimated that 5·3 million people were newly infected with HIV in the year 2000 and 3 million have died in the same year. A total of nearly 22 million people have died of AIDS since the start of the epidemic. In the United Kingdom the prevalence of HIV infection is about 30 000. Over 65 000 in Europe are infected.

The human immunodeficiency virus (HIV) is the cause of the acquired immunodeficiency syndrome (AIDS). This virus was first isolated in 1983 in Paris and a second retrovirus, HIV2, was isolated from West Africa in 1986. The virus contains an enzyme that copies viral RNA into the DNA of the host cell in which the HIV virus then persists in the host cells, particularly monocytes, macrophages, and dendritic cells.

Skin lesions in patients with AIDS

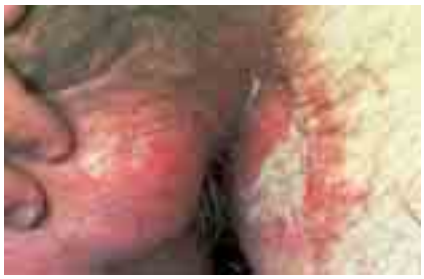
- Skin lesions can develop as a manifestation of primary HIV infection
- Skin lesions can develop as a consequence of immunosuppression (AIDS)

Between these two events there is a latent period that can last from a few months to several years

Stages of AIDS

Primary HIV infection

In 80% of cases there are initial symptoms and signs—"seroconversion illness". There are a variety of symptoms including fever, malaise, headache, nausea, vomiting, and diarrhoea. There is often lymphadenopathy. The skin signs consist of a transient maculopapular eruption associated with erythema and erosions in the mouth in some patients.



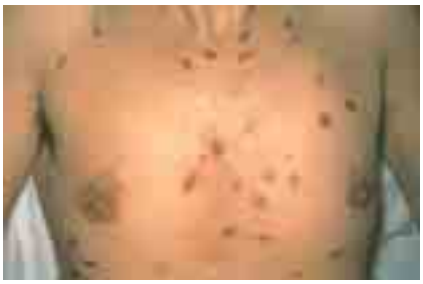
Flexural candidiasis

Early stages

In the early stages 50% of patients have antibiotics to HIV and the p24 antigen can be detected. The proportion of CD4 lymphocytes decreases, and this is associated with the development of secondary changes in the skin. There is also an increase in HIV antibodies so a test for this should be repeated six to eight weeks after the initial illness. Counselling should take place before testing is carried out.

Late stage HIV disease

The skin changes are many and variable. Common inflammatory skin diseases such as psoriasis and seborrhoeic dermatitis will be much more florid. Cutaneous infections are more severe due to the impaired immune response and opportunistic infections also develop. In addition, Kaposi's sarcoma occurs in 34% of homosexual men and in 5% of other cases.



Kaposi's sarcoma

AIDS should therefore be considered in any patient with a florid inflammatory skin disease that is resistant to treatment or severe and extensive infection of the skin.

Skin changes in AIDS

Seborrhoeic eczema
This is common and may be the only evidence of HIV infection initially. It is more extensive and inflamed than usual. The role of *Pityrosporum* organisms is indicated by the response to imidazole antifungal drugs.


Seborrhoeic dermatitis

Psoriasis
Psoriasis is more widespread, severe, and resistant to treatment in patients with late HIV disease. The use of ultraviolet light may lead to an increased risk of Kaposi's sarcoma.

Skin changes in AIDS
- Seborrhoeic eczema
- Psoriasis
- Fungal infections
- Bacterial infections
- Viral infections
- Kaposi's sarcoma
- Drug rashes
- Oral hairy leukoplakia

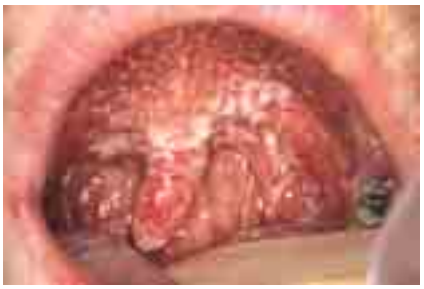
Infections
Any type of opportunistic infection is more likely in patients with AIDS and will generally be more severe. An itching, inflammatory folliculitis occurs in many cases. The cause is unknown, but it is possible that *Demodex* spp. play a part.

Fungal infections
Superficial fungal infections are often much more extensive and invade more deeply into the dermis than usual. There may also be granuloma formation.

Deep fungal infections that are not normally seen in healthy individuals occur in AIDS patients as opportunistic infections. *Cryptococcus neoformans* and *Histoplasma capsulatum* may cause inflammatory papular and necrotic lesions, particularly in the later stages of the disease.

Candidiasis is common and often associated with bacterial infections. It occurs particularly in and around the mouth, on the palate, and in the pharynx. It commonly causes severe vulvovaginitis in infected women.

Pityrosporum organisms occur more frequently and may produce widespread pityriasis versicolor on the trunk or extensive folliculitis.


Pseudomembranous candida

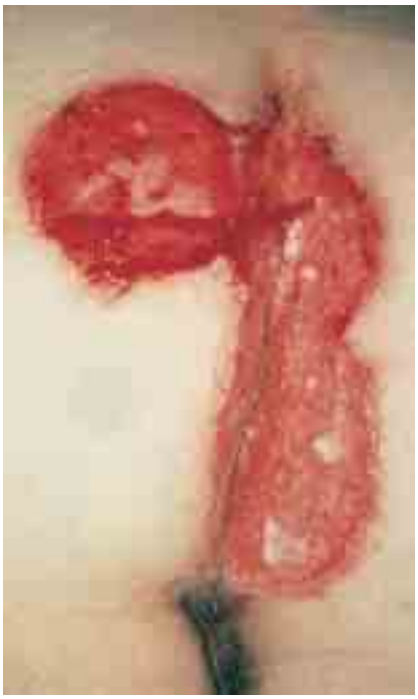
Bacterial infections
Impetigo may be severe, with particularly large bullous lesions occurring.

Mycobacteria may produce widespread cutaneous and systemic lesions. Varieties of mycobacteria that do not normally infect the skin may cause persistent necrotic papules or ulcers.

Viral infections
Both herpes simplex and herpes zoster infections may be unusually extensive, with large individual lesions. In the case of herpes zoster the affected area may extend beyond individual dermatomes. Sometimes persisting ulcerated lesions occur.

Molluscum contagiosum lesions are frequently seen. They are much larger than usual and develop over quite large areas of skin. They are readily identified as small, firm papules with an umbilicated centre. When very large individual molluscum lesions occur they may be due to localised fungal infection, particularly cryptococcus and histoplasmosis.

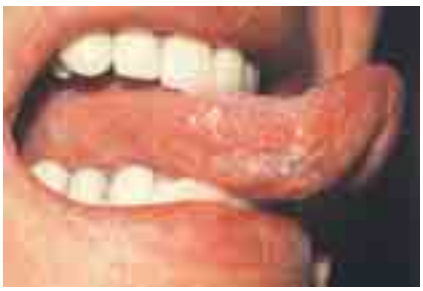
Viral warts may be large and extensive. Perianal and genital warts due to the human papilloma virus (HPV) are common and may be associated with intraepithelial neoplasia of the cervix and sometimes invasive perianal squamous cell carcinoma. The warts tend to become smaller as the immune status of the patient improves with the treatment. It is not unusual for florid viral warts to develop in the mouth.


Aciclovir–resistant perianal herpes simplex infection

Other manifestions

Oral hairy leukoplakia occurs in 30–50% of patients with AIDS. It is characterised by an overgrowth of epithelial plaques on the sides of the tongue with a verrucous surface and a grey/white colour. It is believed to be due to a proliferation of the Epstein–Barr virus.

Infestations with various organisms is not uncommon and the severe widespread crusted lesions of Norwegian scabies may occur.

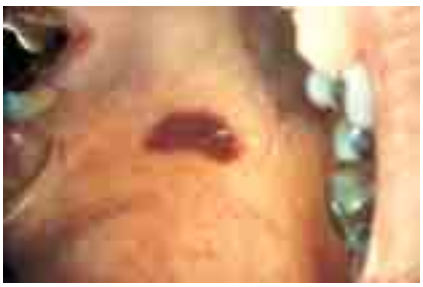


Oral hairy leukoplakia

Kaposi's sarcoma

Kaposi's sarcoma is associated with the later stages of AIDS but can occur earlier. It is associated with herpes virus type 8. It often presents with small polychromic macules on the face, palate, trunk or groin which vary from red and purple to brown. They then develop into larger livid plaques, involving the trunk, limbs, and face, and also the oral mucosa. They are most common on the palate and nose. Sometimes the lesions are very aggressive.

B cell lymphoma may occur in the skin in 10% of AIDS patients. There is also an increased incidence of basal cell and squamous carcinomas.



Kaposi's sarcoma of the hard palate

Drug rashes

Reactions to sulphonamides and antibiotics are not uncommon, usually presenting as a maculopapular eruption. Occasionally this can be severe and associated with Stevens–Johnson syndrome. Myopathy may occur as a reaction to zidovudine.

All the above manifestions of AIDS become less marked as the CD4 count improves with treatment.

AIDS may thus present with a wide variety of skin conditions, commonly with several present at the same time. Any unusually florid skin condition that is resistant to treatment should raise the suspicion that HIV infection may be present.

Further reading

Adler MW. *ABC of AIDS*, 4th ed. London: BMJ Publishing Group, 1997

Penneys NS. *Skin manifestations of AIDS*, 2nd ed. St Louis: Mosby, 1995

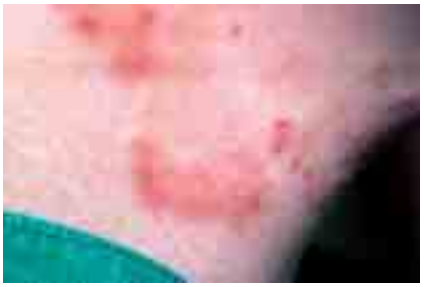
21 Fungal and yeast infections

RJ Hay

Fungal infections

The common fungal infections of the skin are dermatophytosis or "ringworm", superficial candidiasis, and *Malassezia* infections. There are two growth forms of fungi, moulds, and yeasts. Mould fungi produce thread-like hyphae that comprise chains of cells. In dermatophyte fungal infection of the skin, hair, and nails these hyphae invade keratin and are seen on microscopic examination of skin, hair, or nails from infected tissues. Vegetative spores (conidia) develop in culture, and their distinctive shape helps to identify the different species. Skin scrapings or clippings from infected nails can be easily taken and should always be sent to the laboratory for mycological examination and culture in any patient suspected of having a fungal infection.

In yeast infections such as those due to candida, the fungal cells are individual and separate after cell division by a process called budding. In systemic, or deep, fungal infections subcutaneous on deep visceral structures are attacked. However skin involvement can also occur following blood stream dissemination and such lesions may provide a clue to the diagnosis.



Tricophyton rubrum infection of the neck

Why should one suspect a lesion to be due to a fungus?

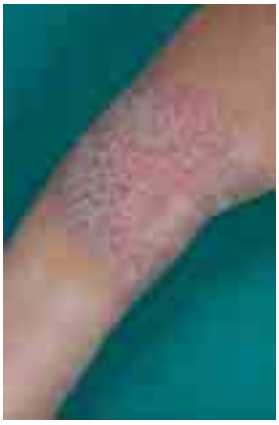
Clinical presentation

Fungal infections usually itch. Those due to zoophilic (animal) fungi generally produce a more intense inflammatory response with deeper indurated lesions than fungal infections due to anthropophilic (human) species. Some lesions, usually those on the trunk, have a prominent scaling margin with apparent clearing in the centre. Hence the name "ringworm".

Children below the age of puberty are susceptible to scalp ringworm and anthropophilic fungi (from humans) have become common in some inner city areas. They can also be infected with zoophilic fungi (from animals), particularly cattle, dogs, and cats. Cattle ringworm can cause an intense inflammatory response in children, producing a "kerion" described below. They rarely develop anthropophilic fungal infection.

Adults. From adolescence onwards infection of the feet is a common occurrence. Tinea cruris in the groin is seen mainly in men and fungal nail infections (onychomycosis) have become particularly common.

Infection from dogs and cats with a zoophilic fungus (*Microsporum canis*) to which humans have little immunity can occur at any age. A patient returned from a skiing holiday with intensely itchy "eczema", which refused to clear. A stray kitten, mewing outside in the dark, had been taken indoors, warmed in their sleeping bags, and infected the whole party with *M. canis*.



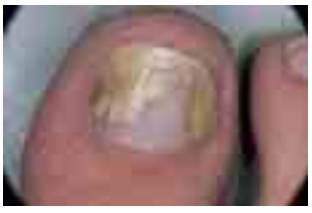
Animal ringworm

Tinea cruris

Nail infections

These occur mainly in adults, usually in their toenails, especially when traumatised—for example the big toes of

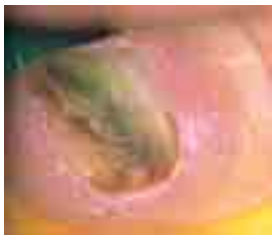
Microsporum canis



Fungal infection of nail

footballers. The nails become thickened and yellow and crumble, usually asymmetrically. The changes occur *distally* and move back to the nailfold. In psoriasis of the nail the changes occur *proximally* and tend to be symmetrical and are associated with pitting and other evidence of psoriasis elsewhere.

Chronic paronychia occurs in the fingers of individuals whose work demands repeated wetting of the hands: housewives, barmen, dentists, nurses, and mushroom growers, for example. Other predisposing factors include diabetes, poor peripheral circulation, and removal of the cuticle. There is erythema and swelling of the nail fold, often on one side with brownish discoloration of the nail. Pus may be exuded. The cause is *Candida albicans* (a yeast) together with secondary bacterial infection.

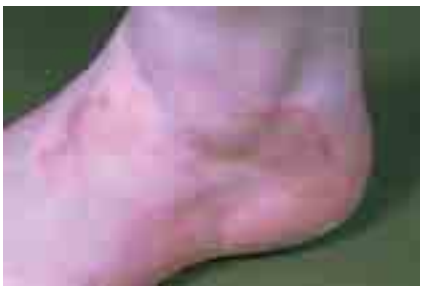

Chronic paronychia

Pushing back the cuticles should be avoided—this is commonly a long term condition, lasting for years. The hands should be kept as dry as possible, an azole lotion applied regularly around the nail fold, and in acute flares a course of erythromycin prescribed.

Feet

Tinea pedis, or athletes foot, is a common disease and its prevalence increases with age. It is easily acquired in public swimming pools or showers and industrial workers appear to be particularly predisposed to this infection. The hands may be affected.

In interdigital tinea pedis the itching, macerated skin beneath the toes is familiar, but when a dry, scaling rash extends across the sole and dorsal surface of the foot (dry type tinea pedis) the diagnosis may be missed. The condition needs to be differentiated from psoriasis and eczema.


Tinea pedis

Hands

Dermatophyte infections often produce a dry rash on one palm. There may be a well defined lesion with a scaling edge.

Trunk

Tinea corporis presents with erythema and itching and a well defined scaling edge. In the groin, tinea cruris, the infection may spread to the adjacent skin on the thighs and abdomen. Intense erythema and satellite lesions suggest a candida infection. In the axillae erythrasma due to *Corynebacterium minutissimum* is more likely. It does not respond to antifungal treatment but clears with tetracycline by mouth.


Tinea corporis


Erythrasma

Tinea versicolor affects the trunk, usually of fair skinned individuals exposed to the sun. It affects mainly the upper back, chest, and arms. Well defined macular lesions with fine scales develop, which tend to be white in suntanned areas and brown on pale skin. It may be confused with seborrhoeic dermatitis, pityriasis rosea, and vitiligo. In skin scrapings the causative


Tinea versicolor


Pityrosporum organisms

organisms, *Malassezia* spp., normally found in hair follicles, can be readily seen.

Scalp and face

Scalp ringworm in children may be caused by anthropophilic fungi such as *Trichophyton tonsurans*, which is spreading in cities in the United Kingdom, or *Microsporum audouinii*. Sporadic cases are caused by *M.canis* which is acquired from cats or dogs. In all cases there is itching, hair loss, and some degree of inflammation which is worse with *M. canis* infections.

Kerion, an inflamed, boggy, pustular lesion, is often due to cattle ringworm and is fairly common in rural areas. It is often seen in children in the autumn when the cows are brought inside for the winter.

Tinea incognito is the term used for unrecognised fungal infection in patients treated with steroids (topical or systemic). The normal response to infection (leading to erythema, scaling, a raised margin, and itching) is diminished, particularly with local steroid creams or ointments. The infecting organism flourishes, however, because of the host's impaired immune response—shown by the enlarging, persistent skin lesions. The groins, hands, and face are sites where this is most likely to occur.

Seborrhoeic dermatitis of adults may also be caused by *Malassezia*.

Microsporum ringworm

Microsporum—Wood's light

Tinea incognito

Yeast infections

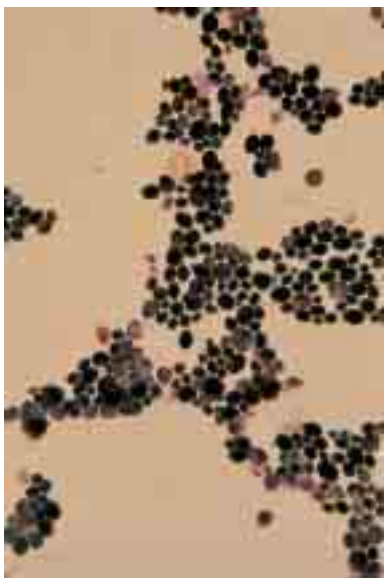
Candida infection may occur in the flexures of infants and elderly or immobilised patients, especially below the breasts and folds of abdominal skin. It needs to be differentiated from: (a) psoriasis, which does not itch; (b) seborrhoeic dermatitis, a common cause of a flexural rash in infants; and (c) contact dermatitis and discoid eczema, which do not have the scaling margin. Candida intertrigo is symmetrical and "satellite" pustules or papules outside the outer rim of the rash are typical. Yeasts, including *Candida albicans*, may be found in the mouth and vagina of healthy individuals. Clinical lesions in the mouth—white buccal plaques or erythema—may develop. Predisposing factors include: general debility, impaired immunity (including AIDS), diabetes mellitus, endocrine disorders, such as Cushing's syndrome, and corticosteroid treatment. Vaginal candidosis or thrush is a common infection of healthy young women; an underlying predisposition is rarely found. The infection presents with itching, soreness, and a mild discharge.

Candida albicans

Deep fungal infection

Fungal infections of the deeper tissues are only rarely associated with skin lesions in the United Kingdom, except in patients with AIDS. Some infections that involve deep tissue, histoplasmosis, cryptococcosis, and infections due to *Penicillium marneffei*, can present with skin lesions. In an HIV positive patient lesions resembling molluscum contagiosum may be the earliest feature of deep fungal infections.

In tropical countries deep fungal infections are more common. These are described in chapter 23. They should be considered in any patient from a tropical country with chronic indurated and ulcerating lesions.



Candida organisms

Treatment

Topical treatment

The most commonly used treatments are the imidazole preparations, such as clotrimazole and miconazole (two to

four weeks) and also topical terbinafine (one to two weeks). The polyenes, nystatin, and amphotericin B are also effective against yeast infection. For damp macerated skin dusting powders may be helpful. In toe web infections a mixture of micro-organisms including dermatophytes and Gram negative bacteria may be present and both require treatment.

Systemic treatment

It is important to confirm the diagnosis from skin scrapings before starting treatment. Terbinafine is a very effective fungicidal drug. It is taken in a dosage of 250 mg once daily for two to six weeks for skin infections, six weeks for finger nail or three months for toenail infections. It is only approved for use in children in some countries. Blood monitoring is only advised in patients with liver disease or impaired renal function. Pregnancy and lactation are relative contraindications. There have been reports of headaches, taste disturbances and, very rarely, liver dysfunction.

Triazole preparations such as itraconazole are effective in both dermatophyte and yeast infections. Cases of liver damage have rarely been reported. Fluconazole is effective in yeast infections. Some drugs interact with azole drugs, the main ones being terfenadine, astemizole, digoxin, midazolam, cyclosporin, tacrolimus, and anticoagulants.

Griseofulvin is mainly used for tinea capitis. The duration of treatment is six to eight weeks for infections of the scalp. The dose is 10–20 mg/kg for children, taken with food. Contraindications to griseofulvin are severe liver disease and porphyrias. The drug interacts with the coumarin anticoagulants.

In countries without access to these drugs simple measures such as antiseptic paints—Neutral Red or Castellan's paint—can be used. Whitfield's ointment (benzoic acid ointment) is easily prepared and is reasonably effective for fungal infections.

Principles of diagnosis and treatment

- Consider a fungal infection in any patient where isolated, itching, dry, and scaling lesions occur without any apparent reason—for example, if there is no previous history of eczema. Lesions due to fungal infection are often asymmetrical
- Skin scrapings or clippings should be sent to the laboratory from lesions, nails, or hair. The skin scales should be removed by scraping the edge of the lesion with a scalpel held at right angles to the skin on to a piece of dark paper—transport packs are available commercially. Clippings can be taken from the nails and as much material as possible should be taken from the nail including subungual debris. Laboratories will report first on the direct microscopy of the material examined after treatment with 10% potassium hydroxide but culture results take at least two weeks
- Lesions to which steroids have been applied are often quite atypical because the normal inflammatory response is suppressed—*tinea incognito*. The patient often states that the treatment controls the itch but the rash persists and may change into a tender form of folliculitis. In such cases microscopy of lesions is usually strongly positive
- Wood's light (ultraviolet light filtered through special glass) can be used to show *Microsporum* infections of hair, as they produce a green-blue fluorescence

Further reading

Midgley G, Clayton Y, Hay RJ. *Medical mycology*. London and St Louis: Mosby-Wolfe, 1997

Elewski BE. *Cutaneous fungal infections*. New York: lgaku-Shoin, 1992

Evans EG, Richardson MD. *Medical mycology: A practical approach*. Oxford: Oxford University Press, 1989

22 Insect bites and infestations

So, naturalists observe, a flea
Hath smaller fleas that on him prey;
And these have smaller fleas to bite 'em.
And so proceeds "Ad infinitum"
Jonathan Swift

It is, of course, the internal parasites of biting insects that cause trouble for humans, rather than "smaller fleas" on their surface.

An ornithologist went bird watching in Guyana, where he sustained widespread "midge bites" on the arms. He was referred on account of nodules that developed a few weeks later, then enlarged and ulcerated. Other lesions occurred further up the arms with regional lymphadenopathy. A biopsy specimen showed histiocytic inflammatory changes, and *Leishmania braziliensis* was isolated from smears; the midges (phlebotomus or sand fly) had acquired the protozoon while feeding on local rodents and transferred it into the ornithologist's skin.

Serious disease from insect vectors is rare in residents of most Western countries but, as in the patient described above, must be considered in those returning from tropical and subtropical countries.


Body louse


Sand fly

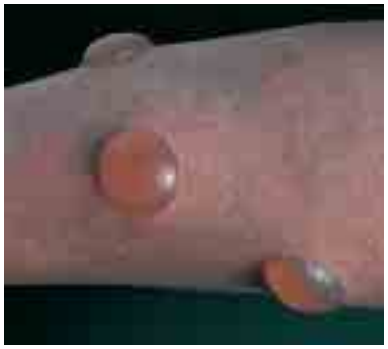

Leishmaniasis

Some diseases with skin lesions resulting from insect bites

Condition	Appearance	Organism	Vector
Cutaneous leishmaniasis	Chronic enlarging nodules with ulceration	Leishmania protozoon (*L. braziliensis*)	Sand fly
Oriental sore	Ulcerating nodules	Leishmania (*L. tropica*)	Sand fly
Kala-azar	Hypopigmented, erythematous, and nodular lesions	Leishmania (*L. donovani*)	Insect vectors
Onchocerciasis	Pruritic nodules	Filaria (*Onchocerca volvulus*)	Black fly (Simuliidae)
Typhus, human	Erythematous rash and systemic illness	Rickettsia (*R. prowazekii*)	Human louse
Typhus, murine		(*R. mooseri*)	Rat flea
Rocky Mountain spotted fever	Maculopapular rash and fever	Rickettsia (*R. rickettsii*)	Ticks
Rickettsial pox	Vesicular eruption like chickenpox	Rickettsia (*R. akari*)	House mouse, louse
Tick typhus	Necrotic lesions, maculopapular rash, and fever	Various rickettsias	Ticks
Scrub typhus	Fever, lymphadenopathy, maculopapular rash	Rickettsia (*R. tsutsugamushi*)	Mites
Relapsing fever	Widespread maculopapular lesions	*Borrelia recurrentis*	Lice, ticks
Lyme disease	May be annular	*Borrelia burgdorferi*	Ticks, black fly
Yellow fever and dengue	Flushing of face, scarlatiniform rash	Arbovirus	Aedes mosquito

Most cases of bites from fleas, midges, and mosquitoes are readily recognised and cause few symptoms apart from discomfort. Occasionally an allergic reaction confuses the picture, particularly the large bullae that can occur from bites on the arms and legs. It may be difficult to persuade patients that their recurrent itching spots are simply due to flea bites and the suggestion may be angrily rejected.

Nevertheless, some patients are convinced that they have an infestation when they do not. Often they will bring small packets containing "insects". Examination shows these to be small screws of wool, pickings of keratin, thread, and so on. Sympathy and tact will win patients' confidence; derision and disbelief will merely send them elsewhere for a further medical opinion. Antipsychotic drugs may help to dispel the delusion of parasitic infestation (delusional parasitosis) and should be used in conjunction with advice from a psychiatrist if possible. These drugs should be used with care and with full awareness of their side effects, particularly in patients with cardiovascular disease and a history of epilepsy. Pimozide has been used in the past but because of its side effects risperidone is preferred.

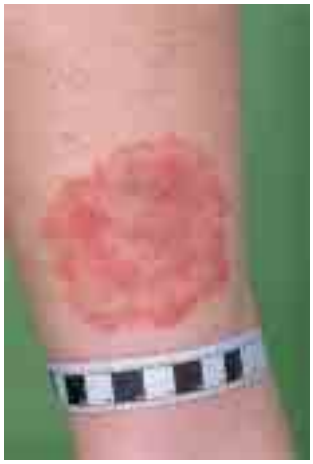


Bullae caused by insect bites

Parasitophobia specimens

- Flea bites, including those from *Cheyletiella* mites in dogs and cats, occur in clusters, often in areas of close contact with clothing, for example, around the waist.
- Grain mites (*Pyemotes*) and harvest mites (*Trombiculidae*) can cause severe reactions.
- Tick, and possibly mosquito, bites can produce infection with *Borrelia burgdorferi*, causing arthropathy, fever, and a distinctive rash (erythema chronicum migrans)—Lyme disease. The condition responds rapidly to treatment with penicillin. Increasing numbers of cases are being reported in the United Kingdom.

Bites on ankles



Erythema chronicum migrans

Papular urticaria

Persistent pruritic (itching) papules in groups on the trunk and legs may be due to bites from fleas, bed bugs, or mites.

A seasonal incidence suggests bites from outdoor insects, while recurrence of the papules in a particular house or room suggests infestations with fleas. The term is sometimes used for other causes of itchy skin.

Harvest mites

Papular urticaria

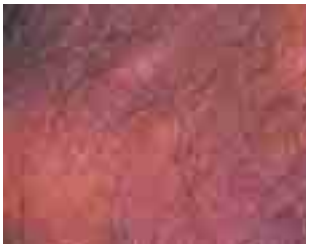
Spider bites

In Europe spider bites rarely cause problems, but sometimes noxious species arrive in consignments of tropical fruit. The patient shown had been bitten by a spider the day before leaving Nigeria and developed a painful necrotic lesion.

Spider bite (Nigeria)

Bites from the European tarantula are painful but otherwise harmless.

In tropical and subtropical countries venomous spiders inject neurotoxins that can be fatal. The "black widow" (*Latrodectus mactans*), "fiddleback" (*Loxosceles veclusa*), and *Atrax* species of Australia are better known examples. Scorpions cause severe local and systemic symptoms as a result of stings (not bites).



Persistent papules in scabies

Infestations

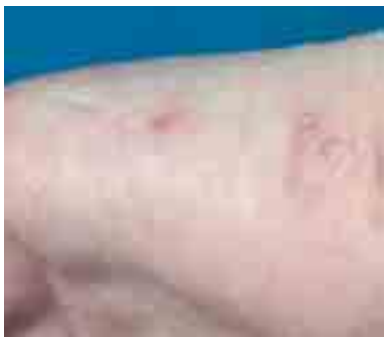
Scabies

The commonest infestation encountered is scabies, and it is easily missed or misdiagnosed. Scabies is due to a small mite, *Sarcoptes scabiei*. The female mite burrows into the stratum corneum to lay

her eggs; the male dies after completing his role of fertilisation, and the developing eggs hatch into larvae within a few days. Intense itching occurs some two weeks later, during which time extensive colonisation may have occurred. The infestation is acquired only by close contact with infected people.

Diagnosis

Finding a burrow—the small (5–10 mm long) ridge, often S shaped—can be difficult as it is often obscured by excoriation from scratching. Without finding a burrow, however, the diagnosis remains uncertain. Isolation of an acarus with a needle or scalpel blade and its demonstration under the microscope convinces the most sceptical patient. Always ask whether there are others in the patient's household and if any of them are itching.



Burrows of scabies

Treatment

10% sulphur in yellow soft paraffin is traditional, effective, and safe. There are several more modern treatments, including 25% benzyl benzoate emulsion, 0·5% malathion cream, 1% gamma benzene hexachloride (lindane) lotion, and 1% permethrin. In children benzyl benzoate should be diluted to 10% and used with care as toxicity results from absorption. In infants over two months old permethrin or 2·5% sulphur ointment can be used. Gamma benzene hexachloride should not be given to children under 10 years or pregnant women in the first trimester. Important points are:

(1) The patient should wash well: a hot bath was formerly advocated but it is now known that this may increase absorption through the skin.
(2) The lotion should be applied from the neck down, concentrating on affected areas and making sure that the axillae, wrists, ankles, and pubic areas are included. If there is any doubt about the thoroughness of application the process should be repeated in a few days.
(3) All contacts and members of the patient's household should be treated at the same time.
(4) Residual papules may persist for many weeks. Topical steroids can be used to relieve the itching.
(5) Secondary infection as a result of scratching may need to be treated.

Scabies—points to note

- There may be very few burrows, though the patient has widespread itching
- The distribution of the infestation is characteristically the fingers, wrists, nipples, abdomen, genitalia, buttocks, and ankles
- Close personal contact is required for infestation to occur, for example, within a family, through infants in playgroups, and through regular nursing of elderly patients
- Itching may persist even after all mites have been eliminated; itching papules on the scrotum and penis are particularly persistent

Demodex

Demodex folliculorum is a small mite that inhabits the human hair follicle, the eggs being deposited in the sebaceous gland. It is found on the central area of the face, chest, and neck of adults. It may have a role in the pathogenesis of rosacea, in which it may be found in large numbers. It may be associated with a pustular eruption round the mouth and blepharitis.

Larva migrans

The patient in the illustration had been on holiday at a tropical coastal town and regularly visited a beach frequented by dogs. Two weeks after returning to Britain he started itching on the buttocks and subsequently noticed a linear, raised area—a condition known as larva migrans, due to the larvae of the hookworm of dogs and cats, *Ancylostoma caninum*. The ova are shed in the faeces and in a warm moist environment hatch into larvae that invade "dead end" hosts. They do not develop any further, so systemic disease does not occur.

Larva migrans

Treatment
This is either by freezing the advancing end of the lesion with liquid nitrogen or by applying thiabendazole (10%) suspension. Similar lesions in patients returning from tropical countries raise the possibility of larva migrans from strongyloides infestation, myiasis from the larvae of flies, or gnathostomiasis.

Visceral larva migrans caused by *Toxocara canis* and *Ascaris lumbricoides* may produce a transient rash.

Pediculosts (lice)
Infestation with lice became less common in the postwar years, but the incidence has recently increased.

There are three areas of the body usually affected by two species of wingless insects—*Pediculus humanus*, infecting the head and body, and *Phthirus pubis*, the pubic louse. The wingless insects feed on blood aspirated at the site of the bite, and each female lays 60–80 encapsulated eggs attached to hairs—"nits" in common parlance.

Head lice are transmitted via combs, brushes, and hats, being more common in girls than boys. The infestation is heaviest behind the ears and over the occiput. If the eyelashes of children are affected this is with "crab lice" (*Phthlrus pubis*); it is not pediculosis.

Body lice are less common in western Europe. Transmission is by clothing and bedding, on which both lice and their eggs may be found in the seams. Poor hygiene favours infestation.

Pubic lice infestation occurs worldwide and is generally transmitted by sexual contact. Infestation of eyelashes may occur with poor hygiene.

As a result of scratching there may be marked secondary infection that obscures the underlying infestation.

Treatment
Gamma benzene hexachloride 1% is usually effective as a single application. Permethrin can also be used.


Pediculosis capitis


Head lice and nits


Phthirus pubis on eyelashes

Further reading

Alexander JO. *Arthropods and human skin*. Berlin: Springer-Verlag, 1984

Busvine JR. *Insects and hygiene*, 3rd ed. London: Chapman and Hall, 1980

Marshall AG. *The ecology of ectoparasitic insects*. New York: Academic Press, 1981

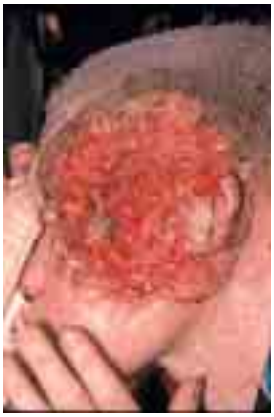
Parish CL, Nutting WB, Schwartzman RM. *Cutaneous infestation of man and animal*. New York: Greenwood, 1983

23 Tropical dermatology

B Leppard

The purpose of this chapter is to provide an overview of tropical diseases that most commonly affect the skin. This will be useful for health workers who may not be familiar with tropical diseases and also as a guide to help those who are already working in the tropics and who see them all the time.

Skin disease is extemely common in the tropics, affecting up to 50% of the population. Most are infections or infestations such as impetigo, ringworm, and scabies. These can easily be treated but continue to be common because of overcrowding, poverty, and the lack of resources given to health care (training of health personnel and lack of basic medicines). To a large extent such diseases can be controlled with very simple measure suitable for use by those with minimal training. Atopic eczema is just as common in urban areas in the tropics as in the west. Skin cancers are uncommon in those with a black skin because of the protective effect of melanin, but are common in albinos.

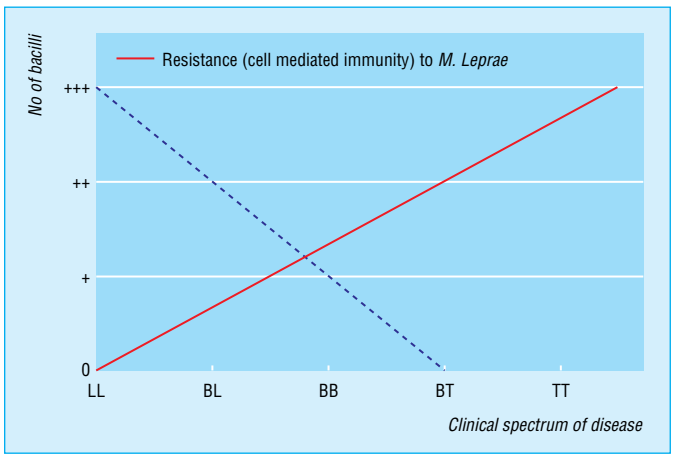


Albino with squamous cell carcinoma

The spectrum of tropical dermatology

All the common inflammatory dermatoses occur in the tropics but may have a different appearance in pigmented skin. Erythema, readily visible in Caucasians, will not be so apparent in black skin.

Infections and infestations occurring in the tropics produce distinctive skin changes These may be due to the presence of the organism, ova, or larvae in the skin. In other diseases a reaction to the organism produces a rash.



Spectrum of clinical disease in leprosy. (BB = borderline leprosy, BL = borderline lepromatous leprosy, BT = borderline tuberculoid leprosy, LL = lepromatous leprosy, TT = tuberculoid leprosy)

Bacterial infections

Impetigo is particularly prone to occur in the tropics and may complicate any area of minor trauma to the skin. It is characterised by erythema, and exudative lesions forming crusts.

Bullae may develop. If possible swabs should be taken for bacteriology and the appropriate antibiotics given.

Erysipelas is a localised streptococcal infection with erythema and tenderness accompanied by fever and malaise. Treatment is with penicillin.

Leprosy

Leprosy is a chronic infection of the skin and nerves by *Mycobacterium leprae*. It is spread by droplet infection and has a long incubation period (anything from two months to 40 years). There is a spectrum of clinical disease depending on the patient's cell mediated immunity to the organism.

Diagnosis

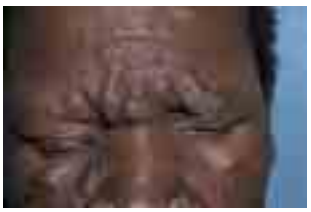
- *Typical clinical findings:*
  (a) In tuberculoid leprosy (TT) there is a single anaesthetic patch or plaque with a raised border.

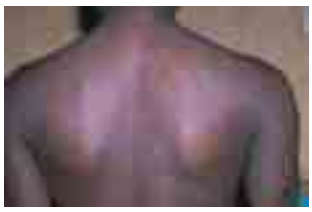
Tuberculoid leprosy

Tuberculoid leprosy

(b) In lepromatous leprosy (LL) there are widespread symmetrical shiny papules, nodules, and plaques which are not anaesthetic.
(c) In borderline leprosy (BT, BB, BL) there are varying numbers of lesions, few in BT and numerous in BL. They may be widespread but are asymmetrical.
(d) Palpably enlarged cutaneous nerves (great auricular nerve in the neck, the superficial branch of the radial nerve at the wrist, the ulnar nerve at the elbow, the lateral popliteal nerve at the knee, and the sural nerve on the lower leg).
(e) Glove and stocking sensory loss causing blisters, ulcers or both on anaesthetic fingers or toes.
(f) Deformity due to invasion of the peripheral nerves with leprosy bacilli, a leprosy reaction or recurrent trauma to anaesthetic limbs.

- *Slit skin smears* measure the numbers of bacilli in the skin (Bacterial Index (BI)) and the % of these that are living (Morphological Index (MI)).


Lepromatous leprosy

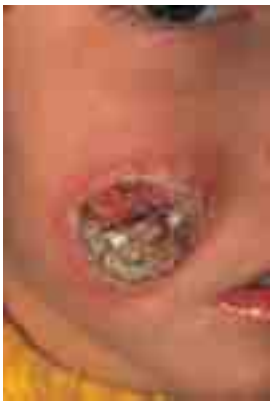

Borderline leprosy

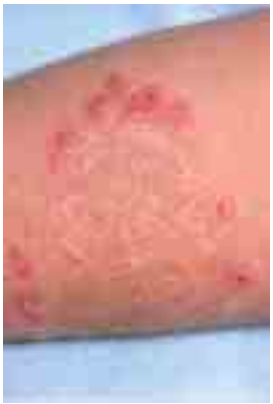

Borderline tuberculoid leprosy

Treatment

Paucibacillary leprosy (BI of 0 or 1+):
Rifampicin 600 mg once a month (supervised) } for
Dapsone 100 mg daily } six months

Multibacillary leprosy (BI of 2+ or more):
Rifampicin 600 mg once a month (supervised)
Clofazamine 300 mg once a month (supervised)
Clofazamine 50 mg/day
Dapsone 100 mg/day
} for two years


Acute leishmaniasis


Chronic leishmaniasis

Cutaneous leishmaniasis

Cutaneous leishmaniasis is due to the protozoa *Leishmania tropica*, and is transmitted by the bite of a sandfly. There is a spectrum of disease depending on the patient's immunity.

Acute leishmaniasis

A red nodule like a boil occurs at the site of the bite. It enlarges, may or may not ulcerate, and heals spontaneously after about one year leaving a cribriform scar.

Chronic leishmaniasis

In a patient with good cell mediated immunity, after the acute leishmaniasis has healed, new granulomata appear at the edge of the scar; these do not heal spontaneously.

Diffuse cutaneous leishmaniasis

This is leishmaniasis in a patient with no immunity to the organism (equivalent to lepromatous leprosy). Extensive skin nodules occur that are full of organisms.


Diffuse cutaneous leishmaniasis

Treatment

Intramuscular injections of sodium stibogluconate 10 mg/kg body weight daily until healing occurs.

Superficial fungal infections

The same fungal infections of the epidermis occur in the tropics as in temperate climates only more so—heat and occlusion of clothing leading to maceration of the skin in which fungi thrive, so expect to see more florid lesions. There are also many fungal infections that are specifically found in the tropics. These include tinea imbricata, tinea nigra, piedra, and favus.

Bacterial Index (BI)

0 = No bacilli seen
1+ = 1–10 bacilli in 100 oil immersion fields
2+ = 1–10 bacilli in 10 oil immersion fields
3+ = 1–10 bacilli in 1 oil immersion field
4+ = 10–100 bacilli in an average oil immersion field
5+ = 100–1000 bacilli in an average oil immersion field
6+ = >1000 bacilli in an average oil immersion field

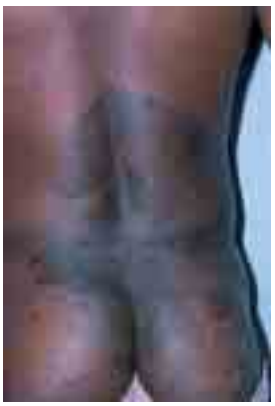
Suspect a dermatophyte fungal infection in any chronic, itching, scaling, slowly developing lesion with epidermal changes.

Tinea imbricata due to *Tricophyton concentricum* is characterised by superficial concentric scaling rings spreading across the trunk. It occurs mainly in Asia but also in other tropical areas.

Tinea nigra occurs in the tropical areas of America, Asia, and Australia. Brown or black macules are seen on the palms and soles. It is due to *Cladosporium werneckii*.

Piedra is a fungal infection of the hair producing hard nodular lesions on the hair shaft. The lesions may be black (due to *Piedra hortai*) or white piedra (due to *Trichosporum beigelii*).

Favus is widespread throughout the Mediterranean, the Middle East, and tropics, but is rare in Africa. It is due to an endothrix fungus—*Trichophyton schoenleinii*—which causes a thick yellow crust with an unpleasant odour. Erythematous areas of scarring occur that must be differentiated from lichen planus and other causes of scarring alopecia.



Superficial fungal infection

Deep fungal infections

In these conditions there is chronic inflammation in the subcutaneous tissues leading to granulomatous and necrotic nodules.

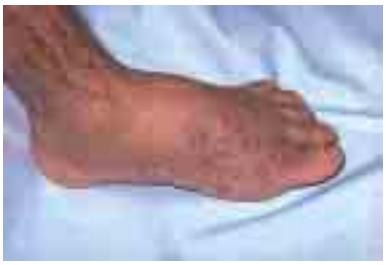
Tinea imbricata

Mycetoma (Madura foot)

This is a chronic infection of the dermis and subcutaneous fat caused by various species of fungus (*eumycetoma*) or bacteria (*actinomycetoma*). Both types look the same with a swollen foot and multiple discharging sinuses, but it is important to differentiate between them because the treatment is different.

Diagnosis

- Examination of the discharging grains (colour will give a clue as to the cause).
- Culture of the grains to identify the causative fungus or bacteria.
- If no grains can be found a skin biopsy will show them.



Madura foot

Treatment

Eumycetoma:

- Itraconazole 200 mg twice daily for at least 12 months if it is affordable.

If not

- Surgical excision of affected tissue if disease is limited.
- Amputation if extensive.

Actinomycetoma:

- Sulfamethoxazole-trimethoprim mixture 960 mg twice daily for up to two years.

Blastomycosis

This condition is caused by the invasion of lymphatic system, lungs, and skin by *Paracoccioddes brasiliensis*. The widespread cutaneous lesions, which vary in appearance and distribution, must be differentiated from tuberculosis and other mycoses such as sporotrichosis, chromomycodis, and coccidiomycosis. It occurs in central and south America.

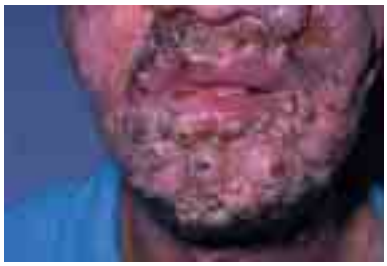
Chromomycosis
This chronic granulomatous condition mainly affects the legs and results from infestation by a variety of parasitic fungi. Large verrucous plaques may require surgical removal.


Chromomycosis

Histoplasmosis
This occurs in West Africa with nodules, ulcers, and bone lesions developing due to infection with *Histoplasma duboisii*. Treatment is with aphotericin B.

Infestations

Tungiasis
Invasion of the skin by sand fleas (*Tunga penetrans*) causes tungiasis in tropical areas of Africa, America, and India. It is most common on the feet, especially under the toes and toenails. The condition looks a bit like plantar warts, but if you watch for a while you will see the eggs being squirted out.


Histoplasmosis in HIV infection

Prevention
Wear shoes.

Treatment
- Carefully winkle the fleas out with a pin (most patients know how to do this themselves).
- If the fleas are very extensive, soak the feet in kerosene or treat with a single dose of ivermectin 200 micrograms/kg body weight.


Tungiasis

Subcutaneous myiasis
Invasion of the skin by the larvae of the tumbu (mango) fly (*Cordylobia anthropophaga*) in central and southern Africa causes this condition. The fly lays her eggs on clothes layed out to dry on the ground. The eggs hatch out two days later on contact with the warm skin when the clothes are put on. The larvae burrow into the skin causing a red painful or itchy papule or nodule, predominantly on the trunk, buttocks, and thighs.

Other flies that cause miyasis are:

- *Dermatobia hominis*—tropical bot fly, in Mexico, central, and south America with tender nodules developing on the scalp, legs, forearms, and face.
- *Aucheronia* sp.—Congo floor maggot, in central and southern Africa. Bites of the larvae cause intense irritation.
- *Callitroga* sp. in central America causing inflamed lesions with necrosis.


Myiasis—larvae

Prevention
Iron the clothes before wearing them.

Treatment
Cover the nodule with petroleum jelly or other grease; the larva will be unable to breath and will crawl out.


Myiasis—papule

Filariasis
This is an infestation with thread-like helminths (Latin "Filum"—a thread). They are widely distributed in many species and live in the lymphatics and connective tissue. Fertilised eggs develop into embryonic worms—microfilariae. These are taken up by insect vectors that act as intermediate hosts in which further development occurs. They are then inoculated into a human host when next bitten by the insect.

Three diseases are caused by filarial worms:

- Lymphatic filariasis due to *Wuchereria bancrofti*, which liberate microfilariae into the blood stream.
- Onchocerciasis due to *Onchocera volvulus*. The microfilariae are liberated into the skin and subcutaneous tissues.
- Loiasis due to *Loa loa*, in which microfilariae are found in the blood.

Lymphatic filariasis affects 120 million people in 73 countries (34% in sub-Saharan Africa). It causes lymphoedema of the legs, genitalia, and breasts. It may be asyptomatic for a long period and the adult worms live for four to six years in the lymphatic vessels and lymph nodes producing thousands of microfilaria each day. These are picked up by mosquitoes when they take a blood meal and are passed on to the next victim when they feed again.

Lymphoema of the legs in filariasis

Treatment

- In endemic areas the whole community should be treated with a single dose of two of the following three drugs once a year for four to six years:
  (a) Ivermectin 400 micrograms/kg body weight
  (b) Diethylcarbamazine (DEC) 6 mg/kg body weight
  (c) Albendazole 600 mg.
- The chronic lymphoedema can be improved by keeping the legs moving, raising the legs when sitting, and prevention of secondary bacterial infection by regular washing and moisturising of the skin.

Diagnosis

- Find the microfilaria on a thick blood smear taken at midnight. This is not a very convenient method of diagnosis.
- Immuno-chromatographic filariasis card test using finger prick blood, which takes less than five minutes to complete. It detects circulating *W. bancrofti* antigens so it can be done at any time of the day or night.
- Polymerase chain reaction to detect parasitic DNA. This is very sensitive and can detect as little as one microfilaria in 1 ml blood.

Onchocerciasis

Onchocerciasis (river blindness) occurs in Africa south of the Sahara and in Central America. It is due to *Onchocerca volvulus* transmitted by the bite of black flies Simuliidae which breed by fast flowing rivers. The inoculation of microfilariae by the bite of a black fly causes intense local inflammation and is followed by an incubation period of many months. The adult worms live in nodules around the hips and cause no harm in themselves. They produce thousands of microfilaria each day which travel to the skin and eyes. In the skin they produce a very itchy rash which looks like lichenified eczema. On the lower legs there is often spotty depigmentation. Involvement of the eyes causes blindness.

Onchocerciasis

Risk factors for being infected

- Living, working, or playing near fast flowing rivers.
- Not wearing enough clothes so that the skin is exposed to insect bites.
- The construction of dams leads to less breeding of black flies in the dam itself but increased breeding in the dam spillways.

"Leopard skin" in onchocerciasis

Diagnosis

- Demonstrate the microfilaria in the skin by skin snips.
- Remove a skin nodule and see the adult worms inside it.
- Polymerase chain reaction to show parasite DNA—not much use in the field.

Treatment

- Spray the breeding areas with insecticides.
- Annual dose of ivermectin 400 micrograms/kg body weight for four to six years. This stops the release of microfilaria from the adult worms.

Loiasis

Loiasis occurs in the rain forests of central and west Africa. It is transmitted by mango flies (*Chrysops*). The adult worms live in the subcutaneous tissues where they can be seen in the skin and under the conjunctiva. The microfilaria are only found in the blood. A hypersensitivity to the worms shows itself as swelling of the skin, particularly of the wrists and ankles (calabar swellings).

Dracontiasis

This condition is due to infestation by *Draculus medinensis* in the connective tissue. It is acquired from drinking water containing the intermediate host, a crustacean, *Cyclops*. Localised papules develop on the lower legs containing the female worm and numerous microfilariae. Treatment consists of very carefully extracting the worm by winding it onto a stick over several weeks. Symptomatic treatment of secondary infection and allergic reactions is also required.

Diagnosis of loiasis

- Find the microfilaria in the peripheral blood between 10 am and 2 pm
- Find the adult worms by ultrasound examination
- High blood eosinophilia

Treatment of loiasis

- A single dose of ivermectin 400 micrograms/kg body weight or a three week course of albendazole 400 mg/kg body weight/bd
- Do not use diethylcarbamazine citrate (DEC) as this can cause death as a result of a reaction to toxins from the rapid destruction of the microfilaria.

24 Practical procedures and where to use them

DWS Harris

Skin lesions are easily accessible for removal or biopsy. The procedure used needs to be appropriate to the site and type of lesion involved. It is important also to keep scarring to a minimum.

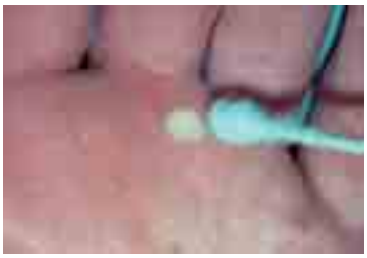
Destruction of skin lesions is carried out with:

- Electrocautery
- Cryotherapy
- Laser treatment

This is suitable for lesions where the diagnosis is certain, as no specimen is available or histology.

Removal of skin lesions results in a specimen for the pathologist to examine. The techniques used are:

- Curettage and cautery
- Surgical excision
- Incisional biopsy which provides a specimen for histology to supplement the clinical diagnosis.



Cryotherapy—applying the liquid nitrogen

Cryotherapy

This involves the destruction of tissues by extreme cold. Current methods used are:

Carbon dioxide

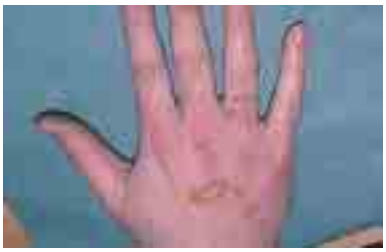
Solid carbon dioxide (temperature −64 °C) is produced by allowing rapid expansion of the compressed gas from a cylinder. This can be mixed with acetone to form a slush that can be applied with a cotton wool bud. A solid carbon dioxide stick, for direct application to lesions, is produced by an apparatus using "sparklet" bulbs.

The lesion must be frozen solid with a 1–2 mm margin of surrounding tissue. After thawing the freezing cycle should be repeated.

Cryotherapy—the frozen wart

Liquid nitrogen (−196 °C)

This can be simply applied using a cotton wool bud dipped in the vacuum flask of liquid nitrogen. Freezing takes a little longer than using spray apparatus. Various types of such apparatus are available with different sizes of nozzle. The larger ones are used for seborrheoic keratoses on the back, for example, and the smaller sizes for small lesions on the face. Freezing takes a few seconds and after thawing a further application can be made if necessary.



Cryotherapy—subsequent slight blistering

Ethyl chloride

This is sprayed directly on the skin, producing lowering of the temperature and temporary analgesia. It is not generally used for treatment.

Nitrous oxide

A cylinder of compressed gas is used to cool a probe to approximately −80 °C. It is usually used for the treatment of warts and requires a 30 second freezing cycle.

Precautions

- Cryotherapy produces pain and inflammation. Blistering and haematoma may occur. This can be diminished by the

Cryotherapy—liquid nitrogen containers

application of a strong steroid cream immediately after freezing, except when treating viral warts as it tends to encourage proliferation of the warts.
- Damage to deeper structures is rare but may occur when freezing the deeper tissues—for example, treating basal cell carcinoma.
- It is possible for adjacent structures to be damaged accidentally, especially with the liquid nitrogen spray. This applies particularly when treating lesions on the face, when it is essential to screen the eyes adequately.

Skin lesions suitable for freezing

Viral warts

These may require several treatments at two to three week intervals. Freezing very small plane warts can result in small depigmented areas, so they may be better treated with wart paint.

Seborrhoeic keratoses

These respond well to cryotherapy, but as they are superficial lesions care must be taken to avoid excessive freezing with resultant scars.

Papillomata and skin tags

These can be easily and permanently treated by compression with artery forceps dipped in liquid nitrogen. Surprisingly, this is generally a painless procedure.

Dysplastic lesions

Early lesions, which are potentially neoplastic or of low grade malignancy, can be effectively treated. This includes solar keratoses, if early and superficial, but follow up is essential. The lesions can progress to squamous carcinoma, and if not responding to cryotherapy they should be excised or removed with curettage and cautery for histological examination.

Bowen's disease

An intraepidermal carcinoma, if confirmed by incisional biopsy, can respond to repeated cryotherapy. Follow up is essential since progression to an invasive squamous carcinoma can occur.

Basal cell carcinoma

The superficial spreading type can be treated with liquid nitrogen, but repeated and often prolonged freezing is required. To be certain of effective treatment a thermocouple probe to record the temperature at the base of the tumour is used. This is not usually a routine procedure in general practice or hospital outpatients. Excision or radiotherapy are more effective methods of treatment.

Cryotherapy—practical points

- Be sure of the diagnosis before cryotherapy, taking a biosy if necessary
- Explain to the patient that inflammation, blistering, or haematoma formation may occur
- Use freezing with care in children and on areas where the skin is thin, such as below the eyelids. In pigmented skin postinflammatory hyperpigmentation can occur

Cryotherapy—freezing a lesion

Electrocautery—pinpoint cautery attachment

Electrocautery—pinpoint cautery attachment

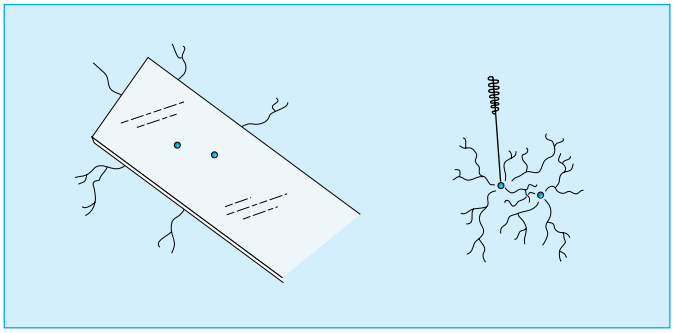
Electrocautery—hyfrecator

Electrocautery—spider naevus

Electrocautery

There are two forms of treatment:

(1) Heat from an electrically heated element, which is used for removal of skin tags and for treatment of the surface after curettage of warts, also seborrhoeic keratoses.
(2) High energy, low current "electrodesiccation" equipment which produces a high energy spark that can coagulate blood vessels or destroy some more papillomata. A fine needle point should be used for small telangiectatic naevi or milia. A larger needle is used for larger surfaces, for example after curettage.



Electrocautery—blanch the lesion to identify feeding vessels, then insert the needle into feeding vessels in the cold state

Laser treatment

Laser—Light Amplification by Stimulated Emission of Radiation—produces high energy radiation. The first laser apparatus was developed from microwave technology in 1960 by the nobel prize winner TH Mamian. It was initially used as a destructive tool to ablate tumours, but now different wavelengths can be directed at specific targets. Blood vessels, for example, take up the blue/green light of the argon laser and the red light of a ruby laser is well absorbed by the green dye of tattoos. Modern developments have resulted in laser equipment that produces minimal scarring and maximum specificity.

Although smaller portable units are available, laser treatment should still only be undertaken by those with appropriate training. The skin lesions most commonly treated by laser are described below.

Treatment of vascular lesions

- **Port wine stains**
  Argon and carbon dioxide lasers cause scarring, so yellow light emitting types such as Krypton, Flashlamp Powered Dye Laser (FIPDL), or Copper Laser are used
- **Telangiectasia**
  This is also treated with the FIPDL, although this can cause transient purpura, or Copper Laser
- **Cavernous haemangioma**
  These can be treated by yellow dye laser followed by surgical excision

Tattoos

Tattoos contain a variety of pigments so that more than one type of laser may be necessary for complete removal. The same pigment may vary in response in different patients. Superficial dark pigment usually responds to the Q switch ruby laser, but deeper pigment may require the Nd:YAG laser or Alexandrite laser. Green pigment is usually removed with a Q switch ruby laser and red pigment with a green light laser such as the Q switched Nd:YAG. It is found that professional tattoos are usually more easily removed than the amateur type.

Curettage—seborrhoeic keratosis

Pigmented lesions

Melanin absorbs light over a wide range of wavelengths, which can result in undesirable loss of skin colour following laser treatment. This can be put to good use in the treatment of benign lentigines and café au lait patches or deeply seated pigmented naevi. A wide range of laser types can be used, including Q switch ruby and Nd:YAG lasers. Congenital pigmented naevi should not be treated unless the biopsy has confirmed that they are benign.

Curettage—actinic keratosis

Hair follicles

Laser equipment is available for removing excess hair and is a very effective cosmetic tool.

Laser surgery

Lasers can be used as a cutting tool and recent studies have shown them to be a very effective means of producing incisions in the skin.

Curettage

Curettage

This is a simple way of removing epidermal lesions. A curette has a metal spoon shaped end with a sharp cutting edge. There are a variety of shapes and sizes suitable for different lesions, from large seborrhoeic keratoses or papillomata to smaller ones for minute keratin cysts. A specimen is provided for histology but completeness of removal cannot be accurately assessed.

Lesions suitable for curetting

- Seborrhoeic keratoses
- Solitary viral warts
- Solar keratoses
- Cutaneous horns
- Small basal cell carcinomas

Local anaesthetic is used and, with the skin stretched, the curette is applied at the edge of the lesion which is then scooped off. It is advisable to work around the edges of larger or more firmly attached lesions. The dermis normally feels firm but when curetting off a keratotic horn or solar keratoses; a soft consistency may indicate dysplastic change. The base can be lightly cauterised to control bleeding, sterilise the site, and prevent recurrence.

Various types of disposable curettes are available and are easy to use.

Curettage and cautery

- Use a sharp curette of appropriate size
- Very firm control of the curette prevents it from suddenly skidding onto normal skin
- Repeat curettage and cautery for neoplastic lesions such as basal cell carcinoma and solar keratoses
- Send the specimen for pathological examination

Incisional biopsy and punch biopsy

It is essential to have a working clinical diagnosis, but wherever there is doubt the pathologists can provide much more precise information regarding the nature and extent of the lesion. For example, a patch of Bowen's disease (intraepidermal carcinoma) may resemble sclerosing superficial basal cell carcinoma and a biopsy will usually distinguish them. Similarly, what seems to be a dysplastic pigmented naevus clinically may, on the one hand, prove to be benign or, on the other hand, turn out to be a malignant melanoma requiring wide excision.

Immunofluorescent staining of a blistering lesion differentiates dermatitis herpetiformis, which is treated with a gluten free diet, from pemphigoid, which requires corticosteroids and often immunosuppressant drugs.

Incisional biopsy

This is suitable for larger lesions and is taken across the margin of the lesion in the form of an elipse. It is essential to include deeper dermis, as the significant changes in, for example, granuloma or lymphoid infiltrate may not be near the surface. An adequate amount of normal tissue should be included, so this could be compared with the pathological area and this also means there is enough normal skin to suture the incision together.

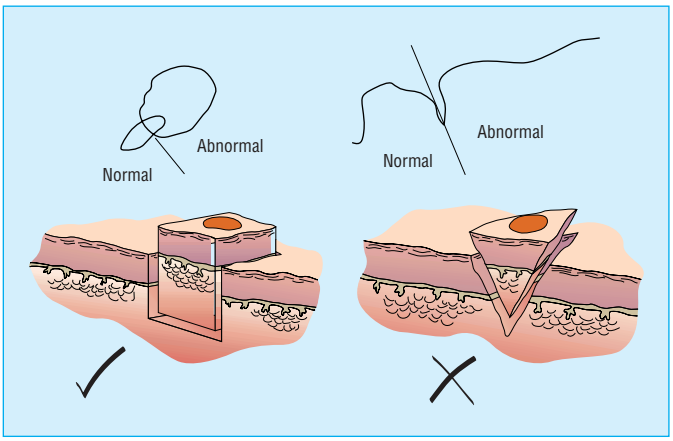
Punch biopsy

The biopsy tool consists of a small cylinder with a cutting rim which is used to penetrate the epidermis by rotation between the operator's finger and thumb. There is minimal danger of damaging deeper structures as the elastic subcutaneous tissues merely rotate with the tool without being cut. The resulting plug of skin is lifted out with forceps and cut off as deeply as possible.

Biopsy—equipment needed

Incisional biopsy—marking lesion

Incisional biopsy—histology



Incisional biopsy—amount of tissue that should be taken

Punch biopsy—injecting local anaesthetic

Punch biopsy—examples of punches

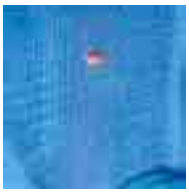
With the smaller sized punches the resulting defect can be treated with electrocautery or left to heal spontaneously. With a punch larger than 3 or 4 mm a single suture can be used. The main disadvantage of a punch biopsy is that it only provides a single small piece of tissue. It may not be representative or may miss an area of substantial change. It tends to leave a more prominent scar than the incisional biopsy.

Punch biopsy—tool inserted

Punch biopsy—plug of skin

Punch biopsy—raising plug of skin to cut

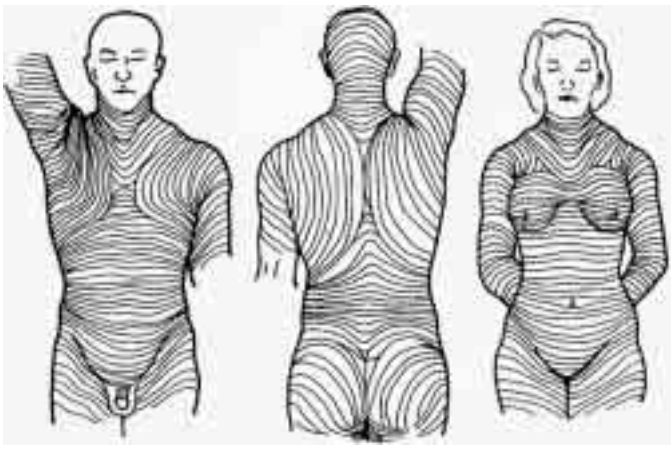
Punch biopsy—treating defect with electrocautery



Punch biopsy—specimen taken

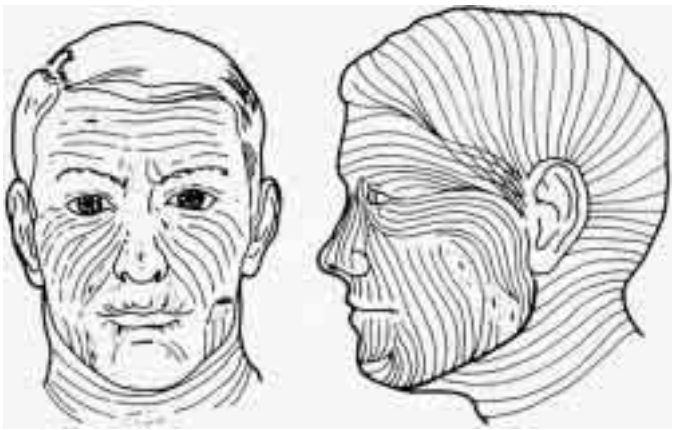
Excision of skin lesions is both curative and diagnostic. It may be the best way of making a diagnosis if there are multiple small papules or vesicles, one of which can be excised intact. Incisions should follow tension or wrinkle lines.

In the case of malignant lesions it is particularly important that the whole lesion is adequately excised. The pathologist can report on the adequacy of excision, but not in multifocal basal cell carcinoma where this cannot be assessed. If there is likely to be any doubt about the excision being complete it is helpful to attach a suture to one end of the excised specimen so the pathologist can describe which border, if any, extends over the excision margin.



Surgical excision—"wrinkle lines" of skin

Surgical excision—"wrinkle lines" of skin



Surgical excision—"wrinkle lines" of skin

Technique

The basic technique consists of making an elliptical incision with the length three times the width. This enables suturing without the formation of "dog ears" at the end. The long axis of the excision should follow the "wrinkle lines" of the skin, which are parallel to the collagen bundle in the dermis. This produces stronger, narrower scars. They are not the same as the deeper lines or fasical attachment or "Lange lines". Lesions on the sternal area, upper chest, and shoulders, where keloid scars often form, should only be excised when it is essential and may be best referred to a plastic surgeon.

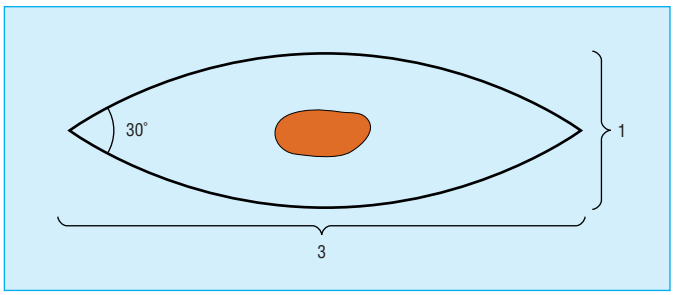
Local anaesthetic is injected subcutaneously but close to the skin. The incision should be vertical rather than wedge shaped. Monofilament sutures cause less inflammation and trapping of serum than the braided variety, but are harder to tie securely.

Methods of suturing and the more specialised techniques of flaps and grafts are outside the scope of this book. It is an asset for the dermatologist to be able to carry out surgical procedures on the skin and suitable courses are generally available.

Surgical excision—elliptical incision

Surgical excision—raising skin

Surgical excision—appearance of lesions under skin



Surgical excision—guidance for making elliptical incision

Surgical excision

- After initially inserting the needle, withdraw the plunger of the syringe to check the needle has not entered the blood vessels. Raising a small "bleb" of local anaesthetic ahead of the needle point helps to prevent this
- It is important to learn appropriate suturing techniques for different sites of the body and size of lesion
- Warn the patient that a scar will result and make sure that this is minimal and does not produce any deformity such as displacement of the eyelid
- Always send an excised lesion for histology. A significant number of lesions diagnosed as being benign clinically have been shown to be malignant. This would of course be missed if they had not been sent to a laboratory

Further reading

Lawrence C. *Introduction to dermatological surgery.* Oxford: Blackwell Science, 1997

Tromovitch TA, Stegman SJ, Glogau RG. *Flaps and grafts in dermatologic surgery.* Chicago: Year Book Medical, 1989

Zachary CB. *Basic cutaneous surgery: a primer in technique.* New York: Churchill Livingstone, 1991

25 Dermatology in general practice

R Balfour, E Crawford

In common with other aspects of general practice the management of dermatological problems has changed considerably in recent years. In particular:

- There is an increased expectation from patients, who want an accurate diagnosis and prompt treatment but also expect treatment for even the smallest lesion.
- The management of inflammatory skin conditions no longer requires many weeks of inpatient treatment. "Daily dressing clinics" in hospital dermatology departments enable patients to continue their daily lives while being treated, but still require visits to hospital. In many practices facilities for carrying out dressings by the practice nurse have been developed.
- The demand for specialist services far outstrips supply. Consequently the majority of patients with skin conditions have to be treated by general practitioners. One positive outcome is that there is a greater emphasis on shared care between general practice and hospital specialist departments. This is facilitated by the appointment of dermatology liaison nurses who are able to supervise patient's treatment in the community and, as the name suggests, liaise with both general practice and hospital departments using the resources of each as appropriate.
- General practitioners are increasingly developing special interests and many have part-time posts in specialist departments. In the case of dermatology this enhances their clinical knowledge, which they can bring to bear on the problems in general practice.

Between 10% and 15% of consultations in general practice are for skin related problems, although the actual number of skin conditions seen is probably much higher than this. In one general practice an analysis of 100 consecutive consultations showed that 38 involved some aspect of dermatology. Increasing knowledge of dermatology enables conditions to be diagnosed and treated. Even if the diagnosis is not known it is important to be able to assess the probable importance of dermatological conditions and differentiate those for whom an urgent referral to hospital is required from those needing a specialist opinion to confirm the diagnosis and treatment but for whom there is no great urgency. This is important with the large demands being made on hospital departments with diminished funding and increasing waiting times. In this respect a good working relationship with the local hospital department is a great asset.

Dermatology in general practice

- In a study of one area of London 55% of adults examined had some form of skin disease*
- Of those with moderate or severe conditions only 24% had made use of medical facilities in the previous six months
- In the same study 30% of patients with skin conditions medicated themselves
- Eight per cent of patients attending a general practitioner for a skin condition are referred to a dermatologist†
- A large reservoir of people in the community with skin conditions do not seek medical attention but can be expected to do so as awareness of skin conditions increases‡
- General practitioners with postgraduate training in dermatology and dermatology liaison nurses linked to specialist centres are needed so that the increasing demand for dermatology services can be met
- In countries with specialist care, basic training in the essentials of daignosis and treatment of skin conditions enables health officers to make a considerable impact on these conditions in the community

*Rea JN, Newhouse ML, Halil T. Skin disease in Lambeth: a community study of prevalence and use of medical care. *Br J Prev Soc Med* 1976;30:107-14
†Carmichael AJ. Achieving an accessible dermatology service. *Dermatol Pract* 1995;3:13-16
‡Savin J. The hidden face of dermatology. *Clin Exp Dermatol* 1999;18:393-5

Hospital consultation

Diagnosis of skin conditions in general practice

As in all aspects of medicine, knowing how the more common conditions present will provide a basis for recognising the unusual variants or less common lesions. Psoriasis, eczema, and other forms of dermatitis usually present no problems. Where they do a review of the history often gives valuable clues. Atopic eczema in childhood may explain the development of a widespread itching rash in a young adult and an otherwise unexplained rash may well be accounted for after a review of

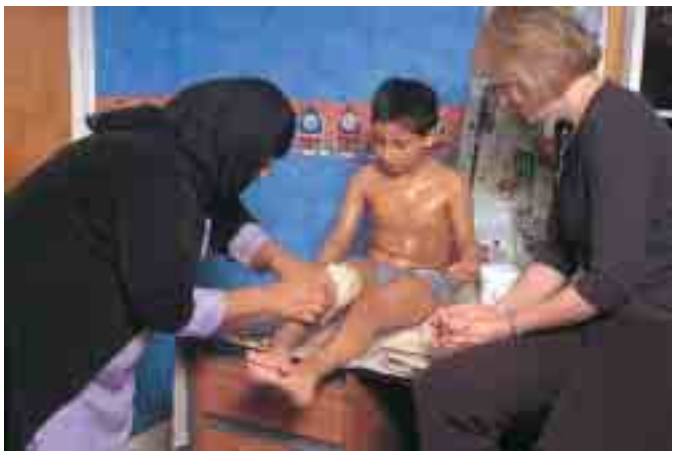
current drugs the patient is taking. Knowledge of the dermatological conditions associated with systemic disease, and conditions that may mimic them, is clearly important. A bilateral malar rash with photosensitivity in a woman should suggest the possibility of lupus erythematosus and appropriate investigations instituted. However, a completely typical case of rosacea does not require extensive investigation. The most useful diagnostic aids in general practice are:

- Skin scrapings and a sellotape strip should be sent for mycology whenever there is an area of itching inflammatory change, particularly in the flexures, that is not responding to treatment.
- A swab for bacteriology should be sent from any area of dermatitis that develops crusting and exudate.
- An incisional biopsy can be carried out to confirm a significant diagnosis, for example in a patch of Bowen's disease This is not usually needed to make a diagnosis of granuloma annulare, which has a very characteristic presentation. All lesions removed by excision or curettage or cautery should be sent for histology.
- Patch testing is not practicable in general practice as a rule because of the large number of reagents and specialised nursing skills required. Patients suspected of having contact dermatitis should be referred to the appropriate unit.

Demonstration of dressings by liaison nurse

The management of skin conditions in general practice

One great advantage of general practice is that there is continuity of care and the family doctor has a much more complete overall picture of the patient, their family and social circumstances than can be acquired in a hospital consultation. Increasingly dressings and other treatments are being used by practice nurses in conjunction with the dermatology liaison nurse when necessary. This applies to inflammatory skin conditions such as psoriasis and eczema as well as leg ulcers, but also to conditions such as Darier's disease, dermatitis herpetiformis, and lupus erythematosus where regular supervision and blood tests may be required. There is no reason why continuing treatment with drugs such as ciclosporin and methotrexate cannot be carried out in general practice once the diagnosis and treatment regime have been established. Regular blood tests are mandatory when these drugs are being used.



Dressings being applied at home by mother

Procedures in general practice

The details of techniques decribed in chapter 24.

- *Excisions* and other forms of minor surgery are probably best undertaken by a member of the practice who has developed expertise in this area, and received some level of training. It is particularly important to be aware of which lesions and anatomical sites are the most difficult for minor surgery. It is probably wise not to attempt excisions over the sternum and shoulders, particularly in young people, where keloid scars will probably result. In general it is best to refer any malignant lesion to a hospital department for excision.
- *Curettage and cautery* is a straightforward procedure for the removal of superficial and well defined lesions such as seborrhoeic keratoses and viral warts. It is as well to send all specimens for histology—to confirm the diagnosis and make sure that an unsuspected malignancy is not missed. It is important that a clinical probable diagnosis is made so as to

Mother, liaison nurse, and health visitor discuss treatment

avoid excising lesions where such treatment is not required or when it is inappropriate.

- *Cryotherapy* is suitable for the treatment of warts, seborrheoic keratoses, solar keratoses, and conditions such as Bowen's disease. If there is doubt as to the diagnosis or the condition fails to respond to treatment a specialist opinion should be sought. It usually most satisfactory to have a cryotherapy clinic for which liquid nitrogen is regularly supplied and a suitably trained nurse is available to carry out the treatment. It is important that all patients are seen by by a doctor before treatment is started so that the diagnosis and suitability for treatment can be confirmed.

Self-help groups

There are now groups that provide an invaluable source of information and advice for a wide range of conditions. A number of these are listed in the appendix.

Mother applying wet wraps at home

Assessment of circulation in leg ulcers by liaison nurse

Further reading

Barker DJ, Millard LG. *Essentials of skin disease management.* Oxford: Blackwell Scientific, 1979

Lamberg SI. *Dermatology in primary care: a problem oriented guide.* Philadelphia: Saunders, 1986

Stone LA, Lindfield EM, Robertson SJ. *A colour atlas of nursing procedures in skin disorders.* London and St. Louis: Mosby–Wolfe, 1989

26 Formulary

The zinc topped tables used for many years to prepare tar "spreads" in a teaching hospital dermatology department were recently thrown out—a sign of the times and an indication of both the increasing use of systemic treatment and much more effective forms of phototherapy. There is still an important place for topical treatment and "dressing clinics" to play a vital role in the treatment of skin diseases and enabling affected individuals to continue their daily lives as far as possible.

The link between hospital departments and community services has been greatly increased by the development of the role of "liaison nurses". These nurses, with experience and training in the treatment of skin disease, visit patients in their homes to supervise treatment in conjuction with the general practitioner and the practice nurse. As they are based in the hospital they can call on any specialist opinion or treatment needed.

A great variety of preparations is available for the treatment of skin conditions, and those most commonly used are described. There are numerous effective alternatives.

The techniques for the treatment of specific conditions are described in the appropriate chapters.

Topical treatment

General treatments

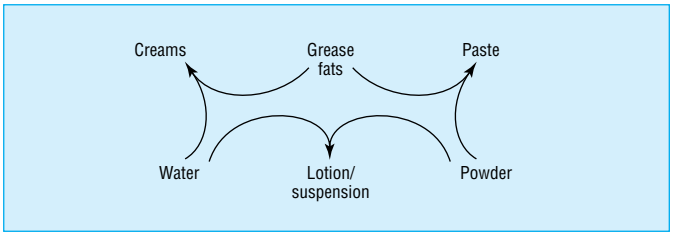
The epidermis is capable of absorbing both greasy and aqueous preparations or a mixture of the two. The type of lesion influences which type is used.

- Dry, scaling skin—greasy ointments.
- Crusted, weeping lesions—creams which are an emulsion of greases in water.
- For occlusion or long action—pastes which are powder (for example, zinc oxide) in an ointment.
- For the face and scalp—gels

Composition of bases

The consistency and properties of ointment and cream depend on the ratio of oil or grease to water (that is, whether they are oil in water or water in oil) and the emulsifying agents used. For example, emulsifying ointment contains soft white paraffin, emulsifying ointment, and liquid paraffin.

The oils and greases range from mineral oil through soft paraffin to solid waxes. Some are naturally occurring, such as lanolin and beeswax. Creams or ointments may be used on their own as emollients or as vehicles for active ingredients.



Composition of bases

Emollients

There are numerous preparations for softening dry scaling skin and it is largely a matter of personal preference as to which one is used.

Official preparations

- Soft white paraffin—greasy; protects skin and is long lasting.
- Emulsifying ointment—less greasy; mixes with water and can be used for washing.
- Aqueous cream—oil in water emulsion; useful as a vehicle, as an emollient, and for washing.
- Liquid paraffin: white soft paraffin, equal parts—spreads easily and is less greasy than white soft paraffin.
- Hydrophilic ointment—contains propylene glycol; mixes with water and spreads easily.
- Lanolin (hydrous wool fat)—the natural emollient from sheep; mixes with water and greases, softens the epidermis, but can also cause allergic reactions.

Proprietary preparations

Proprietary preparations are numerous, varied, and more expensive than the standard preparations. They may also contain sensitisers—lanolin and preservatives (hydroxybenzoate, chlorocresol, sorbic acid)—and can cause allergies. Some examples are E45 cream (Crookes), Oilatum cream (Stiefel), and Lacticare (Stiefel), Unguentum Merck (Merck), Aquadrate (Norwich Eaton), and Diprobase (Schering-Plough).

Bath additives

Bath additives comprise dispersible oils such as Oilatum (Westwood, United States), Aveeno (Bioglan), Balneum (Merck), Alpha Keri (Westwood, United States), Emulsiderm, and Dermol (Dermal).

Topical steroids

Topical steroids provide effective anti-inflammatory treatment but have the disadvantage of causing atrophy (due to decreased fibrin formation) and telangiectasis. They are readily absorbed by thin skin around the eyes and in flexures. On the face the halogenated steroids produce considerable telangiectasia, so nothing stronger than hydrocortisone should be used (except in lupus erythematosus). They can cause hirsutism and folliculitis or acne. Infection of the skin may be concealed (tinea incognita, for example) or made worse.

Side effects can be avoided by observing the following guidelines:

- Avoid long term use of strong steroids.
- Potent or very potent steroids should be applied sparingly and often for a short time, then a less potent preparation less often as the condition improves.
- Use only mildly potent steroids (that is, hydrocortisone) on the face.
- Use preparations combined with antibiotics or antifungals for the flexures.

Topical steroids come in various strengths and a wide variety of bases—ointments, creams, oily creams, lotions, and gels—which can be used according to the type of lesion being treated.

Their pharmacological activity varies and they are classified according to their potency, the synthetic halogenated steroids being much stronger than hydrocortisone:

- Mildly potent—hydrocortisone 0·5%, 1%, and 2·5%
- Moderately potent—Eumovate (GSK), Stiedex LP (Stiefel)
- Potent—Betnovate (GSK), Cutivate (GSK), Locoid (Brocades), Synalar (Zeneca)
- Very potent—Dermovate (GSK).

In Britain a full list showing relative potencies appears in MIMS. Combinations with antiseptics and antifungals, are listed below:

Mildly potent
- Vioform HC (Zyma)
- Terra-Cortril ointment (Pfizer), containing oxytetracycline and hydrocortisone
- Fucidin H cream or ointment (Leo), containing fucidic acid and hydrocortisone
- Canesten HC (Baypharm)
- Daktacort cream (Janssen)

Moderately potent
- Betnovate N (betamethasone and neomycin) (GSK)
- Synalar N (neomycin; Zeneca)
- Trimovate cream (clobetasone butyrate, nystatin, and oxytetracyclin; GSK)
- Fucibet (betamethasone, fucidic acid; Leo)

Very potent
- Dermovate-NN (clobetasone, with neomycin and nystatin; GSK).

Antiseptics and cleaning lotions
Simple antiseptics are very useful for cleaning infected, weeping lesions and leg ulcers.

Potassium permanganate can be used by dropping four or five crystals in a litre of water or in an 0·1% solution that is diluted to 0·01% for use as a soak. It will stain the skin temporarily and plastic containers permanently.

Silver nitrate 0·25% is a simple, safe antiseptic solution that, applied as a wet compress, is useful for cleaning ulcers.

Flamazine (Smith and Nephew) is silver sulfadiazine cream, used for leg ulcers, pressure sores, and burns.

Hydrogen peroxide (6%) helps remove slough but tends to be painful. Hioxyl (Quinoderm) is a proprietary cream for desloughing.

Iodine (2·5%) is an old fashioned, effective preparation as a tincture in alcohol and Betadine (Napp) is a proprietary equivalent.

Shampoos. Ceanel concentrate (Quinoderm) contains cetrimide 10%. Ionil T (Galderma) has benzalkonium chloride and coal tar solution, and Betadine (SSL) contains povidone iodine. Shampoos containing selenium sulphide (Selsun, Abbot) and ketokonazole (Nizoral, Janssen) can be used for seborrhoeic dermatitis and also for pityriasis versicolor of the skin.

There are numerous other antiseptic, cleansing, and desloughing agents such as cetrimide, chlorhexidine, benzalkonium chloride, benzoic acid, and enzyme preparations such as Varidase (Lederle), a streptokinase and streptodornase preparation.

Tar preparations
These are mainly used for treating psoriasis as described in chapter 3.

Tar has an anti-inflammatory effect and seems to suppress the epidermal turnover in lesions of psoriasis. The various tar pastes are generally too messy to use at home and are most suitable for dermatology treatment centres. Standard tar paste contains a strong solution of coal tar 7·5% in 25 g of zinc oxide, 25 g of starch, and 50 g of white soft paraffin.

There are numerous proprietary preparations that are less messy and do not stain but are not so effective. They are useful for treating less severe psoriasis at home. Examples are: Alphosyl cream (Stafford-Miller), Pragmatar (Bioglan), Psoriderm (Dermal); alphosyl HC (Stafford-Miller) and Carbo-Cort (Lagap) contain hydrocortisone as well.

Dithranol can be used in a paste containing salicylic acid, zinc oxide starch, and soft white paraffin. It has to be applied carefully avoiding contact with the surrounding skin, as it can cause severe irritation. It is best to start with a low concentration.

For short contact treatment relatively clean preparations in a range of concentrations are available, such as Dithrocream (Dermal), Anthranol (Stiefel), and Psoradrate cream (Stafford-Miller).

Ichthammol is a useful soothing extract of shale tar. It can be made up as a 1% paste in yellow soft paraffin with 15% zinc oxide.

Bath preparations are useful for dry skin and widespread psoriasis. Coal tar solution (20%) can be used or Polytar Emollient (Stiefel) or Psoriderm.

Tar shampoos are useful for treating psoriasis of the scalp. Polytar (Stiefel), T-Gel (Neutrogena), Capasal (Dermal), and Alphosyl (Stafford-Miller) are some examples.

Keratolytics
These can be used for hyperkeratotic lesions. They soften and help remove excess keratin. If used for extensive areas or in infants systemic absorption can occur. A useful preparation is salicylic acid 2–4% in aqueous cream. Salicylic acid with betamethasone ointment (Diprosalic ointment, Schering-Plough) can be used for hyperkeratotic lesions where inflammation is present.

Antipruritics
Useful anti-pruritics for persistent itching include menthol (0·5%) or phenol 1% in aqueous cream, and calamine lotion, which contains arachis oil.

Barrier and protective preparations
These preparations protect against softening and maceration from moisture in flexures, for example the groins. They also have an occlusive effect. They are essentially bases with zinc oxide or silicone (as dimethicone). There are many preparations; some of the most commonly used are:

- Zinc cream BP, contains zinc oxide, arachis oil, and lanolin
- Zinc and castor oil ointment BP
- Conotrane (Yamanoouchi) and Siopel (Bioglan), contain dimeticone (dimethicone)
- Metanium (Roche), contains titanium dioxide
- Sudocrem (Forest) and Drapolene (Pfizer), contain lanolin.

Treatment for specific situations
Sunscreens
These give a degree of protection—mainly to ultraviolet B but also to ultraviolet A. They depend on their effect on a physical barrier (usually titanium dioxide) and chemicals that combine with epidermal cells, usually esters of PABA or oxybenzone.

Camouflage
Scars, congenital naevi and other blemishes that cannot be removed can be covered with suitable creams. Proprietary preparations are available.

Antiperspirants
Aluminium chloride for hyperhidrosis: aluminium chloride 20% (Driclor, Stiefel, or Anhydrol, Dermal Laboratories).

Depigmenting agents
2% hydroquinone cream is available without prescription as "fade-out". Preparations containing corticosteroids are also prescribed but not available as proprietary preparations.

Antimitotic agents
5-Flourouracil cream is useful for treating incipient malignancies—that is, solar keratoses, but not actual carcinomas. It is available as Efudix cream (Roche), which is applied daily for one to two weeks. It produces a variable degree of inflammation that is allowed to subside before the treatment is repeated.

Infestations
(1) *Scabies*. The correct procedure for treatment is more important than the preparation used. Benzyl benzoate 25% application BP is still available and is cheap but tends to irritate the skin. Malathion is available as Derbac (SSL), Prioderm (SSL), and Quellada M (Stafford-Miller), and Permethrin as Lyclear cream (Kestrel) preparations are more effective and less likely to irritate. 6% sulphur in white soft paraffin or permethrin are recommended for young children and pregnant or lactating women. The procedures for treatment set out on page 107 should be followed and clearly explained to the patient. For resistant cases ivermectin (Mectizan, MSD) by mouth is available on a named patient basis.
(2) *Pediculosis*. Preparations containing malathion, carbaryl, and permethrin are used either as shampoos or lotions. Lotions are most effective and should be left on the skin for 12 hours before washing off. The same preparations are available as for treating scabies, with the addition of 0·5% malathion lotion as Suleo-M (SSL). Recently a lotion of phenothrin (Full Marks, SSL) has become available for treating head and pubic lice.

Preparations for the mouth
Steroids—Adcortyl in Orabase (Squibb) or Corlan pellets (Evans). Both these preparations contain corticosteroids.

Antifungals—Daktarin (Janssen) or Fungilin lozenges (Squibb); Nystan (nystatin suspension, Squibb). Corsoidyl (chlorhexidine, GSK), and Difflam (3m Riker) are useful mouthwashes.

Topical immunosuppressants
Tacrolimus (Protopic, Fujisawa) has recently become available as an ointment in two strengths, 0.03% and 0.1%. It has not been evaluated in children under the age of two or in pregnant women. It is recommended that it is only used by dermatologists or those with considerable experience in treating eczema. Although the exact mode of action is unknown it does diminish T cell stimulation by Langerhan cells and diminishes the production of inflammatory mediators from mast cells. It should be used in moderate to severe atopic eczema that has not responded to either treatment. Skin irritation with burning, erythema, and pruritis are the most common side effects. In view of its immunosuppressive activity any infection should be treated first and it should be used with caution if there is a risk of viral infection or if inoculations using attenuated or live organisms are being used.

Pimecrolimus (Elidel, Steeple Novartis) is a similar preparation recommended for intermittent treatment of eczema can also be used as an initial treatment for any flare up of eczema. It diminishes cytokine activity long term relieving both the erythema and pruritis of eczema.

In common with topical steroids any immunosuppressive drug should be used with caution as viral infections are likely to be present or the patient is undergoing inoculation with live or attenuated organisms.

Systemic treatment

Antibiotics are probably the most commonly used systemic treatment. Long term antibiotics are needed for acne and cellulitis. Antifungal and antiviral drugs are indicated if topical treatment is ineffective, particularly in the immunosuppressed, and when the infection has been confirmed by laboratory tests.

Immunosuppressant drugs have had a considerable impact on the treatment of autoimmune and connective tissue diseases and diminished the need for systemic steroids—previously the only treatment available. They are increasingly used for extensive and persistently inflamed dermatoses, particularly psoriasis and eczema.

Antibacterial drugs
All penicillins may cause allergic rashes, which may be severe, and the broad spectrum penicillins, amoxicillin, ampicillin, and co-amoxiclav, are particularly likely to cause an intense rash in patients with glandular fever. They tend to accumulate in patients with renal failure and may reduce the excretion of methotrexate which is used in the treatment of psoriasis.

Phenoxymethylpenicillin (penicillin V) is useful in Gram positive infections and erysipelas.

Flucloxacillin is used to treat infections due to penicillinase producing organisms. It is used in impetigo and cellulitis.

Amoxicillin and ampicillin are broad spectrum antibiotics but are destroyed by penicillinase. Co-amoxiclav is a combination of amoxacillin and clavullinic acid. It is effective against a wide range of organisms and beta lactamase producing staphylococci as well.

Cephalosporins are not affected by penicillinase and are effective against both Gram positive and Gram negative infections.

Ciprofloxacin is used for infections with both Gram positive and Gram negative organisms such as pseudomonas.

Erythromycin is used for the treatment of acne and is useful in Gram positive infections. Resistant strains of staphylococcus are appearing.

Metronidazole is useful for treating anaerobic infections and trichomonas infections. It is useful for rosacea that is not responding to conventional treatment.

Antifungal drugs
Topical treatment is usually effective but for fungal infection of the nails and intractable infections of the skin systemic treatment may be required.

Griseofulvin (500 mg daily) is a well established treatment for fungal infections of the skin, hair, and nails. Although it should not be used in pregnancy, it can be used in children. It can cause lupus erythematosus to flare up.

Terbinafine (250 mg daily) is an effective systemic antifungal drug that does not affect the liver. It is used for both nail and skin infections.

Imidazole and triazole drugs include itraconazole and ketoconazole, which are effective for dermatophyte infections of the skin and pityriasis versicolor.

Antiviral drugs

Discovery of drugs that inhibit viral DNA polymerase and inhibit their proliferation in vivo means that effective treatment for herpes simplex and zoster is now possible. They are effective at the early stages of infection and should be started as soon as symptoms appear. Aciclovir (Zovirax, GSK) is available as a cream.

Aciclovir is effective against both herpes simplex and zoster. The standard dose is 200 mg five times daily for five days. In varicella infections and herpes zoster 800 mg is given five times daily for seven days. It can also be given by intravenous infusion, and should be applied as soon as symptoms appear. In addition, it can be used for prophylaxis, particularly in the immunocompromised patients and atopics who are liable to fulminating infection.

Famciclovir and valaciclovir are similar and are recommended for treating herpes zoster.

Antihistamines

These drugs are used in urticaria and acute allergic (type I immediate hypersensitivity) reactions. The newer long acting and non-sedating antihistamines are useful for treatment during the day and can be combined with one of the sedating type at night if pruritus is preventing sleep.

Non-sedating antihistamines only cross the blood–brain barrier to a slight extent. They may cause arrhythmias, particularly terfenadine.

- Acrivastine (Semprex, GSK) 8 mg three times daily
- Cetirizine (Zirtek, UCB Pharma) 10 mg once daily
- Fexofenadine (Telfast, Hoechst) 120 or 180 mg once daily
- Loratadine (Clarityn, Schering-Plough) 10 mg once daily.

Sedating antihistamines

There are many available and which is used is largely a matter of personal preference. The sedating effect, which is enhanced by alcohol, means that they are best taken at night. They also potentiate CNS depressants and anticholinergic drugs. They tend to have anticholinergic effects, causing dry mouth, blurred vision, tachycardia, and urinary retention. Those commonly used are:

- Chlorphenamine (Piriton Stafford-Miller 4 mg daily)
- Cyproheptadine (Periactin (MSD) 4 mg up to four times daily)
- Hydroxyzine (Atarax (Pfizer) 10–25 mg at night; can be used during the day if drowsiness is not a problem)
- Promethazine (10 or 25 mg at night or twice daily)
- Trimeprazine (Vallergan (Castlemead) 10 mg two to three times daily).

Corticosteroids

In addition to topical preparations, systemic steroids may be required for the treatment of severe inflammatory skin conditions such as erythroderma developing from psoriasis or eczema. They are also used in vasculitis and erythema multiforme as well as connective tissue diseases. They are often required for the treatment of pemphigoid and pemphigus together with immunosuppressant drugs.

The side effects must be borne in mind, particularly for any long term treatment. Most important are given below.

Water and electrolytes

Sodium and water retention with loss of potassium.

Musculoskeletal

Osteoporosis, aseptic necrosis of the femoral head, growth retardation in children, and muscle wasting.

Ophthalmic effect

Cataract formation and increased tendency to glaucoma.

Other effects

Increase in blood pressure, peptic ulceration and fat redistribution, and impaired glucose intolerance.

Retinoids

These vitamin A derivatives have proved very effective in the treatment of psoriasis and acne but are not without risk of side effects. The most serious is that they are teratogenic and must be discontinued for at least three months after stopping treatment in the case of isotretinoin and five years after taking acitretin.

All patients should be warned of possible side effects and women of childbearing age must be using an effective form of contraception, which must have been used for at least a month before treatment has started as well as having a pregnancy test carried out. Liver function tests and fasting cholesterol and triglycerides should be carried out on all patients. After prolonged treatment in adolescence, radiological tests should be carried out to ensure that there is no extraosseous calcification. The most important side effects are:

- Abnormal liver function tests
- An increase in cholesterol and triglycerides
- Occasional increases in electron spin resonance and lowered white count.

Clinical side effects

Drying and roughening of the skin and mucous membranes, particularly the lips, can occur. There may also be thinning of the hair and nails. Photosensitivity eruptions can develop. Occasionally muscle and joint pains occur.

Acitretin

This drug is used for severe psoriasis including pustulosis of the hands and feet. It has also been used in other forms of keratosis such as Darier's disease and pityriasis rubra pilaris.

Isotretinoin

This drug is used for severe acne vulgaris that has not responded to antibiotics or other treatments. It is therefore often used in adolescence and it is important to be aware of the musculoskeletal effects and possible mood changes.

Immunosuppressants

Methotrexate

This drug is useful in severe psoriasis that is not responding to topical treatment. The main disadvantage is its adverse effect on the liver, which precludes its use in those who have alcoholic liver disease but who are often those most needing systemic treatments. Idiopathic immunosuppression can occur so a test dose must always be given and a full blood count carried out 48 hours later before treatment has started. There may be gastrointestinal upsets and osteometitis as well.

Methotrexate interacts with anti-inflammatory and antiepileptic drugs. A full list of drug interactions should be consulted before treatment is started.

After a 2·5 mg test dose and full blood count 48 hours later, the regular dose is 5–15 mg by mouth once a week. Full blood count and liver function tests should be carried out once a week for the first six weeks and thereafter once a month during treatment. Folinic acid should be given at the same time, as this prevents bone marrow depression. In many centres a liver biopsy is considered mandatory before treatment is started since blood tests will remain normal for some time during the development of hepatic fibrosis.

Azathioprine

This drug is used for systemic lupus erythematosus, pemphigus, and bullous pemphigoid. It enables the dosage of systemic steroids to be reduced. The most serious side effect is bone marrow suppression. This may occur quite rapidly, particularly in those with diminished ability to metabolise the drug. This is carried out by thiopurine methyl transferase (TMT). The level of this enzyme should therefore be determined before treatment is started and those at low levels given a lower dosage. Those who inherit high activity may require higher doses. Other side effects include gastrointestinal upset, liver toxicity, and an increased tendency to infection.

Ciclosporin

This drug has proved helpful in severe psoriasis within inflammatory lesions and, secondly, in the treatment of severe atopic dermatitis. There are a number of drug interactions and it is important to check renal function and monitor both blood urea and serum creatinine.

Other drugs

Dapsone

This drug was originally developed for treating leprosy but was proved very effective in dermatitis pityformis and some other conditions, such as pyoderma gangrenosum. It may cause haemolytic anaemia, and other side effects include bone marrow suppression, hepatitis, and peripheral neuropathy. Regular blood checks are essential.

Hydroxychloroquine

This drug is used in both systemic and discoid lupus erythematosus as well polymorphic light eruption and porphyria cutanea tarda. The most serious side effect is retinopathy but this does not occur if the dose does not exceed 6·5 mg/kg lean body weight.

Psoralens

These drugs are used in conjunction with long wavelength ultraviolet light as psoralen with ultraviolet A (PUVA) therapy described on page 67. It is used for the treatment of severe psoriasis. It has also proved effective in some cases of atopic eczema, T cell lymphoma of the skin, and occasionally in lichen planus. There is a risk of cataract formation, and a full blood count as well as antinuclear factor tests should be carried out.

Preparations for treating acne and varicose ulcers are described in the appropriate sections.

Further reading

Arndt KA. *Manual of dermatological therapeutics*, 5th ed. NewYork: Little, 1995

Appendix: Patient support groups

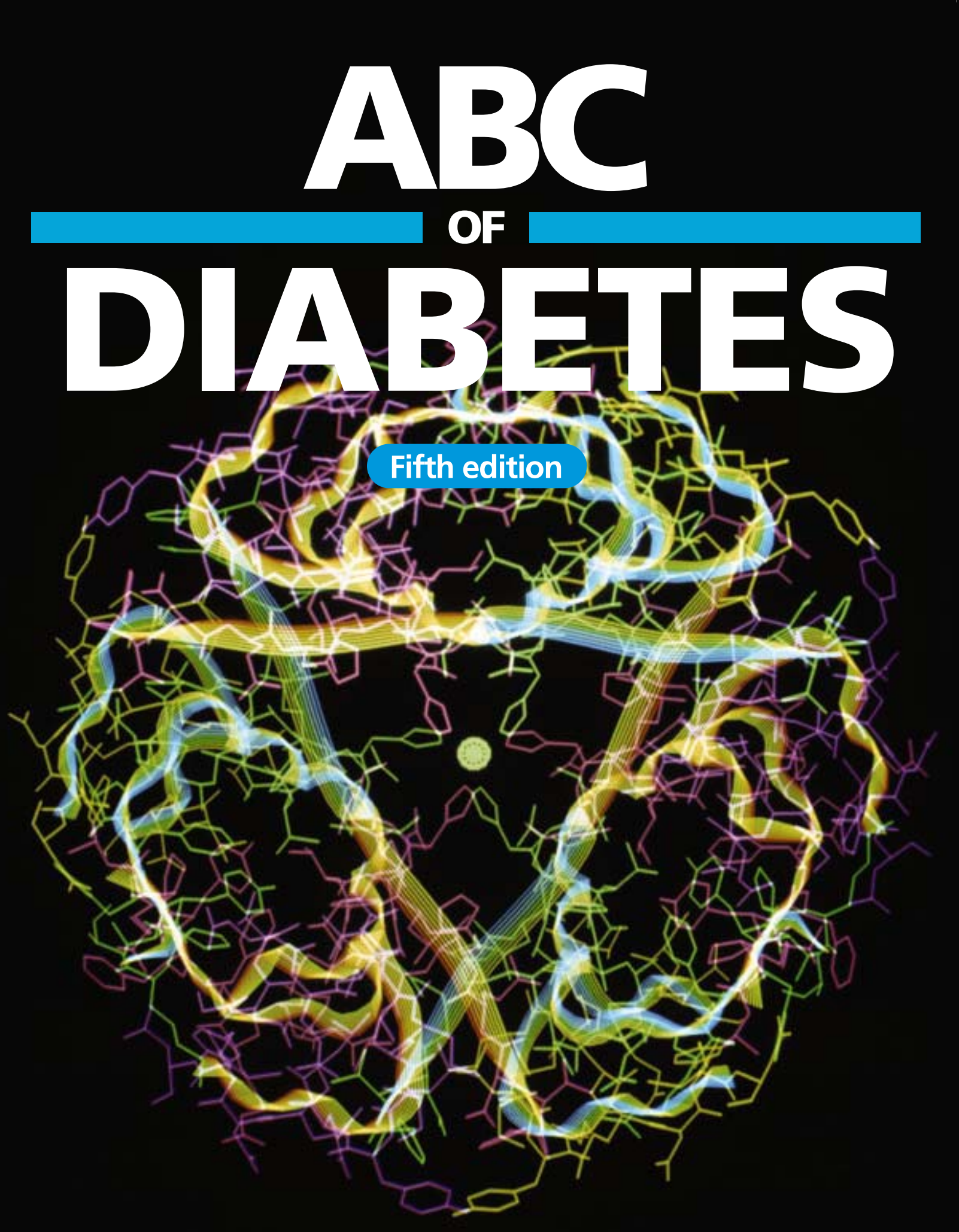
Acne Support Group	(0870) 870 2263
Behcet's Syndrome Society	(01488) 71116
Allergy UK	(020) 8303 8583
Cancer BACUP	(0808) 800 1234
British Association of Skin Camouflage	(01625) 267880
British Red Cross Skin Camouflage Service	(020) 7201 5172
British Leprosy Relief Association	(01206) 562286
Congenital Melanocytic Naevus Support Group	(0151) 281 9716
Darier's Disease Support Group	(01646) 695055
Dystrophic Epidermolysis Bullosa Research Association	(01344) 771971
Ectodermal Dysplasia Society	(01242) 261332
Ehlers–Danlos Support Group	(01252) 690940
Hairline International	(01564) 775281
Herpes Viruses Association	(020) 7609 9061
Ichthyosis Support Group	(020) 7461 0356
Latex Allergy Support Group	(07071) 225838
Lupus UK	(01708) 731251
Lymphoma Association (LA)	(0808) 808 5555
Marfan Association UK	(01252) 810472
Myositis Support Group	(023) 8044 9708
National Eczema Society	(0870) 241 3604
Neurofibromatosis Association	(020) 8547 1636
Pemphigus Vulgaris Network	(020) 8690 6462
Primary Immunodeficiency Association (PIA)	(010) 7976 7640
Pseudoxanthoma Elasticum (PXE) Support Group	(01628) 476687
Psoriatic Arthropathy Alliance	(01923) 672837
Psoriasis Association	(01604) 711129
Raynaud's and Scleroderma Association Trust	(01270) 872776
Scleroderma Society	(020) 8961 4912
Shingles Support Society	(020) 7607 9061
Telangiectasia Self Help Group	(01494) 528047
Tuberous Sclerosis Association	(01527) 871898
Vitiligo Society	(020) 7840 0855

Index

Page numbers printed in *italics* refer to boxed material

ABC OF DIABETES

Fifth edition



Peter J Watkins

BMJ Books

ABC OF DIABETES

Fifth Edition

ABC OF DIABETES

Fifth Edition

PETER J WATKINS

Honorary Consultant Physician, King's Diabetes Centre, King's College Hospital, London

BMJ Books

ЗНАНИЕ

First edition 1983
Second edition 1988
Third edition 1993
Fourth edition 1998
Fifth edition 2003

by BMJ Publishing Group Ltd, BMA House, Tavistock Square, London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0-7279-16939

Typeset by Newgen Imaging Systems (P) Ltd, Chennai, India.
Printed and bound in Spain by Graphy Cems, Navarra.

The cover image shows molecular graphics of insulin hexamer with permission of JC Revy/Science Photo Library

Contents

Acknowledgments

Any ideas or inspiration which these pages may contain have inevitably been learnt or borrowed from others. I am indebted to the late Professor JM Malins and Dr MG FitzGerald, through whose enthusiasm I was first introduced to diabetes, and to the late Dr David Pyke through whose energy this interest has been fostered over many years. Close collaboration with colleagues at King's both past and present has made possible many of the joint ventures described here, and I am grateful to them all. Our registrars and research fellows and above all our patients have been a constant source of inspiration.

I am particularly grateful to the many colleagues who have assisted me with the preparation of this book, especially Professor Stephanie Amiel (RD Lawrence Professor of Diabetic Medicine), Dr Michael Edmonds and Dr Stephen Thomas (consultant physicians), Dr Tyrrell Evans (general practitioner), Dr Phin Kon (renal physician), Dr William Marshall (Reader in Clinical Biochemistry), Dr Joanna Raeburn (associate specialist), Ms Helen Reid (Diabetes Specialist Nurse), and Mrs Eileen Turner (Consultant Nurse Specialist).
Dr Simon Page (consultant physician in Nottingham) has helped me considerably by his many valuable comments in reading the manuscript. My wife Mrs Val Watkins has throughout provided me with invaluable support and encouragement.

Consultant physicians to the Diabetes Centre at King's from 1938 (from left to right) Dr Wilfred Oakley (1905-1998); Dr David Pyke (1921-2001) and Dr Peter Watkins

Introduction

Advances in clinical science over a single professional lifetime during the second half of the 20th century have led to improvements in understanding the causes and complications of diabetes, together with alleviation of suffering to an extraordinary degree, unimaginable even 25 years ago. Many of the clinical improvements have been initiated at innovative centres across the United Kingdom.

In the 1960s and 1970s physicians had to stand by helplessly watching their patients overwhelmed by complications of the disease. Prevention of blindness by photocoagulation and renal support treatment for those in renal failure became possible in the 1970s, while development of specialist foot clinics during the 1980s succeeded in halving the amputation rate. The sad outcome for pregnancies even 20 years after the discovery of insulin when the fetal mortality rate was more than 25%, has been transformed so that now more than 95% of these pregnancies succeed. And now, the landmark Diabetes Control and Complications Trial (DCCT) of Type 1 diabetes in the United States, and more recently the astonishing achievement of the late Professor Robert Turner in completing the United Kingdom Perspective Diabetes Survey (UKPDS) of Type 2 diabetes have demonstrated how to reduce the incidence and progression of diabetic complications by good treatment.

Yet there is still more. The present technology of managing diabetes was undreamt of until the last quarter of the 20th century. The introduction of home blood glucose monitoring with new non-invasive technologies now in sight, has made possible the achievement of "tight control", while at the same time advances in understanding and reversing diminished awareness of hypoglycaemia are reducing its hazards. The invention of insulin pens and more recently the development of insulin pumps has contributed in great measure to improving the quality of life of those with the burden of lifelong diabetes. Furthermore after the British discoveries of the chemical (Frederick Sanger, 1955) and physical structure (Dorothy Hodgkin, 1969) of insulin followed by the revolution in molecular science, man-made insulin analogues have been introduced, giving further advantages in achieving good blood glucose control while minimising hypoglycaemia.

The initially controversial "invention" of the diabetes specialist nurse by Dr Joan Walker in Leicester in the 1950s is arguably one of the most important advances in health care, not only for those with diabetes but across the whole of medicine. The tremendous benefits in the delivery of care especially to those with diabetes and other chronic diseases have been accompanied by recognition of community needs and improvements in crossing the primary/secondary care interface. It is now to be hoped that improvements in information technology, more sophisticated audit, and provision of a national eye screening programme may emerge from the National Service Framework of 2002/2003.

RD Lawrence 1892-1968. Founder of the diabetic clinic at King's in the 1920s, founder of the British Diabetic Association in 1934

Rapid clinical advances of this magnitude require substantial support. Diabetes UK, founded as the Diabetic Association by Dr RD Lawrence and his patient HG Wells in 1934 (later the British Diabetic Association), has uniquely supported both patients and their needs as well as clinical and scientific research. More recently the Juvenile Diabetes Foundation has made substantial contributions. Furthermore the pharmaceutical industry has been both innovative in its own laboratories as well as supportive of both patients and clinicians.

It gives particular pleasure to reproduce some parts of the personal account by Mrs B-J (with her permission) of her own diabetes over the last 70 years of attendances at King's College Hospital. She describes vividly aspects of treatment and some of the problems faced by people with diabetes, and one can see clearly how many improvements there have been during her lifetime. Her account should give tremendous encouragement to those now starting on their own life with diabetes.

The ABC is intended as a strictly practical guide to the management of diabetes and its complications and is directed to all those doctors, nurses, and health professionals, other than established specialists, who see diabetic patients, and medical students should find some value in its pages. Many of the innovations of the end of the 20th century are described in this fifth edition of the ABC in the hope that it will help in the delivery of the very best standards of care to those who need it in the 21st century.

1 What is diabetes?

Diabetes once diagnosed is for life. The perseverance and self discipline needed over a lifetime can often tax even the most robust of people to the limit. Those caring for them also require perseverance and an understanding of humanity combined with a cautious optimism, to guide those with diabetes through the peaks and troughs of their lives.

Ebers papyrus: early clinical description of diabetes (Egyptian, 1500 BC)

Definition

Diabetes occurs either because of a lack of insulin or because of the presence of factors that oppose the action of insulin. The result of insufficient action of insulin is an increase in blood glucose concentration (hyperglycaemia). Many other metabolic abnormalities occur, notably an increase in ketone bodies in the blood when there is a severe lack of insulin.

Diagnosis

The diagnosis of diabetes must always be established by a blood glucose measurement made in an accredited laboratory.

Glucose tolerance test

The glucose tolerance test is not normally needed in routine clinical practice, and then only if uncertainty exists in younger patients, or to establish an exact diagnosis in pregnancy. For reliable results, glucose tolerance tests should be performed in the morning after an overnight fast, with the patient sitting quietly and not smoking; it is also important that the patient should have normal meals for the previous three days and should not have been dieting. False results may also occur if the patient has been ill recently or has had prolonged bed rest. Blood glucose concentrations are measured fasting and then one and two hours after a drink of 75 g of glucose in 250-350 ml water (in children 1·75 g/kg to a maximum of 75 g), preferably flavoured, for example, with pure lemon juice. Urine tests should be performed before the glucose drink and at one and two hours. Interpretation of blood glucose values according to WHO criteria is shown in the table.

Gestational diabetes

This term embraces the criteria for both diabetes and impaired glucose tolerance when discovered during pregnancy (see page 80).

Glucose tolerance tests may also show:

Renal glycosuria—this occurs when there is glycosuria but normal blood glucose concentrations; this is a benign condition, only rarely indicating unusual forms of renal disease. It is worth issuing these patients with a certificate to prevent them from being subjected to repeated glucose tolerance tests at every medical examination.

Steeple or lag curve—this is described when fasting and two hour concentrations are normal, but those between are high, causing glycosuria; this is also a benign condition, which most commonly occurs after gastrectomy but may occur in healthy people.

Impaired glucose tolerance

This is defined in the table. Patients are managed at the discretion of the physician. In general, no treatment is given to

WHO criteria for the diagnosis of diabetes

1 Symptoms of diabetes plus casual venous *plasma* glucose ⩾ 11·1 mmol/l. Casual is defined as any time of day without regard to time since last meal. The classic symptoms of diabetes include polyuria, polydipsia, and unexplained weight loss
2 Fasting *plasma* glucose ⩾ 7·0 mmol/l or whole blood ⩾ 6·1 mmol/l. Fasting is defined as no calorie intake for at least 8 hours
3 2 hour *plasma* glucose ⩾ 11·1 mmol/l during oral glucose tolerance test using 75 g glucose load

In the absence of symptoms, these criteria should be confirmed by repeat testing on a different day. If the fasting or random values are not diagnostic, the 2 hour value post-glucose load should be used

Note:
Fasting plasma glucose < 6·1 mmol/l—normal
Fasting plasma glucose ⩾ 6·1 and < 7·0 mmol/l—impaired fasting blood glucose
Fasting plasma glucose ⩾ 7·0 mmol/l—provisional diagnosis of diabetes; the diagnosis must be confirmed (see above)

Adapted from *Diabetes Care* 1997;20:1183-1195

Glucose tolerance test

	Glucose concentration (mmol/l)		
	Venous whole blood	**Capillary whole blood**	**Venous plasma**
*Diabetes mellitus**			
Fasting	⩾ 6·1	⩾ 6·1	⩾ 7·0
2 hours after glucose load	⩾ 10·0	⩾ 11·1	⩾ 11·1
Impaired glucose tolerance			
Fasting	< 6·1	< 6·1	< 7·0
2 hours after glucose load	⩾ 6·7 < 10·0	⩾ 7·8 < 11·1	⩾ 7·8 < 11·1

*In the absence of symptoms at least one additional abnormal blood glucose concentration is needed to confirm clinical diagnosis—for example, 1 hour value of 11 mmol/l or more

elderly people, but diet, exercise and weight reduction are advisable in younger subjects. Over 10 years, approximately half of those with impaired glucose tolerance will develop diabetes, one-quarter will persist with impaired glucose tolerance, and one-quarter will revert to normal. Pregnant women with "impaired glucose tolerance" must be treated as if they were diabetic; for interpretation of the test in pregnancy seen page 80.

Types of diabetes

Type 1 diabetes (previously insulin dependent diabetes) is due to B-cell destruction, usually leading to absolute insulin deficiency). It can be immune mediated or idiopathic.

Type 2 diabetes (previously non-insulin dependent diabetes) ranges from those with predominant insulin resistance associated with relative insulin deficiency, to those with a predominantly insulin secretory defect with insulin resistance.

Type 1 and Type 2 diabetes are the commonest forms of primary diabetes mellitus. The division is important both clinically in assessing the need for treatment, and also in understanding the causes of diabetes which are entirely different in the two groups.

Other specific types of diabetes

- *Genetic defects of β cell function*—chromosome 12 hepatic nuclear factor-1α (HNF-1α) (formerly maturity onset diabetes of the young (MODY) 3), chromosome 7 glucokinase defect (formerly MODY 2), chromosome 20 HNF-4α (formerly MODY 1), mitochondrial DNA mutation
- *Genetic defects in insulin action*—Type A insulin resistance (genetic defects in insulin receptor), lipoatrophic diabetes, genetic defects in the PPARγ receptor
- *Gestational diabetes*
- *Diseases of the exocrine pancreas*—pancreatitis, pancreatectomy, carcinoma of pancreas, cystic fibrosis, fibro-calculous pancreatopathy, haemochromatosis
- *Endocrinopathies*—acromegaly, Cushing's disease, Conn's syndrome, glucagonoma, phaeochromocytoma, somatostatinoma
- *Drug induced* (these agents in particular exacerbate hyperglycaemia in patients with established diabetes)—corticosteroids, diazoxide, β adrenergic agonists (for example, intravenous salbutamol), thiazides, α interferon
- *Uncommon forms of immune mediated diabetes*—stiff man syndrome, anti-insulin receptor antibodies (Type B insulin resistance)
- *Infections*—congenital rubella, cytomegalovirus
- *Other genetic syndromes sometimes associated with diabetes*—Wolfram syndrome, Down's syndrome, Turner's syndrome, Klinefelter's syndrome, Prader-Willi syndrome

Type 1 diabetes

Type 1 diabetes is due to destruction of B-cells in the pancreatic islets of Langerhans with resulting loss of insulin production. A combination of environmental and genetic factors that trigger an autoimmune attack on the B-cells is responsible, occurring in genetically susceptible individuals. Thus, among monozygotic identical twins only about one-third of the pairs are concordant for diabetes in contrast to the situation in Type 2 diabetes where almost all pairs are concordant. The process of islet destruction probably begins very early in life and is known to start several years before the clinical onset of diabetes.

HLA status

The major histocompatibility complex antigens are adjuncts to several types of immunological activity. Ninety percent of Type 1 diabetic patients show either DR3 or DR4, or both together, while DR2 is protective against diabetes.

Autoantibodies and cellular immunity

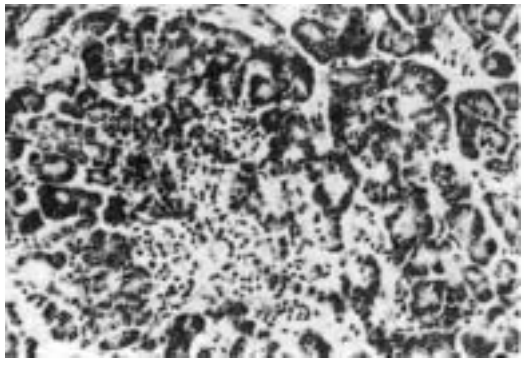
Islet cell antibodies are present at diagnosis in most Type 1 diabetic patients and gradually decline and disappear during the following years. Antibodies to specific proteins have more recently been identified: these include antibodies to glutamic acid decarboxylase (GAD, a 64-kDa antigen); and even closer association is found in the presence of antibodies to tyrosine phosphatase (37 kDa, IA-2). The presence in a non-diabetic individual of three or more antibodies (islet cell antibodies, anti-GAD antibodies, anti-IA-2 antibodies, anti-insulin autoantibodies) indicates an 88% chance of developing diabetes within 10 years.

The presence of insulinitis at the onset of Type 1 diabetes represents the role of inflammatory cells (for example, cytotoxic T cells and macrophages) in B-cell destruction. Macrophages also produce cytokines leading to activation of lymphocytes known to be present at the onset of Type 1 diabetes.

Attempts have been made to prevent the onset of Type 1 diabetes. Immune suppression can to some extent preserve islet function, but permanent remissions are not normally achieved and the treatment is in any case too dangerous for routine use.

Comparison of Type 1 and Type 2 diabetes

Type 1 diabetes	Type 2 diabetes
Inflammatory reaction in islets	No insulitis
Islet B-cells destroyed	B-cells function
Islet cell antibodies	No islet cell antibodies
HLA related	Not HLA related
Not directly inherited	Strong genetic basis (some cases)



An islet with lymphocytic infiltration (insulitis)

The use of nicotinamide to prevent diabetes by altering macrophage function has not proved to be of benefit. Giving insulin itself may conserve islet function; the results of trials are awaited.

Associated autoimmune disorders
The incidence of coeliac disease, Addison's disease, hypothyroidism, and pernicious anaemia are increased in Type 1 diabetic patients, and appear to occur especially in those with persisting islet cell antibodies.

Risks of inheriting diabetes
A child of a mother with Type 1 diabetes has an increased risk of developing the same type of diabetes, amounting to 1-2% by 25 years; the risk is about three times greater if the father has this disease. If both parents have the disease the risk is further increased and genetic counselling should be sought by these rare couples.

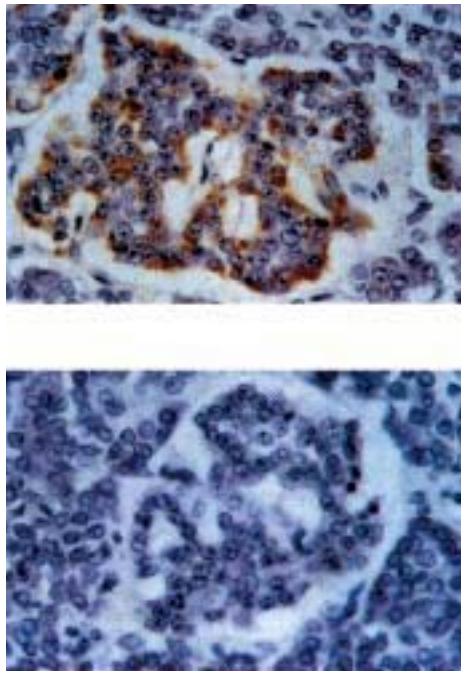
Type 2 diabetes
There are numerous causes of Type 2 diabetes, which is now known to include a wide range of disorders with differing progression and outlook. The underlying mechanism is due either to diminished insulin secretion—that is, an islet defect, associated with increased peripheral resistance to the action of insulin resulting in decreased peripheral glucose uptake, or increased hepatic glucose output. Probably as many as 98% of Type 2 diabetic patients are "idiopathic"—that is, no specific causative defect has been identified. Whether decreasing insulin secretion or increasing insulin resistance occurs first is still uncertain, but the sequence of events may vary in different individuals. Obesity is the commonest cause of insulin resistance. Other rare insulin resistant states are shown in the table.

Some adults (especially those not overweight) over 25 years of age who appear to present with Type 2 diabetes may have latent autoimmune diabetes of adulthood (LADA) and become insulin dependent. Autoantibodies are often present in this group of patients.

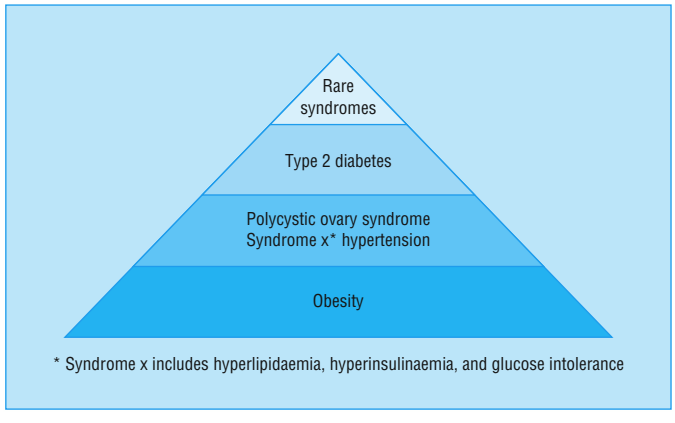
Type 2 diabetes is a slowly progressive disease: insulin secretion declines over several decades, resulting in an insidious deterioration of glycaemic control which becomes increasingly difficult to achieve.

Obesity
Relative insulin resistance occurs in obese subjects, perhaps because of down regulation of insulin receptors due to hyperinsulinaemia. Obese subjects have a considerably increased risk of developing Type 2 diabetes. Fat distribution is relevant to the development of diabetes, so that those who are "apple shaped" (android obesity, waist-hip ratio $>0{\cdot}9$) are more prone to Type 2 diabetes than those who are "pear shaped" (gynoid obesity, waist-hip ratio $<0{\cdot}7$).

The importance of leptin in the evolution of lifestyle related obesity is unclear. Leptin is a single chain peptide produced by adipose tissue and its receptors are expressed widely throughout the brain and peripheral tissues; when injected into leptin deficient rodents it causes profound hypophagia and weight loss. Plasma leptin levels rise in parallel with body fat content. Although very rare cases of morbid obesity due to leptin deficiency have been reported, and are shown to respond to leptin injections, there is in general an absence of measurable biological responses to leptin, which at present has no role in the management of obesity.

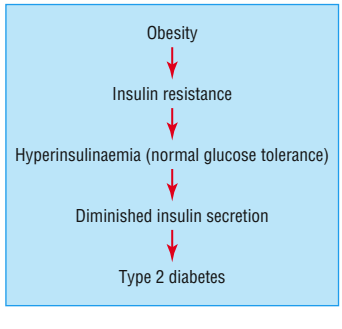


A pancreatic islet after 50 years of Type 1 diabetes: (Top) in this slide A-cells stained for glucagon are intact; (Bottom) in this slide, which is stained for insulin, B-cells are completely absent



Insulin resistance and disease

Increased risk for Type 2 diabetes
- People over 40 years of age
- People of Asian or African-Caribbean ethnic origin
- Overweight people
- Family history of diabetes
- History of gestational diabetes
- History of large baby (birth weight exceeding 4 kg)



Natural history of Type 2 diabetes

Birthweight and Type 2 diabetes
Recent observations suggest a relationship between low birthweight and the development in middle age of insulin resistance, Type 2 diabetes, and coronary artery disease. Those who are smallest at birth and largest at one year of age are most at risk.

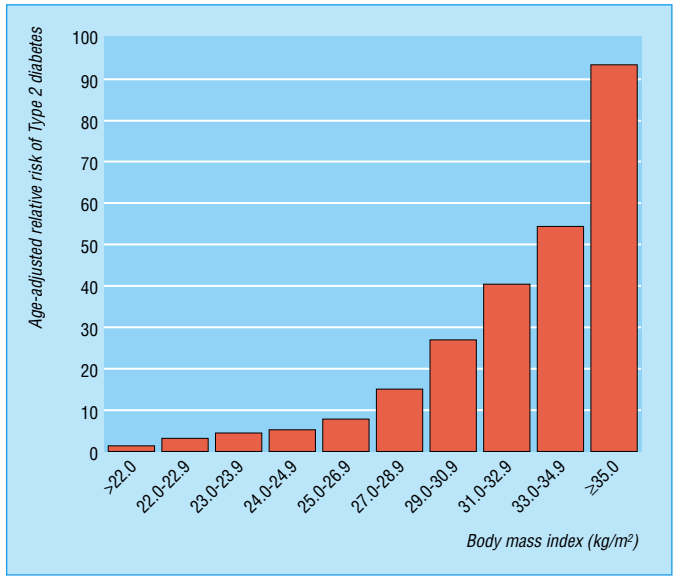
Genetics of Type 2 diabetes
Type 2 diabetes has a strong genetic component, manifest in the high concordance of diabetes in monozygotic twins, familial clustering and differences in prevalence between ethnic groups. An increasing number of specific genetic defects are becoming recognised and some are described below.

Type 2 diabetes in children and young people
Hitherto, childhood diabetes was witnessed in some ethnic minorities and in those with the rare inherited MODY syndromes described below. Growing recognition now exists of a substantial increase of this disease in the prosperous industrialised nations. In the United States, between 8% and 45% of recently diagnosed cases of diabetes among children and adolescents are Type 2, and the problem is increasing. It is most likely to occur at 12 to 14 years of age, more frequently in girls, and is strongly associated with obesity, physical inactivity and a family history of Type 2 diabetes. When young people of lean physique are discovered to have Type 2 diabetes, it is important to attempt to identify whether they may represent those with LADA and thus in need of insulin. There is also evidence that in approximately one-quarter of such patients diabetes is due to a specific genetic defect including those of the MODY group described below or other rare genetic syndromes.

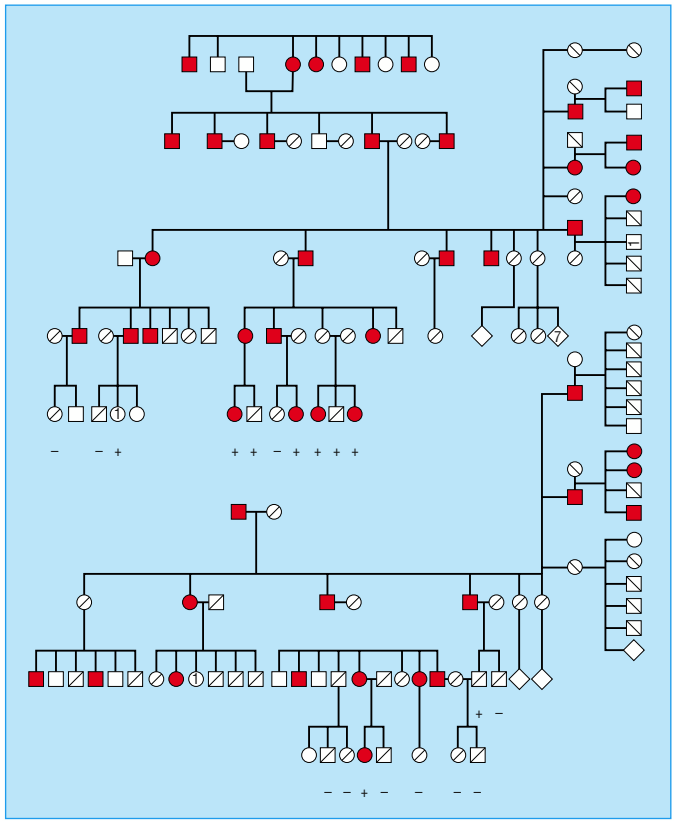
Dominantly inherited Type 2 diabetes (MODY)
Seven genetic syndromes, three of which are shown in the box at the top of page 2, cause MODY—defined as an early onset of dominantly inherited Type 2 diabetes. Two (or at the very least one) members of such families should have been diagnosed before 25 years of age, three generations (usually first-degree) should have diabetes, and they should not normally require insulin until they have had diabetes for more than five years.

Mitochondrial diabetes
Mitochondrial diabetes and deafness is a rare form of diabetes maternally transmitted, and is related to the A3243G mitochondrial DNA mutation. Diabetes is diagnosed in the fourth to fifth decades, usually in thin patients with symptoms. Patients respond better to sulphonylureas than to diet alone. Diabetic microvascular complications do occur.

Insulin resistant diabetes
Some rare insulin resistant states exist in which hundreds or even thousands of units of insulin may be ineffective. They are often associated with lipodystrophy, hyperlipidaemia, and acanthosis nigricans. Type A insulin resistance is due to genetic defects in the insulin receptor or in the post-receptor pathway. Type B insulin resistance occurs as a result of IgG autoantibodies directed against the insulin receptor; it is often associated with other autoimmune disorders such as systemic lupus erythematosis, and it is much commoner in women of African descent. Management of these conditions can be very difficult and specialist texts should be consulted.



Relative risk of Type 2 diabetes according to body mass index in US women aged 30 to 55 years



A family with dominantly inherited Type 2 diabetes. HNF-1α defect (chromosome 12), formerly MODY 3. Diabetic patients are shown in black

Prevalence

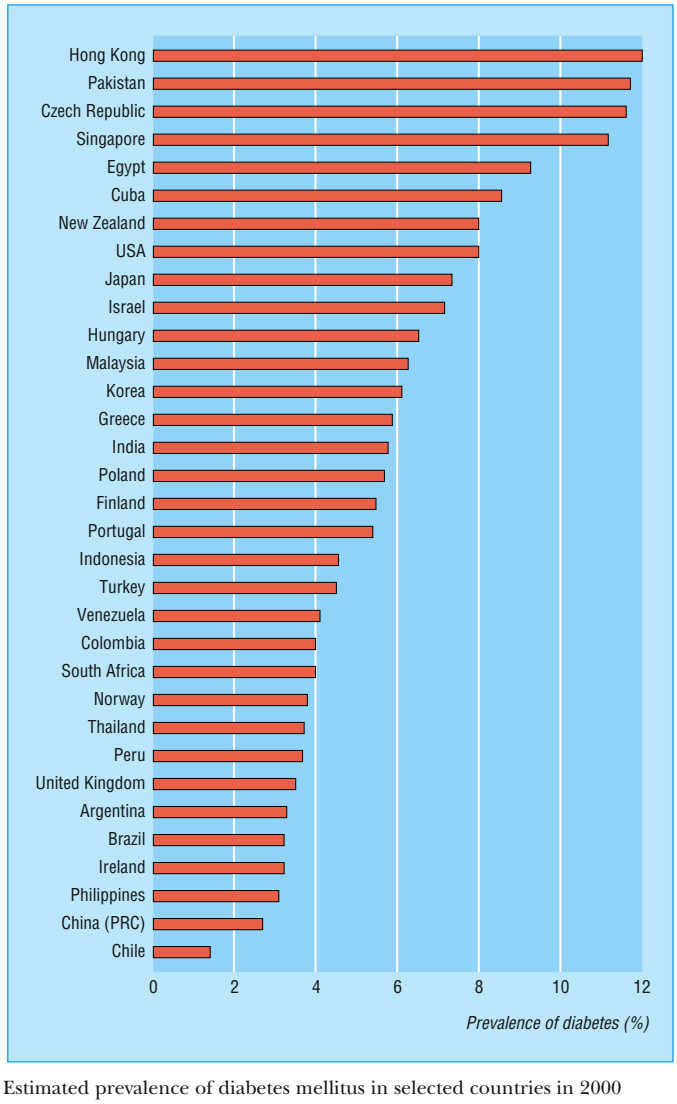
In the United Kingdom more than three percent of the population have diabetes, and about the same number again can be found on screening in population studies. Among schoolchildren about two in 1000 have diabetes.

Diabetes can occur at any age. Type 2 diabetes is most common after middle age and occurs most often at 50-70 years of age, affecting both sexes equally. The peak incidence of Type 1 diabetes is at 10-12 years with a small male predominance. Nevertheless, elderly people can also have Type 1 diabetes, and some children have Type 2 diabetes.

Worldwide, the incidence of Type 2 diabetes is increasing rapidly: in 1995, it was estimated that there were 135 million people with diabetes, this may rise to about 300 million by 2025, increasing particularly in developing countries.

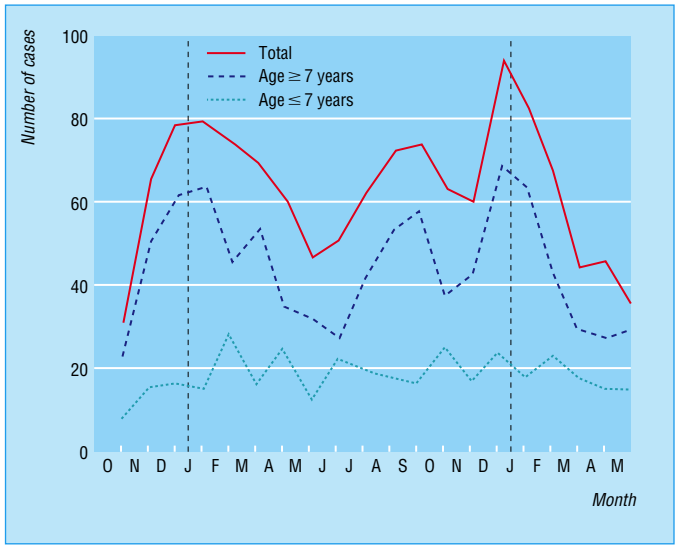
Ethnic variations

The prevalence of Type 2 diabetes is particularly high in Asian and African-Caribbean people and presents a considerable health burden in some inner urban areas. Thus in the United Kingdom 20% of Asians and 17% of African-Caribbeans over 40 years of age have Type 2 diabetes. Children not infrequently have Type 2 diabetes. Asian people have a particularly high risk of developing diabetic nephropathy and coronary artery disease, and a very low risk of foot ulceration; those among the

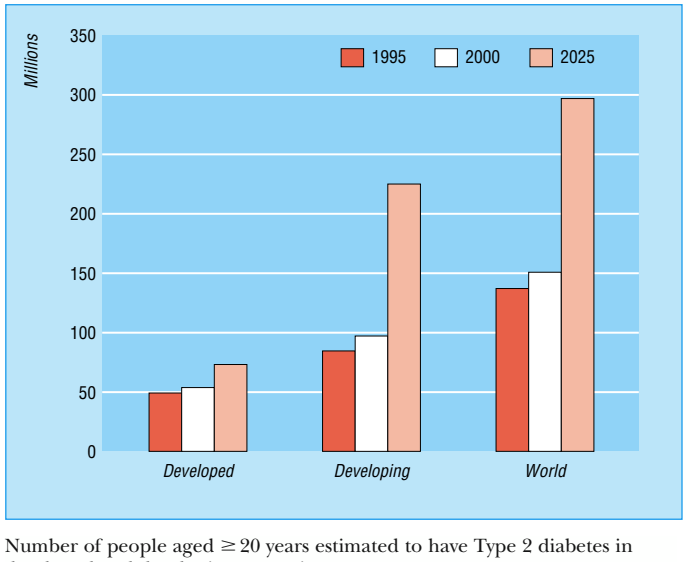


Estimated prevalence of diabetes mellitus in selected countries in 2000

Age of onset of Type 1 diabetes in 3537 children from the British Diabetic Association (now Diabetes UK) register



Seasonal incidence of the onset of Type 1 diabetes showing that nearly three times as many of the older children develop the disorder in the winter months, suggesting some role for viral infections



Number of people aged ≥ 20 years estimated to have Type 2 diabetes in developed and developing countries

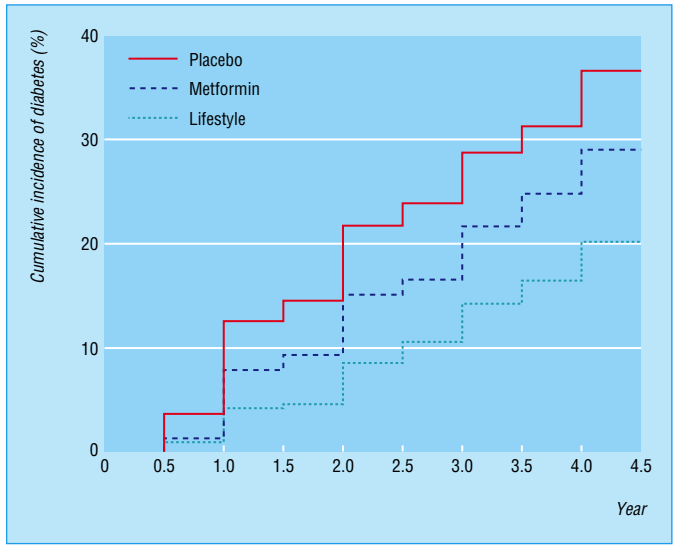
black races are prone to severe hypertension that can be difficult to treat, and also have a strong tendency to develop gestational diabetes.

Prevention of Type 2 diabetes

Lifestyle changes in those prone to Type 2 diabetes can effectively delay the onset of this disease. Several studies in different countries have demonstrated the feasibility of achieving this by a programme of weight reduction, improved diet (less fat, less saturated fat, and more dietary fibre) and increased physical activity. Recent investigations show that the development of diabetes can be approximately halved if these lifestyle changes are maintained over four years.

Diabetic complications

Patients with long-standing diabetes, both Type 1 and Type 2, may develop complications affecting the eyes, kidneys or nerves (microvascular complications) or major arteries. The major arteries are affected in people with diabetes, causing a substantial increase both in coronary artery disease and strokes as well as peripheral vascular disease. The greatest risk of large vessel disease occurs in those diabetic patients who develop proteinuria or microalbuminuria, which is associated with widespread vascular damage. The distribution of arterial narrowing tends to be more distal than in non-diabetic people, whether in coronary arteries or in the peripheral arteries affecting feet and legs. Medial arterial calcification (Monckeberg's sclerosis) is also substantially increased in patients with neuropathy and in those with renal impairment. The functional effects of vascular calcification are uncertain.



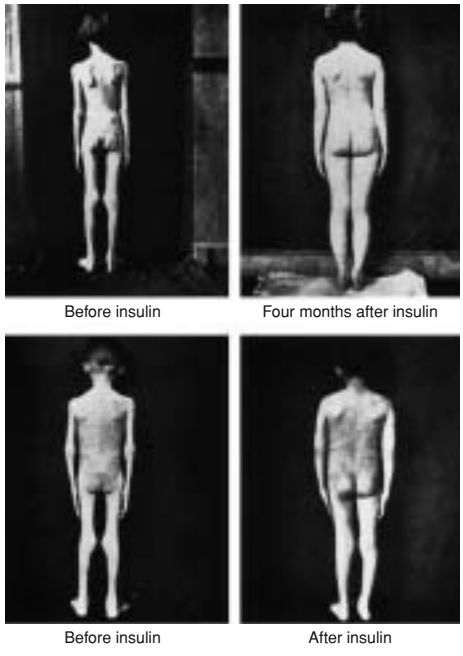
Cumulative incidence of diabetes according to the Diabetes Prevention Programme Research Group. The diagnosis of diabetes was based on the criteria of the American Diabetes Association. The incidence of diabetes differed significantly among the three groups ($P < 0.001$ for each comparison), showing that lifestyle interventions are particularly effective in diminishing the development of Type 2 diabetes

The illustration of an islet cell is reproduced from Gepts W *Insulin: islet pathology, islet function, insulin treatment*, Loft R, ed. Nordisk Insulinlaboratorium. The bar chart showing relative risk of type 2 diabetes according to body mass index in US women uses data from Colditz GA, et al. *Ann Intern Med* 1995;122:461-86. The figure showing a family with dominantly inherited Type 2 diabetes is adapted from Fajans SS, et al. History, genetics and pathogenesis of HNF-4a/MODY1: a 40-year prospective study of the RW pedigree. In *Frontiers in Diabetes*. Basel: Karger, 2000. The age of onset chart is adapted from *Diabetes in Epidemiological perspective*, Mann JI, et al, eds. Churchill Livingstone, 1983. The bar chart showing number of people over 20 years estimated to have Type 2 diabetes in developed and developing countries is adapted from King H, Rogtic G. Global status of diabetes and recommendations for international action. *International Diabetes Monitor*. Copenhagen: IFDOR (Novo Nordisk). The seasonal incidence is adapted from Bloom A, Ireland J, Watkins PJ. *A Colour Atlas of Diabetes*. Wolfe Publishing Ltd, 1992. The estimated prevalence of diabetes in countries in 2000 is adapted from the executive summary of *Diabetes Atlas 2000*, with permission from the International Diabetes Federation. The figure showing cumulative incidence of diabetes according to the Diabetes Prevention Program Research Group is adapted from Diabetes Prevention Program Research Group. *New Engl J Med* 2002;346:393-403.

2 Clinical presentation: why is diabetes so often missed?

Thirst, tiredness, pruritus vulvae or balanitis, polyuria, and weight loss are the familiar symptoms of diabetes. Why then is the diagnosis so often missed? Of 15 new patients with diabetes presenting in our diabetic ward for the first time with ketoacidosis, 14 had had no tests for diabetes after a total of 41 visits to their doctors. Almost all these serious cases of ketoacidosis could have been prevented.

Patients do not, of course, always describe their symptoms in the clearest possible terms, or else their complaints may occur only as an indirect consequence of the more common features. Many patients describe dry mouth rather than thirst, and patients have been investigated for dysphagia when dehydration was the cause. Polyuria is often treated blindly with antibiotics; it may cause enuresis in young people and incontinence in elderly people and the true diagnosis is often overlooked. Complex urological investigations and even circumcision are sometimes performed before diabetes is considered.



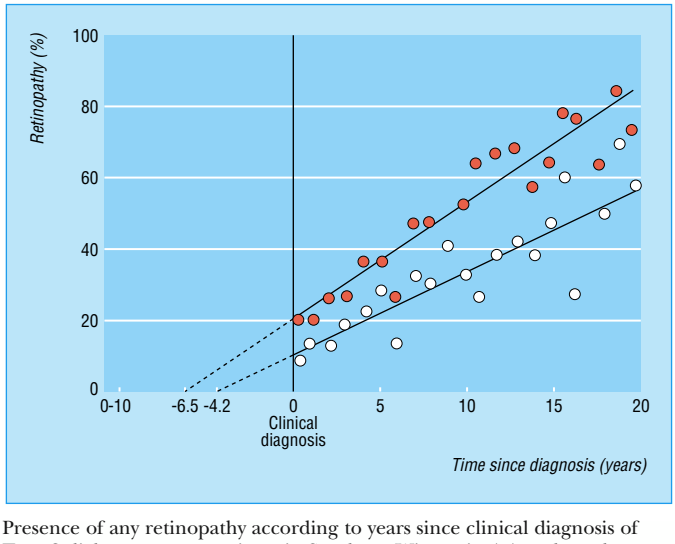
Insulin dependent diabetes, 1922

Confusion in diagnosis

Some diabetic patients present chiefly with weight loss, but even then the diagnosis is sometimes missed, and I have seen two teenagers referred for psychiatric management of anorexia nervosa before admission with ketoacidosis. Perhaps weakness, tiredness, and lethargy, which may be the dominant symptoms, are the most commonly misinterpreted; "tonics" and iron are sometimes given as the symptoms worsen.

Deteriorating vision is not uncommon as a presentation, due either to change of refraction causing myopia (mainly in Type 1 diabetes) or to the early development of retinopathy (mainly in Type 2 diabetes). Foot ulceration or sepsis in older patients brings them to accident and emergency departments and is nearly always due to diabetes. Occasionally painful neuropathy is the presenting symptom, causing extreme pain in the feet, thighs, or trunk.

Glycosuria itself is responsible for the monilial overgrowth which causes pruritus vulvae or balanitis; some older men are first aware of diabetes when they notice white spots on their trousers. In hot climates drops of sugary urine attract an interested population of ants, and at least one patient now attending the clinic at King's College Hospital presented in this way before he came to England.



Presence of any retinopathy according to years since clinical diagnosis of Type 2 diabetes among patients in Southern Wisconsin (●) and rural Western Australia (○). Solid lines represent data fitted by weighted regression; lines are extrapolated to indicate the time at which onset of observable retinopathy is estimated to have occurred, demonstrating that diabetes was likely to have been present for several years before the clinical diagnosis was made

Patterns of presentation

Symptoms are similar in the two types of diabetes (Type 1 and Type 2), but they vary in their intensity. The presentation is most typical and the symptoms develop most rapidly in patients with Type 1 diabetes; they usually develop over some weeks, but the duration may be a few days to a few months. There is usually considerable weight loss and exhaustion. If the diagnosis is missed, diabetic ketoacidosis occurs. Type 1 diabetes occurs under 40 years of age in approximately 70% of cases but can occur at any age, and even in older people.

Presenting symptoms (%) in 547 consecutive cases of diabetes seen by Professor John Malins

				Symptoms							
Age	Males	Females	Total	None	Thirst	Wasting	Fatigue	Pruritus vulvae	Sepsis	Visual	Other
0-39	28	27	55	9	62	14	4	9	2		
40-59	98	108	206	22	22	9	9	21(2)*	7	6	4
60+	100	186	286	22	22	5	8	22(6)*	6	10	5

*Cases of balanitis in males

Symptoms in patients with Type 2 diabetes are similar but tend to be insidious in their onset; sometimes these patients deny any symptoms, although they often admit to feeling more energetic after treatment has been started. These patients are usually middle aged or elderly, but increasingly children, especially those of ethnic minorities, or those who are inert and overweight, are developing Type 2 diabetes. Microvascular and macrovascular complications are frequently already present when Type 2 diabetes is diagnosed. Type 2 diabetes is commonly detected at routine medical examinations or on admission to hospital with another illness.

Type 2 diabetes—presentations

- Diabetic symptoms 53%
- Incidental 29%
- Infections (for example, candida) 16%
- Diabetic complications 2%

Identifying patients in need of insulin

Patients in need of treatment with insulin must be identified early. This is done by judging the patient's clinical features; blood glucose concentrations alone offer a relatively poor guide, although most patients with a blood glucose concentration greater than 25 mmol/l are likely to need insulin.

Features suggesting need for insulin are:

- a rapid development of symptoms
- substantial weight loss—patients are usually thin and demonstrate a dry tongue or more severe dehydration
- weakness
- the presence of ketonuria.

If their condition worsens, vomiting can occur and they rapidly become ketoacidotic; these patients are drowsy, dehydrated, overbreathing, and their breath smells of acetone (although many people are unable to detect this smell).

The following groups of patients are likely to need insulin:

- almost all children and most of those under 30-40 years of age
- women who present during pregnancy
- diabetic patients whose tablet treatment has failed
- all patients who have undergone pancreatectomy.

If there is any doubt give insulin. It can never be wrong to do so, and if the decision was mistaken it can easily be reversed.

Identifying patients in need of insulin

Symptoms
- Rapid onset
- Substantial weight loss
- Weakness
- Vomiting

Signs
- Usually thin
- Dry tongue
- Weak

Ketoacidosis
- Drowsiness
- Dehydration
- Overbreathing
- Breath smelling of acetone

Age
- Any, more likely under 30 years

Blood glucose concentration
- Any

Other indications
- When tablets have failed during pregnancy
- When diet has failed during intercurrent illness
- In patients who have undergone pancreatomy
- Ill patients need admission
- Others may start insulin at home
- If there is any doubt use insulin

Opportunistic screening

The diagnosis of diabetes should no longer be missed. New patients attending their doctor, whether the family doctor, at a hospital outpatient clinic or accident and emergency department, should have a blood glucose measurement as a matter of routine, especially if their symptoms are unexplained. Only a few diabetic patients are wholly without symptoms and their diabetes should be detected by screening at any medical examination. Opportunistic screening for diabetes in this way is a duty.

Della Robbia panel from the Ospedale del Ceppo, Pistoia, 1514

The personal story of Mrs B-J's diabetes

Mrs B-J was born in 1922 and developed Type 1 diabetes at the age of 10 years in 1932. She saw Dr RD Lawrence at diagnosis and in 1989 wrote an account of her own diabetes which will be presented over several chapters in this book.

Presentation and diagnosis

I was always a lively, energetic child so nobody was particularly surprised when I seemed to be growing tall and thin at the end of the summer of 1932. Because of my weight loss, my mother took me to our family doctor who thought I might have TB. He told her to put me to bed for one week, then take me back to him with a urine specimen. Up till then I was perfectly fit and well, but being in bed without exercise, I soon lost my appetite and only wanted oranges and drinks. At night, my mother put water in several quart milk bottles by my bed, but I had drunk it all—about ten pints—before my parents came up to bed. This meant dozens of trips to the toilet each night as a chamber pot could not cope with it.

After about three days, my mother went back to the doctor with a specimen and he said that I had diabetes and must go into hospital the next day. My mother was upset but also relieved, as she had a deep fear of TB.

That night a neighbour called to see how I was. My mother did not realise that I could hear their conversation and told her what was wrong. In a loud, shocked voice this lady asked, "Is she going to die?" I was immediately interested, and hearing my mother say that I wouldn't if I did not eat sweets, cake, biscuits, etc. for the rest of my life, I resolved there and then that I would do just that. I never wanted again to feel as awful as I did just then. I think that eavesdropping probably affected me for the rest of my life, and since then I have had no desire for sweet things except as part of my diet, or for warding off hypos.

The next morning my parents took me by taxi to King's College Hospital.

The first illustration is from Geyelin, HR, Marrop, C. *J Med Res* 1922;2:767-9. The figure showing presence of retinopathy according to years since diagnosis is adapted from *Diab Care* 1992;15:815-21 with permission of American Diabetes Association. The table showing presenting symptoms of diabetes is adapted from Malins J. *Clinical diabetes mellitus*. London: Eyre and Spottiswoode, 1968, and the box showing Type 2 diabetes presentation uses data from UKPDS. *Diabetes Med* 1998;5:154-9.

3 Aims of treatment: a healthy lifestyle

Diabetes is easy to diagnose, but can be managed with negligent ease by those inclined to do so RB Tattersall, 1990

The first concerns in treating diabetic patients are to save life, alleviate symptoms, and enhance the quality of an independent life. Thereafter treatment aims to minimise the long-term complications and reduce early mortality.

Treatment aims

- Save life
- Alleviate symptoms
- Prevent long-term complications
- Reduce risk factors:
  - smoking
  - hypertension
  - obesity
  - hyperlipidaemia
- Educate patients and encourage self-management
- Achieve goals of St Vincent declaration (see page 82)

Aims of treatment

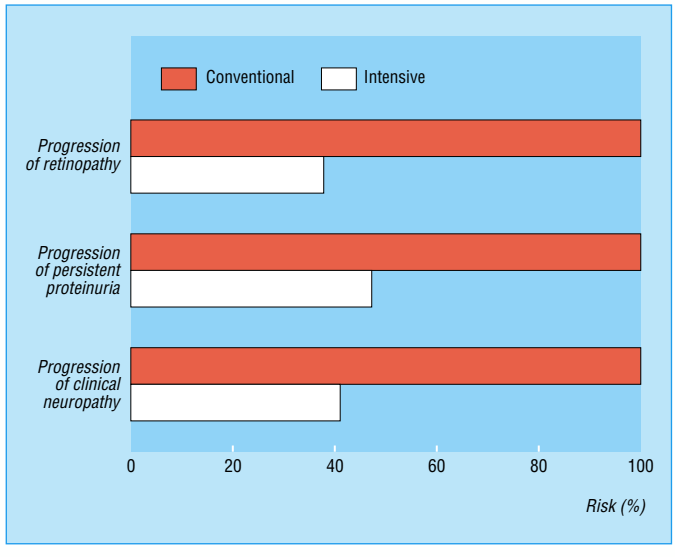
Alleviation of symptoms and improvement in quality of life
This is achieved by reducing hyperglycaemia; patients who need insulin immediately (those with Type 1 diabetes) were described in the previous chapter. All others normally begin on diet alone, moving to diet and oral hypoglycaemic agents, or diet and insulin as indicated. All treatments must be adjusted to ensure that patients are symptom-free. Education of patients plays an important role in enhancing the quality of life, and needs to be maintained over many years.

Maintainenance of health by reduction of risk factors and preventing the development of diabetic complications
The needs here are for:

- achievement of optimal blood glucose control
- detection and control of hypertension
- assessment and control of hyperlipidaemia
- assessment of the need for antiplatelet medication
- cessation of smoking
- regular complications screening procedures (described on page 45).

Management of long-term diabetic complications

Management of other medical problems affecting the patient



Risk reduction for complications in young Type 1 diabetic patients under intensive diabetic control: results from the DCCT

The aims of controlling diabetes

Once diabetes treatment has been established, there is a need to agree the level of control to be achieved in each individual patient. Once symptoms have been eliminated, targets for optimal control (shown in the table) should be discussed and agreed, but it is not always possible to reach ideal goals and pragmatic decisions have to be made.

The following criteria need consideration:

- ensure that symptoms have been eliminated
- lean patients should gain weight
- obese patients should lose weight
- children should grow normally
- prevention of long-term diabetic complications.

Targets for control of diabetes

	Very good*	Acceptable	Less than ideal
Body mass index (kg/m^2)	<25	<27	>27
HbA_{1c} (%) (normal 4.0-6.0)	<6·5	6·5-7·5	>7·5 (>8·0 poor)
Blood glucose in Type 2 diabetes† (mmol/l):			
Fasting	<5·5	<8.0	≥10.0
Postprandial	<9·0	<10·0	≥10·0

*This is the ideal and may be difficult, impossible, or unnecessary to achieve in certain patients (for example, elderly people) Individual targets should be established for each patient
†The optimal range in Type 1 diabetes is about 4·0-9·0 mmol/l

Healthy lifestyle

People with diabetes can help themselves considerably by attention to healthy eating, appropriate exercise levels and weight reduction, and cessation of smoking. These measures

are of great benefit, and may also substantially reduce the need for medication. Behaviour change strategies may be needed to help patients to implement them.

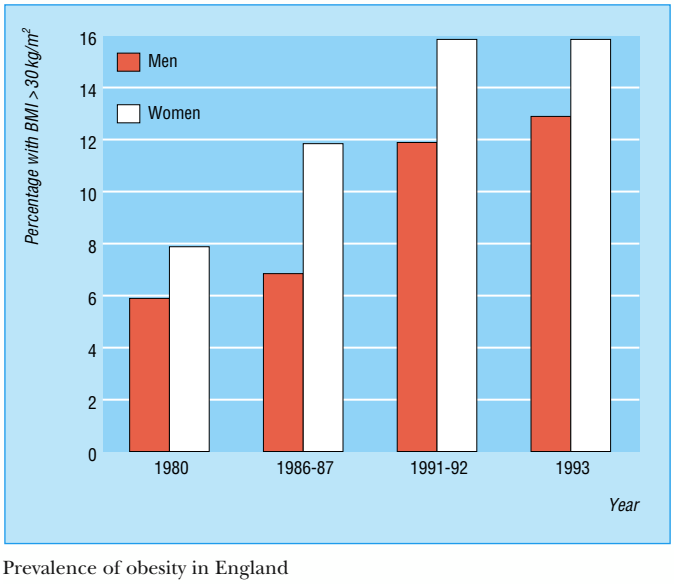
Healthy eating

Healthy eating is the cornerstone of diabetic treatment, and control of the diet should always be the first treatment offered to Type 2 diabetic patients before drugs are considered. Eliminating sugar (sucrose and glucose) lowers blood glucose concentrations in both Type 1 and Type 2 diabetic patients, and although recent dietary recommendations suggest that eating small amounts of sugar is of little consequence, this practice is not recommended. Artificial sweeteners can be used. Good dietary advice is essential to the proper care of diabetic patients; ill considered advice can be very damaging or else it is ignored. I recall one patient who kept to the same sample menu for many years before she reported it to be rather boring. The diet needs to be tailored to the patient's age and weight, type of work, race, and religion.

Recommendations for Type 2 diabetic patients

Diets for overweight Type 2 diabetic patients should aim to eliminate all forms of sugar and restrict the total energy intake. Many of the patients are overweight, and their main goal is to lose weight, although this aim is difficult to achieve. It is important to try to ensure that when patients reduce their intake they do not replace it by an increase of fatty foodstuffs, notably a high intake of cheese. The present emphasis is on reducing total calorie intake, with special emphasis on fat reduction and a proportionately more generous allowance of carbohydrate than in previous years. It has been suggested that as much as half the energy content of the diet may be derived from carbohydrate, while the fat intake is drastically reduced, although these diets in practice require rather difficult and radical changes in the types of food normally eaten. The use of polyunsaturated fats is desirable. These diets are of value and help to reduce blood glucose concentrations if enough fibre is taken. Bran, All Bran, wholemeal bread, and beans have a relatively high fibre content, and are therefore recommended, but foodstuffs with a very high fibre content, such as guar gum, are unpalatable.

For some elderly patients it is enough simply to eliminate all forms of sugar from the diet. Their blood glucose concentrations then fall and symptoms may resolve.



Prevalence of obesity in England

Simple dietary guidelines

- Never take any form of sugar
- Do not take too much fat
- There is no need to restrict most meat, fish, or vegetables
- Control your weight

There is no need to buy proprietary diabetic foodstuffs. Most forms of alcohol (other than sweet wines and liqueurs) are suitable for diabetics, with the usual restrictions for the overweight

A diabetic diet: elimination of sugar/glucose/sucrose

Do not eat or drink:

- Sugar or glucose in any form and do not use sugar in your cooking
- Jam, marmalade, honey, syrup, or lemon curd
- Sweets or chocolates
- Cakes and sweet biscuits
- Tinned fruit
- Lucozade, Ribena, Coca-Cola, Pepsi-Cola, lemonade, or other fizzy drinks

You may use artificial sweeteners, such as saccharin, Sweetex, Hermesetas, Saxin, but NOT Sucron, and any sugar-free drinks including squashes and Slimline range

Fibre content of diet

The following will increase the fibre content of the diet:

Bread	Wholemeal or stoneground—wholemeal for preference If these are not available use HiBran or wheatmeal or granary loaves
Biscuits and crispbreads	Ryvita, Macvita, and similar varieties. Digestive, oatcakes, coconut, and bran biscuits, etc. Crackawheat
Breakfast cereals	Porridge, Weetabix, Weetaflakes, All Bran, Bran Buds, Shredded Wheat, Oat Krunchies, muesli, Alpen, and similar cereals
Wholemeal flour or 100% rye flour	Should be used with white flour for making bread, scones, cakes, biscuits, puddings, etc
Fresh fruit and vegetables	Should be included at least twice daily. The skin and peel of fruit and vegetables such as apples, pears, plums, tomatoes, and potatoes should be eaten
Dried fruit and nuts *Brown rice, wholemeal pasta*	Eat frequently
Pulse vegetables	Such as peas and all varieties of beans

Optimal control may not be needed and it is best to interfere as little as possible with the patient's usual way of life.

Diets for Type 1 diabetic patients

Greater finesse is required in managing the diets of Type 1 diabetic patients; if they eat too much, diabetic control deteriorates; if they eat too little they become hypoglycaemic. The important principles are that carbohydrate intake should be steady from day to day and that it should be taken at fairly regular times each day. If this discipline is not followed diabetic control becomes difficult, although new approaches to the management of Type 1 diabetes such as dose adjustment for normal eating (DAFNE) (see page 29) may permit flexibility in which calculation of carbohydrate intake is used to calculate the insulin dose, thus freeing the patient from a rigidly controlled dietary intake. Severe carbohydrate restriction is not necessarily required; indeed, if the diet is fairly generous patients are less likely to resort to a high fat intake, which may be harmful in the long term.

The actual requirement for carbohydrate varies considerably; it is unsatisfactory to recommend less than 100 g daily, and control may become more difficult if more than 250 g daily is allowed. The smaller amounts are more suitable for elderly, sedentary patients while the larger amounts are more appropriate for younger, very active people particularly athletes who may need considerably more. Although it has been observed that not all carbohydrate-containing foodstuffs are equally absorbed and that they do not have the same influence on blood glucose values, it is impracticable to make allowances for such variations other than recommending that sugar (sucrose) should be avoided except for the treatment of hypoglycaemia.

For social convenience it is customary to advise that most of carbohydrate should be taken at the main meals—breakfast, lunch, and dinner—even though these are not necessarily the times when, according to blood glucose profiles, most carbohydrate is needed; for example, less carbohydrate at breakfast and more at mid-morning and lunch often improves the profile. Snacks should be taken between meals—that is, at elevenses, during the afternoon, and at bedtime—to prevent hypoglycaemia. At least the morning and night snacks are essential and should never be missed.

For the convenience of some, and for those adopting the DAFNE method of controlling Type I diabetes and therefore needing to calculate the carbohydrate content of their meals, 10 g of carbohydrate is described as "one portion" so that a 170 g carbohydrate diet is described to patients as one of "17 portions". Patients sometimes find it valuable to know the carbohydrate values of different foodstuffs.

A sample meal plan for a Type 1 diabetic

	Carbohydrate portions	Recommended food and drink
Breakfast	1	Fruit
	1	Wholemeal cereal
	1	Milk
	1	Wholemeal bread
		Egg/grilled bacon
		Tea/coffee
Mid-morning	1	Fruit/plain biscuit
		Tea/coffee/diet drink
Lunch		Lean meat/fish/egg/cheese
	2	Potatoes/bread/rice/pasta
		Vegetable salad
	2	Fruit/sugar-free pudding
Mid-afternoon	1	Fruit/plain biscuit
		Tea/coffee/diet drink
Dinner		Lean meat/fish/eggs/cheese
	2	Potatoes/bread/rice/pasta
		Vegetable salad
	2	Fruit/sugar-free pudding
Bed-time	1	Bread/fruit/plain biscuit
		Tea/coffee/diet drink
	Total 15	

Alcohol

- Alcohols containing simple sugar should not be drunk by people with diabetes, especially sweet wines and liqueurs
- Dry wines and spirits are mainly sugar-free and do not present special problems
- Beers and lagers have a relatively high sugar and calorie content and their amount needs to be both limited and counted as part of the controlled carbohydrate intake
- Sugar-free beers are high in calorie and alcohol content and therefore have some limitations to their usefulness, whereas "low alcohol" beers are high in carbohydrate
- Profound hypoglycaemia may be provoked in those who take large amounts of alcohol, and omit their normal diet, especially in those taking sulphonylureas; this can be dangerous
- Normal social drinking is usually free from this hazard but care is still needed
- Reduction in alcohol intake is sometimes an important part of helping weight loss

Foods suitable during intercurrent illness

For patients who are feeling ill but need to maintain their carbohydrate intake, the following are useful (each item contains 10 g of carbohydrate):

- ⅓ pint (0.15 l) tinned soup
- 1 glass fruit juice
- 1 scoop of ice cream
- 1 glass of milk

The following each contain 20 g of carbohydrate:

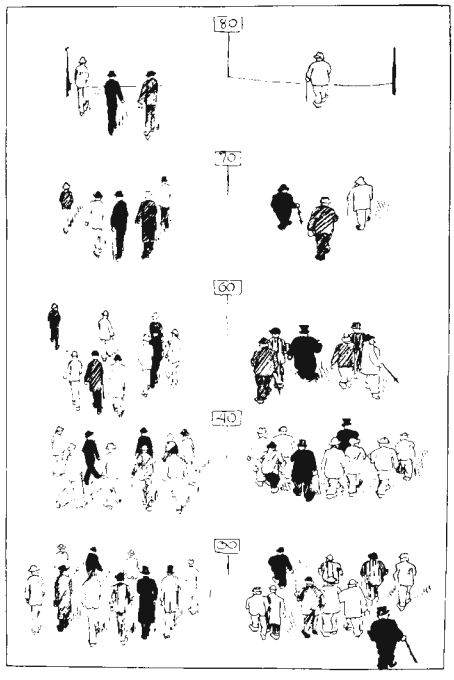
- 2 teaspoons Horlicks and milk
- 2 digestive biscuits
- 1 Weetabix and a glass of milk
- 1 ordinary fruit yoghurt
- "Build-up" made with ½ a pint (0.25 l) of milk and ½ a sachet

Foods suitable during intercurrent illness

Weight control towards optimal levels yields considerable health benefits to all, notably in this context to those who have the combined disadvantages of being overweight and having Type 2 diabetes. Exercise has a central role in weight reduction and health improvement. The proven benefits include reduced insulin resistance (hence enhanced insulin sensitivity) leading to better glycaemic control which may even be independent of actual weight reduction. Risk factors for cardiovascular disease,

which include high blood pressure, also diminish. Indeed, the prevention of Type 2 diabetes itself in those at high risk has been amply demonstrated (see page 6). People with osteoarthritis, chronic heart failure, and chronic lung disease all benefit from appropriate exercise programmes and weight reduction, and there are advantages to those recovering from myocardial infarction. A healthier life is also gained by the very old and by the overweight child. For those with Type 2 diabetes it is recommended that exercise of moderate intensity should be undertaken for about 30 minutes each day. This can include walking, as well as both aerobic and resistance exercise.

The effects of exercise in Type 1 diabetes present the hazard of hypoglycaemia and it is not a specific contributor to improvement of diabetes control. Advice is required on the use of insulin and the need for additional food (in particular carbohydrate) before, during, and after periods of exercise especially (since hypoglycaemia may develop after cessation of exercise) for those engaged in major sports and athletics. The challenge for sportsmen can be extreme but nevertheless people with Type 1 diabetes are known for huge achievements. Great credit went to Sir Steven Redgrave for his ingenious food and insulin regimen which enabled him to win a rowing gold medal in the 2000 Olympic Games.



How 10 fat men and 10 lean men fare on the journey through life (Joslin, 1941)

Smoking

The addiction of smoking is now well established. Its harmful effects are numerous, and include a substantial increase in cardiovascular and peripheral vascular disease as well as the best known consequences of lung cancer and chronic obstructive pulmonary disease. In diabetes, higher rates of both nephropathy and retinopathy have been well documented.

Nicotine replacement therapy using proprietary sublingual preparations, chewing gum, self adhesive patches, or alternatively amfebutamone tablets can help, especially if used in conjunction with the counselling which is provided by smoking clinics. Detailed use of these medications is described in the *British National Formulary* (*BNF*).

The histogram of risk reduction for complication in young Type 1 diabetic patients under intensive diabetic control is adapted from Watkins PJ, et al. *Diabetes and its management*, 5th ed. Oxford: Blackwell Science, 1996. The histogram showing prevalence of obesity in England is adapted from Nutrition and Obesity Task Force. Obesity: reversing the increasing problems of obesity in England. London: Department of Health, 1995. The illustration of how 10 fat men and lean men fare through life is from Joslin EP. *Diabetic manual*, 1941, Lea and Febiger.

The story of Mrs B-J continued: the diet

I was put in a Women's ward and I was given my first dose of insulin. The bed was in the centre of the ward and I soon became the ladies' pet. They threw sweets on to my bed, which I politely refused, no doubt recalling how I had forsworn such poison.

I stayed in hospital for three weeks, and each day I was given lessons in diet. I had a red exercise book in which I set out different diets, stating the weight and value of each carbohydrate item. I had a chart with various foods listed. Those printed in black were called "black lines" and had to be limited by weight to equal the "black lines" allowed at each meal. The protein foods were "red lines" and could be taken ad lib.

4 Treatment of Type 2 diabetes mellitus

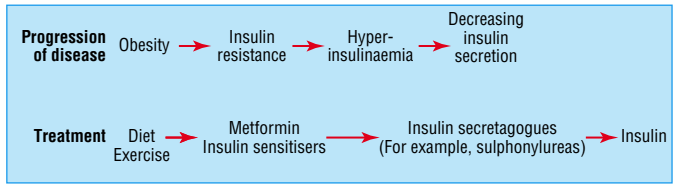
Type 2 diabetes is a complex disorder generally affecting older people who are often overweight and likely to suffer other medical problems as well. Its management presents considerable challenges to medical and nursing staff, whose care must be directed at the sum of the problems of the individual patient. Management now requires not only attention to blood glucose control, but also to the treatment of hypertension and hyperlipidaemia, as well as introducing the necessary measures for reducing cardiovascular risk factors.

Optimal treatment of Type 2 diabetic patients, especially those who are symptom-free, overweight and have in addition several cardiovascular risk factors, exercises our clinical skills and judgments to the limit. There needs to be a sense of reality within the consultation, bearing in mind the potential dangers of unacceptable polypharmacy accompanied by low adherence to prescribed treatment as well as a sense of guilt experienced by those who fail to achieve ideal targets set by physicians. Awareness of the priorities and intentions of individual patients needs to be given consideration, and patients need to agree on the objectives for treatment. Recommendations for treatment must be clinically relevant for the individual patient, who should be involved in choosing which of the many therapeutic options to select after explanation of advantages and risks. The difficulties of controlling Type 2 diabetes tend to increase with the passage of time as the disease progresses. Management is often difficult and needs to be pragmatic: the late Professor John Malins when asked how this should be done used to quote the advice given by Chekhov to his actors—that it should be "done as well as possible".

Approaches to management

There are three distinctive aspects in management, each of which requires entirely different approaches

- To alleviate symptoms and improve quality of life, achieved by reducing hyperglycaemia and weight
- To maintain health by reduction of risk factors (especially hypertension, hyperlipidaemia, and smoking) and by screening programmes to diminish the development of diabetic complications
- Management of diabetic complications
- Management of other medical problems



Natural history of Type 2 diabetes

Glycaemic control

Natural history

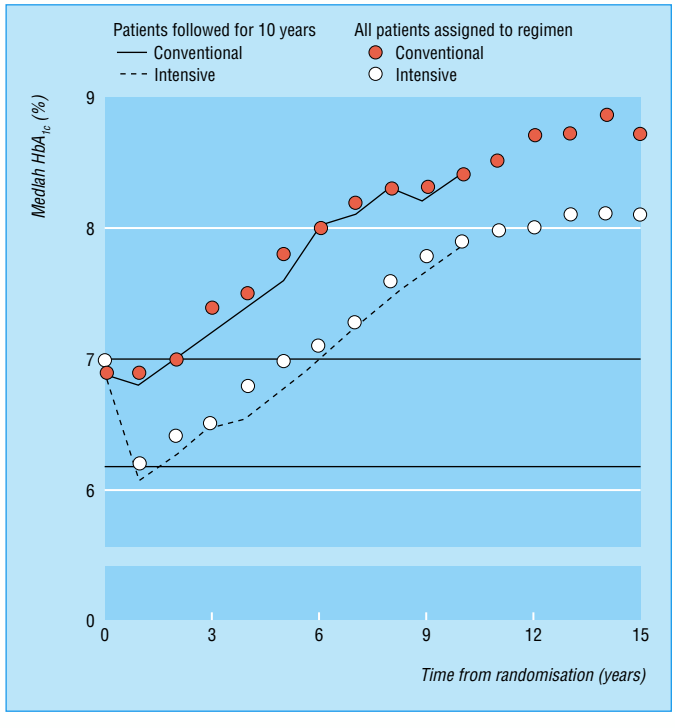
Type 2 diabetes is an insidiously progressive disease. Gradually decreasing insulin secretion leads to a slow increase in hyperglycaemia and a rise of HbA_{1c} values, often despite vigorous clinical attempts to maintain control. Thus, while control during the early years is often straightforward, it becomes increasingly difficult with the passage of time, so that the appropriate need for tablets and insulin requires continuing consideration.

Non-obese patients

Such patients require different consideration from the obese. They are much more likely to require insulin early in the course of treatment, and indeed apparent presentation as Type 2 diabetes may be deceptive when they progress to Type 1 diabetes as cases of latent autoimmune diabetes of adulthood (LADA). Sulphonylurea treatment is used initially while metformin treatment is inappropriate for these patients. Some of them cling desperately to minute diets with the large doses of sulphonylureas as weight and health decline: these patients regain their health rapidly when insulin treatment is started and indeed it should not be delayed.

Obese patients

These patients require a different approach. The need for healthy eating and exercise in an attempt to reduce weight are paramount yet difficult to achieve. When these measures fail,



Cross sectional and 10-year cohort data for HbA_{1c} in patients receiving intensive or conventional treatment, from UKPDS (Type 2 diabetes; see page 42)

metformin is the first choice, and will to a small extent diminish the weight gain which comes almost inevitably with improved glycaemic control. A sulphonylurea or meglitidine analogue is added when metformin alone fails. The use of thiazolidinediones is described below.

Patients who remain unwell and often symptomatic (thirst and nocturia especially) and who continue to lose weight should be switched to insulin without delay.

Achieving glycaemic control and reducing risk factors

- Healthy lifestyle advice—healthy eating plan, exercise, and weight reduction plan.
- Oral hypoglycaemic agents should be given only when dietary treatment alone has failed after a proper trial period, usually lasting at least three months. They should not normally be given as the initial treatment (this is a common error).
  - **Sulphonylureas** stimulate insulin secretion
  - **Meglitidine analogues** stimulate insulin secretion
  - **Biguanides (metformin)** reduce hepatic gluconeogenesis and enhance glucose uptake
  - **Thiazolidinediones** enhance insulin sensitivity
  - **α glucosidase inhibitors (acarbose)** reduce absorption of complex carbohydrates.
- Pharmacological agents to assist weight reduction:
  - **Orlistat** inhibits pancreatic lipase and reduces fat absorption
  - **Sibutramine** is a monoamine reuptake inhibitor, causing reduced appetite
- Antihypertensive and lipid lowering agents (see chapter 17).

Sulphonylureas

Seven sulphonylureas are available. They are remarkably safe and free from side effects, although rare toxic effects have been reported, including rashes and jaundice. Only one sulphonylurea should be used at a time since there is nothing to be gained from any combination of these drugs and there is no evidence that any one drug is likely to be more successful than another.

Selecting a sulphonylurea is largely a matter of personal choice, though it is now usual to use one of the shorter acting, metabolised drugs such as gliclazide or glipizide, which are suitable for all ages and for those with renal impairment as well. Glibenclamide, which has the advantage of once-daily use, is still suitable for younger patients, but is contraindicated in the elderly. Glimepiride is also given once daily and may cause less hypoglycaemia. Excessive doses can cause hazardous (even fatal) hypoglycaemia, and it is thus usual to start treatment with the smallest useful dose. If hypoglycaemia does occur in a patient taking a sulphonylurea, the drug should be stopped or at the very least the dose substantially reduced.

Chlorpropamide is now obsolete. It has a very long half life, thus increasing the risk of hypoglycaemia, and many patients experience an unpleasant facial flush on drinking very small amounts of alcohol.

Meglitidine analogues

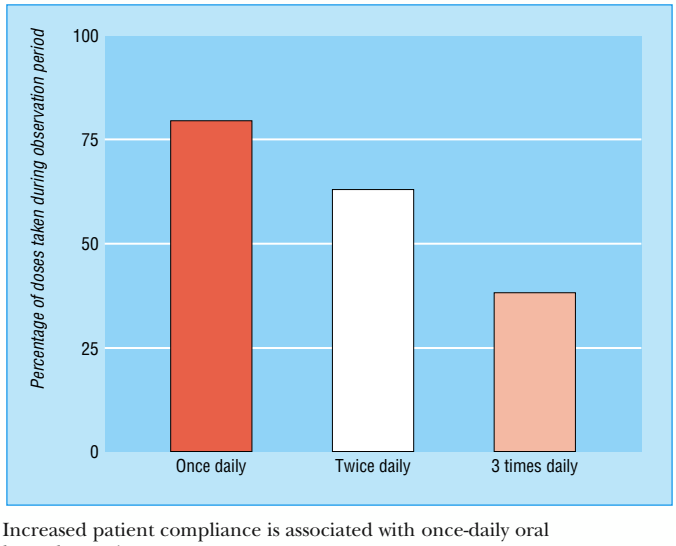
The mode of action of this group of drugs is similar to that of sulphonylureas though acting at a different site. Their advantages are the rapid onset and short half life, efficacy when taken just 15 minutes before meals, and a duration of effect of no more than three hours. They are omitted if no meal is taken. There may be some benefit in reducing postprandial glycaemia and in theory at least there might be less hypoglycaemia.

Daily dose ranges for oral hypoglycaemic agents

Oral hypoglycaemic agents	Dose range (mg)
Sulphonylureas	
Glibenclamide (× 1/day)	2·5-15
Gliclazide	40-320
Glimepiride (× 1/day)	1-4
Glipizide	2·5-20
Gliquidone	15-180
Tolbutamide	500-2000
*Meglitidine analogues**	
Nateglinide	180-540
Repaglinide	1·5-16
Biguanide	
Metformin	1000-2000
Thiazolidinediones	
Pioglitazone (× 1/day)	15-30
Rosiglitazone	4-8
α *glucosidase inhibitors*	
Acarbose	50-600

*Meglitidine analogues are taken three times daily; before meals. Most of the other drugs can be started as single daily doses, but as requirements increase are more effective in divided doses

The hypoglycaemic effect of early sulphonamides was observed in the 1940s, and in the next decade first tolbutamide (1956) and then chlorpropamide (1957) were introduced into clinical practice. They act chiefly by stimulating insulin release from the B-cells of pancreatic islets



Increased patient compliance is associated with once-daily oral hypoglycaemic agents

Nateglinide is one of a new class of oral hypoglycaemic agents, namely an amino acid derivative. Insulin release after meals is both faster and of shorter duration than that with either sulphonylureas or repaglinide, giving less postprandial hyper-insulinaemia and less reactive hypoglycaemia. It is licensed only for use in combination with metformin, but not for monotherapy or substitution for conventional sulphonylureas.

Biguanides: metformin

Biguanides act chiefly by reducing hepatic glucose production. They also enhance peripheral glucose uptake, and to some extent reduce carbohydrate absorption. Metformin is the only biguanide available in the United Kingdom. It is the drug of choice in the treatment of overweight Type 2 diabetic patients when diet alone has failed. UKPDS found some evidence for a reduction in mortality after the use of metformin.

Lactic acidosis is a serious consequence of the inappropriate use of metformin. It is contraindicated in any patient with renal failure, and serum creatinine should be monitored. A creatinine concentration above 150 μmol/l indicates that the drug should be stopped. Metformin should not be used in any seriously ill or shocked patient, nor in those with heart failure, serious liver disease or a very high alcohol intake. It is not appropriate for the treatment of thin diabetic patients nor for use in frail elderly patients.

Use of metformin

- Drug of first choice for overweight Type 2 diabetes
- May reduce mortality (UKPDS)
- Never use when creatinine is >150 μmol/l
- Danger of lactic acidosis if given to:
  - renal failure patients
  - patients with liver disease
  - patients with a high alcohol intake
- Not to be used for thin patients or those in heart failure
- During intravenous contrast procedures:
  - stop metformin for 48 hours beforehand and do not restart until 48 hours after procedure completed

Side effects of metformin

- Nausea
- Diarrhoea
- Metallic taste

These effects can be minimised by starting with a low dose and taking tablets during meals. The effects generally resolve with time

α *Glucosidase inhibitors*

These agents block the enzyme responsible for the breakdown of complex carbohydrates in the gut and can effectively reduce the increase in blood glucose after a meal. Acarbose acts in this way and can be used alone or in combination with other oral hypoglycaemic agents. Its hypoglycaemic effect is relatively small and the severe flatulence which develops (to some extent avoidable by starting with small amounts) deters many patients from using it.

Thiazolidinediones

This newly introduced group of hypoglycaemic agents act by reducing insulin resistance and by activation of the peroxisome proliferator activated receptor γ expressed predominantly in adipose tissue.

These drugs are licensed for use with metformin if this alone has failed to control the diabetes, or with a sulphonylurea if metformin is either not tolerated or contraindicated (for example, in renal failure). In the European Union they are not licensed for use alone or in combination with insulin, and should not be given to patients with a history of heart failure, or during pregnancy. They may cause oedema, a minor reduction of haemoglobin, and a small increase of HDL cholesterol. There are very rare reports of liver dysfunction, and liver function should be monitored before, and every two months after, starting treatment, for the following 12 months.

Thiazolidinediones should be used in patients expected to be insulin resistant, namely those who are overweight and likely to be hyperlipidaemic and hypertensive as well

Drugs for management of obesity

There is a limited place for the use of medication in assisting with weight reduction in the obese diabetic patient. The use of such drugs is restricted to those whose BMI is 28 or more and who are between the ages of 18 and 65 years; they should only be prescribed for individuals who have lost at least 2·5 kg body weight by diet and exercise during the preceding month. Patients should continue to be supported by their advisers and counsellors throughout treatment. Orlistat inhibits fat absorption by inhibition of pancreatic lipase. Weight reduction

Useful drug combinations

- Sulphonylurea (or metaglitidine analogue) with metformin
- Sulphonylurea (or metaglinide analogue) with thiazolidinedione (if metformin is contraindicated or not tolerated)
- Metformin with thiazolidinedione
- Nateglinide with metformin
- Metformin with insulin (for overweight patients)
- Acarbose can be used in association with any of the above

indicating a successful response should be greater than 5% after 12 weeks, in which event prescription may be continued for one year to a limit of two years, otherwise treatment should cease. Unpleasant oily leakage and steatorrhoea can occur.

Sibutramine also acts centrally as a serotonin and noradrenaline reuptake inhibitor and enhances the satiety response. It is used as an adjunct to weight maintenance after weight loss. Full details of its use and contraindications are to be found in the *BNF*.

Guar gum

Guar gum preparations, taken in adequate quantity three times daily before meals, can reduce postprandial blood glucose concentrations. Flatulence is common and often unacceptable. Guarem is the only available preparation. It has a very limited role.

Hypoglycaemia

Only sulphonylureas and meglitidine analogues cause hypoglycaemia, but it should not be allowed to occur at all—it almost invariably indicates excessive dosage. Those most at risk are elderly people who may make dosage errors or fail to take their normal meals. Hypoglycaemia in this situation can be fatal. The shorter acting sulphonylureas cause the least hypoglycaemia and are therefore best for older people (see below). Management of hypoglycaemia is described in chapter 8. β Blockers may not only exacerbate hypoglycaemia, but also occasionally inhibit the early warning symptoms.

Indications for insulin in Type 2 diabetes

Approximately 6% of non-obese and 2% of obese Type 2 diabetic patients need to start insulin each year. Predicting the need for insulin is difficult: those of lean body mass, especially in the presence of islet cell antibodies, are at greatest risk.

Whether to give insulin to Type 2 diabetic patients is one of the most important yet difficult decisions to be made in treating these patients. Failure to give insulin to some patients results in protracted and needless malaise if not actual danger. On the other hand, giving insulin inappropriately can cause needless problems, notably from hypoglycaemia and weight gain.

Indications for giving insulin to Type 2 diabetic patients who are inadequately controlled despite adherence to their recommended diet and oral hypoglycaemic agents are as follows:

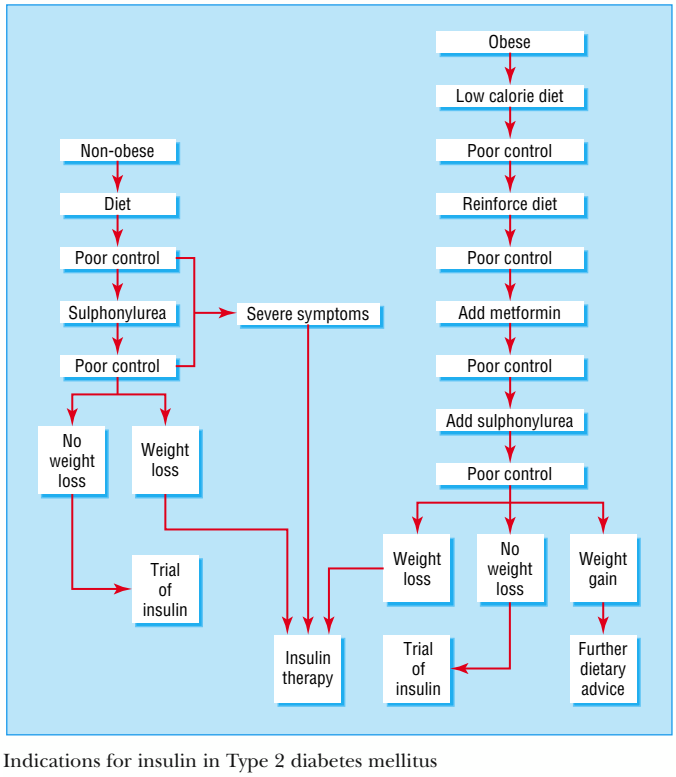
- Continuing weight loss (even if this is insidious), and persistent symptoms, or both. Insulin treatment in these patients almost always results in a substantial improvement in health.
- A non-obese patient without symptoms whose weight is stable and who is conscientious with existing medication. Diabetic control will usually improve, and about half of the patients will enjoy an improvement in well-being.
- An obese patient without symptoms whose weight is stable presents an even more difficult problem. The correct management is to ensure that they are taking their medication, together with intensification of diet, but sometimes insulin may be needed simply to improve control of diabetes in order to reduce long-term complications during the following decade or more. A reduction of HbA_{1C} of approximately 2% together with weight gain of around 5-7 kg can be expected. Unfortunately improvement in glycaemic control is not always achieved. Patient choice is important here, and some prefer not to take insulin after all explanations have been presented. Reluctant patients can be

Drug interactions

- Alcohol can cause serious hypoglycaemia when used with sulphonureas and lactic acidosis in those taking metformin
- Aspirin, sulphonamides, and monoamine oxidase inhibitors may enhance the hypoglycaemic action of sulphonylureas, but in practice problems are rarely seen
- Selective serotonin reuptake inhibitors used in the treatment of depression may provoke hypoglycaemia
- Serious hyperglycaemia is provoked by corticosteroids, dopexamine (inotropic support agent) and intravenous β agonists (salbutamol, terbutaline, ritrodrine)
- Thiazide diuretics (other than minimum dosage, for example, bendrofluazide 2·5 mg) can exacerbate hyperglycaemia
- The immunosuppressive drug ciclosporin can also exacerbate hyperglycaemia
- Protease inhibitors used in the treatment of patients with HIV can cause a syndrome of lipodystrophy, hyperlipidaemia, and insulin resistance leading to severe exacerbation of hyperglycaemia or even causing diabetes
- Clozapine may provoke hyperglycaemia
- β blockers may exacerbate hyperglycaemia or hypoglycaemia depending on dose, concomitant medication, nutritional state, severity of illness, and the patient's age
- Other less common drug interactions are described in the *BNF*

Indications for insulin in Type 2 diabetes

- Insulin is usually contraindicated in overweight patients whose weight is increasing—giving insulin will make this worse
- Patients who continue to lose weight usually need insulin
- Achievement of tight control in order to prevent complications is obviously more appropiate in younger than in older patients, so the patient's age needs to be considered in deciding whether or not to start giving insulin
- Many older patients, however, benefit greatly from insulin treatment, with an improvement of well-being, and insulin should not be withheld on grounds of age alone



Indications for insulin in Type 2 diabetes mellitus

given a three-month trial of insulin and then make their decision, which experience shows to be usually affirmative. Those with a short life expectancy do not necessarily benefit, and those with other medical disorders will require individual consideration.

- Insulin is often required in patients with intercurrent illness. Many disorders, notably infections, increase insulin resistance, leading to the temporary need for insulin. Withdrawal of insulin after recovering from the illness is important provided adequate control is achieved and maintained.

Corticosteroids always exacerbate hyperglycaemia and often precipitate the need for insulin. This should not deter doctors from prescribing them when they are needed.

Combination treatment with insulin and metformin

Metformin can be given together with insulin to overweight Type 2 diabetic patients: this can to a small extent limit the inevitable weight gain following introduction of insulin. A combination of sulphonylureas with insulin gives little benefit and has the added disadvantage that patients must continue with both modes of treatment.

Insulin regimens suitable for Type 2 diabetic patients are described in chapter 5.

The figure showing the cross sectional and 10-year cohort data for HbA_{1c} in patients receiving intensive or conventional treatment is adapted from UKPDS *Lancet* 1998;352:837-53 with permission from Elsevier Science. The histogram showing increased patient compliance is adapted from Paes AH, Bakker AS, Soe Agnie CJ. Impact of dosage frequency upon patient compliance. *Diabetes Care* 1997;20:1512-17.

5 Insulin treatment

I was like a dried tree, but you have given me new life.
An Ethiopian villager, after starting insulin.

The astonishing power of insulin to restore health and well-being to rapidly deteriorating newly diagnosed Type 1 diabetic patients is as remarkable now as it was in 1922. After Banting gave insulin to Elizabeth Hughes in that year, she wrote to her mother that "it is simply too wonderful for words this stuff." Insulin to this day always has this effect; the challenge now is to optimise control in order to maintain health throughout life.

Insulin is also needed to enhance well-being and control in many Type 2 diabetic patients when the natural progression of their disease has lead to loss of optimal control. The potential to reduce the development of long-term diabetic complications as demonstrated by the UKPDS (see page 42) has led to a recent explosion in conversions from tablets to insulin. The difficult decisions which surround the need for insulin in this situation, together with benefits, uses and misuses of insulin have been described in the previous chapter.

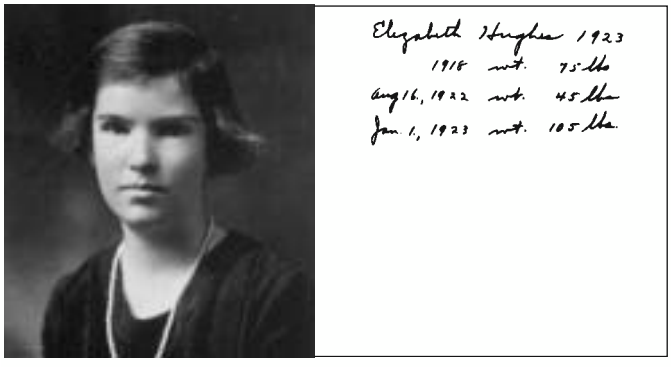
The use of insulin must be tailored to meet individual requirements. The aim is to achieve the best possible control in the circumstances, avoiding at all costs the disabling hypoglycaemia which can occur if control is excessively tight. In some elderly patients and those who lack motivation, it is therefore wise to aim only at alleviating symptoms and not to attempt very strict control.

An Ethiopian patient carrying his diabetes equipment to the clinic

Types of insulin

Soluble insulins
These were first introduced in 1922. They have a rapid onset of action (within 15-30 minutes) and a relatively short overall duration of action of six to eight hours. They play an important part in both daily maintenance of diabetic patients by subcutaneous injection, and also in managing emergencies, when they can be given intravenously or intramuscularly. Other insulin preparations are not suitable for intravenous or intramuscular use.

New recombinant insulin analogues
These have a very rapid onset and very short action, and have been developed by altering the structure and thus the property of the insulin. The preparations available in the United Kingdom



Elizabeth Evans Hughes (1907-1981). Banting's prize patient, who found insulin "unspeakably wonderful." The photograph is from Banting's scrapbook

Insulins available in the United Kingdom

Insulins are available as human, pork or beef preparations, or as insulin analogues.

Very short acting insulin analogues
- Insulin Aspart (Novo Rapid)
- Insulin Lispro (Humalog)

Short acting neutral soluble insulins
- Human Actrapid
- Human Velosulin
- Humulin S
- Insuman Rapid
- Pork Actrapid
- Pork Neutral
- Beef Neutral

Medium acting isophane insulins
- Human Insulatard
- Humulin I
- Insuman Basal
- Pork Insulatard
- Pork Isophane
- Beef Isophane

Medium acting insulin zinc suspensions
- Human Monotard
- Humulin Lente
- Lentard MC (beef or pork)
- Beef Lente

Long acting insulin zinc suspensions
- Human Ultratard
- Humulin Zinc

Long acting insulin analogue
- Insulin Glargine

at present are Insulin Lispro (Humalog) and Insulin Aspart (Novo Rapid). They have some advantages because they may be given immediately before meals (or even immediately after meals if necessary). By virtue of their very short action, there is less hypoglycaemia before the next meal, and when they are used before the main evening meal nocturnal hypoglycaemia is effectively reduced.

There is a risk of postprandial hypoglycaemia if they are used before a meal with a very high fat content because of the delayed gastric emptying. Duration of action is short and does not normally exceed three hours, and their use is therefore inappropriate if the gap between meals exceeds about four hours. Preprandial blood glucose levels are slightly higher than with conventional soluble insulins.

They are also ideal for use in continuous subcutaneous insulin infusion pumps (CSII).

Protamine insulins

These are medium duration insulins introduced in Denmark during the 1930s. Isophane insulin is the most frequently used insulin in this group.

Insulin zinc suspensions

These were first introduced during the 1950s; there are several preparations with widely ranging durations of action. There are limited indications for using insulins with a very long duration of action (ultratard).

Insulin glargine

This is a new prolonged action, soluble insulin analogue (clear solution) forming a microprecipitate after subcutaneous injection. Its onset of action is after about 90 minutes, it has a prolonged plateau rather than a peak, and lasts 24 hours or more. Thus it mimics more closely the basal insulin secretion of healthy people. When taken at bedtime it reduces the incidence of nocturnal hypoglycaemia, and also reduces the prebreakfast hyperglycaemia. It does not appear to reduce symptomatic or severe hypoglycaemia during the day, and there is no significant beneficial effect on overall diabetic control. More extensive clinical experience in using this insulin is still needed.

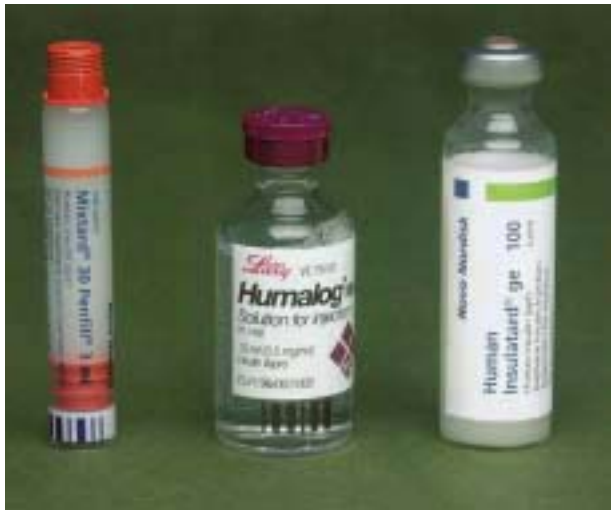
Insulin mixtures

Some preparations of insulin are presented as proprietary mixtures in either vials or pen cartridges, eliminating the need for patients to mix insulins in the syringe. The most popular mixture contains 30% soluble insulin and 70% isophane, whereas the whole range also includes ratios 10%/90%, 20%/80%, 40%/60%, and 50%/50%. These insulin mixtures represent a considerable advantage for many patients, especially those who find it difficult to mix insulins in the syringe or those whose visual acuity is impaired. Details of the types of insulin available in the United Kingdom are shown in the box.

Selection of insulin

The choice of insulin preparation is based on the duration of action. Although insulins can be broadly classified as having very short, short, medium or long duration of action, their effect varies considerably from one patient to another and can be discovered in the individual patient only by trial and error. There are several preparations of medium acting insulins, but those most often used are either one of the isophane preparations or less frequently Human Monotard zinc insulin preparation (see box on page 19).

Most patients (85%) now use insulin of human sequence, a few prefer porcine preparations, while use of some insulin



Some insulins

Insulin mixtures

These are all mixtures of a short acting soluble insulin (or very short acting insulin analogue) with a medium acting isophane insulin (or insulin analogue). The number after the insulin name indicates the percentage of the short acting insulin, for example, "30" or "M3" indicates 30% soluble insulin mixed with 70% isophane insulin.

- Human Mixtard 10 (pen only)
- Human Mixtard 20 (pen only)
- Humulin M2 (pen only)
- Human Mixtard 30
- Humulin M3
- Human Mixtard 40 (pen only)
- Human Mixtard 50
- Humulin M5
- Insuman Comb 50
- Insuman Comb 15
- Insuman Comb 25
- Pork Mixtard 30
- Pork 30/70 Mix

Insulin analogue mixtures

- Humalog Mix 25 (pen only)
- Humalog Mix 50 (pen only)
- NovoMix 30 (pen only)

Insulins are available in vials for use with syringe and needle; in cartridges for use in insulin pens; or in preloaded pens. The insulins listed above are available in one or more of these preparations.

analogues with specific indications is increasing (see also chapters 6 and 8). Some preparations of bovine insulins are still available for the few patients who prefer them.

Insulin regimens

Starting insulin in patients with Type 1 diabetes

Some patients start treatment with twice-daily insulin injections using either a mixture containing premixed short and medium acting insulins twice daily or medium acting insulin alone; 8 units twice daily, 15 to 30 minutes before meals is a suitable initial dose for most patients; others will start with a three or four times daily regimen. Only those who are seriously weakened or ill need hospital admission and treatment either with intravenous insulin and fluids or multiple insulin injections. Many patients who present with acute diabetes enter partial remission soon after diagnosis, known as the "honeymoon" phase, when a small dose of almost any insulin is enough to maintain control. The practice of withdrawing insulin at this stage is not encouraged because after a few months the need for insulin is almost inevitable.

Maintenance regimens

Most Type 1 diabetic patients who want to achieve very good control will need at least thrice-daily injections. Multiple injections (three or four times daily) may improve control, reduce the risk of serious hypoglycaemia, and to some extent increase flexibility (for example, the timing of the midday meal) and are often needed in pregnancy. Suitable insulin regimens are as follows:

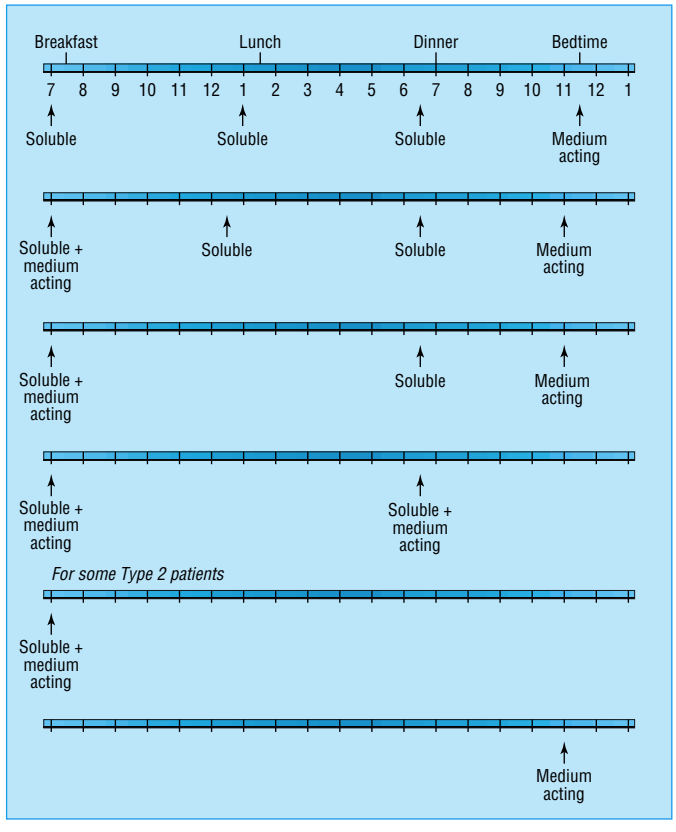
Twice daily: short and medium acting insulins or occasionally medium acting insulin alone are taken twice daily before breakfast and the main evening meal.

Three times daily: the mixture of neutral soluble and medium acting insulins is taken before breakfast; neutral soluble insulin alone before the evening meal; medium acting insulin alone before bedtime. This insulin regimen has the advantage that the noon injection is not required and is thus favoured by many. Fasting blood glucose is also improved using this regimen.

Four times daily: neutral soluble insulin alone or a short acting insulin analogue is taken before each of the three main meals, and medium acting insulin at bedtime. (Occasionally the long acting Human Ultratard insulin is used, though this has not proved to be as advantageous as it should be in theory.) Sometimes there is a further advantage in adding a medium acting insulin to the prebreakfast soluble insulin.

For some Type 2 diabetic patients, when control on oral medication fails, a single daily injection may suffice; the use of medium acting insulin at bedtime alone has gained popularity and by lowering fasting blood glucose may achieve an acceptable profile throughout the day. This regimen can be usefully combined with concurrent use of metformin. If this fails, insulin needs to be delivered on a twice daily basis or more often, as described above. Premixed insulin mixtures are valuable for many Type 2 diabetic patients.

When changing from one insulin regimen to another some trial and error by regular blood glucose monitoring is always needed. In converting a patient to the four times daily regimen the normal dose should be divided by four and a slight adjustment made to give more than one-quarter before breakfast and less than one-quarter before bedtime.



Insulin regimens: short acting insulin analogues can replace conventional soluble insulins

Administration of insulin

Insulin "pens" and syringes

The use of insulin "pen" devices which deliver metered doses of insulin from an insulin cartridge is now favoured by most

Examples of insulin pens

patients. They are portable and simplify the procedure of measuring the insulin dose. The required dose can be dialled, and some pens feature audible and palpable dose graduations which are of value to those with impaired vision. Some versions of the pen are preloaded and disposable.

Plastic insulin syringes with needle attached are still preferred by some patients, and they are still required by those who prefer to mix individual insulins in the syringe. They can be reused several times and between use can be stored in a refrigerator.

Insulin pumps

Sophisticated insulin pumps infuse insulin subcutaneously over 24 hours, with facilities for preprandial boosts. They are worn on a belt and attached to a subcutaneous cannula. They are expensive and not at present available on the NHS although they should be. They are of value for selected patients with Type 1 diabetes. For more detail regarding their use, see page 29.

Insulin pump

Inhaled insulin

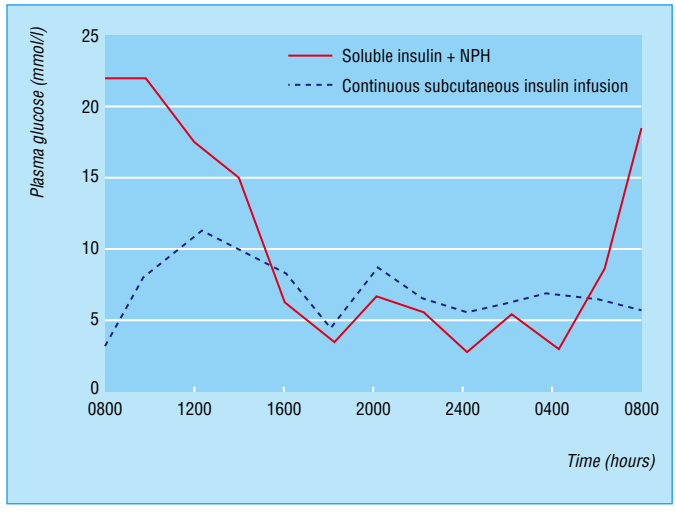
While nasal administration of insulin proved unsuccessful, the use of inhaled insulin looks promising. Absorption though inefficient is adequate to reduce hyperglycaemia, and this route of administration may prove to be of value, notably in Type 2 diabetic patients. The practicality of this technique is still under investigation.

Pancreas transplantation

For those needing a kidney transplantation, pancreas transplantation can be performed simultaneously, eliminating the need for insulin injections and rendering glucose tolerance normal or very nearly normal. Five years after transplantation, 60% of patients remain well without needing insulin injections (see page 70).

Islet cell transplantation

The feasibility of islet transplantation has been demonstrated, and with novel immunosuppression techniques islet survival has improved considerably, eliminating the need for injected insulin over increasing periods of time up to three years. Intensive research of this technique is in progress, and at present it is only available for those participating in carefully organised research trials now conducted in several centres including some in the United Kingdom.



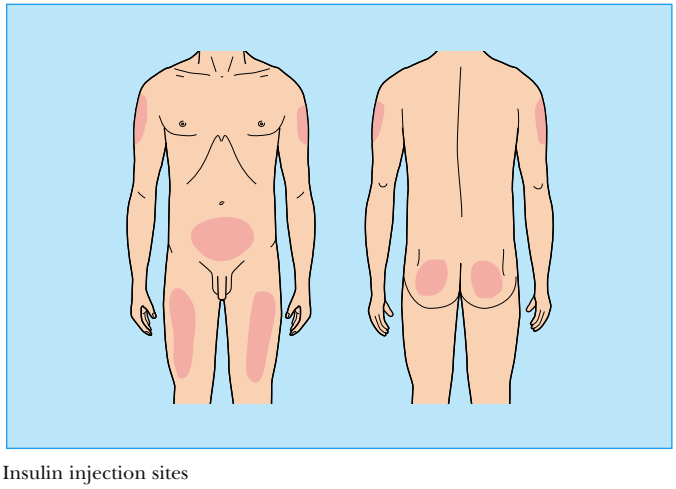
Effect of continuous subcutaneous insulin infusion on plasma glucose

Insulin injection sites

Insulin for routine treatment is given subcutaneously by intermittent injections or by continuous infusion. Insulin can be injected subcutaneously almost anywhere if there is enough flesh. The best site is the front of the thigh. The lower abdominal wall, buttocks, and upper arms may also be used. Patients who want to wear sleeveless clothes should normally avoid using the arms in case unsightly marks or fat hypertrophy should appear; some may then prefer to confine injections to the lower abdomen.

It is important to vary the injection sites from day to day, using for example, each thigh alternately over as wide an area as possible. Absorption of insulin varies from one site to another, being most rapid from the abdominal site, and less rapid from the arms and least from the legs. If there are any difficulties with "control" it is advisable to use one area consistently—for example, the thigh.

In diabetic emergencies soluble insulin is given intravenously, or occasionally intramuscularly (see chapter 9).



Insulin injection sites

Injection of insulin

Drawing up insulin from the vials

1 Clean the top of the insulin bottle with industrial methylated spirit.
2 Draw air into the syringe to the number of units of insulin required and inject this into the insulin bottle.
3 Draw the required dose of insulin into the syringe, and before withdrawing the needle from the insulin bottle, expel the air bubble if one has formed.

If clear and cloudy insulins are to be mixed:

1 Inject the correct number of units of air first into the cloudy insulin bottle.
2 Withdraw the needle from the cloudy bottle.
3 Inject the air into the clear bottle, and withdraw the insulin into the syringe.
4 Finally, insert the needle into the cloudy bottle and withdraw the insulin.

Injecting insulin

1 The skin needs to be clean, but application of spirit, which hardens the skin, is not necessary.
2 Stretching the skin at the injection site is the best way to obtain a painless injection; in thin people it may be necessary to pinch the skin between thumb and forefinger of the hand.
3 The needle should be inserted briskly at 90 degrees to the skin, to its whole length.
4 Inject the insulin by depressing the plunger.
5 Withdraw the needle briskly.

Loading insulin cartridge into pen

Checking dose and expelling the air

Depressing the plunger

Inserting the needle

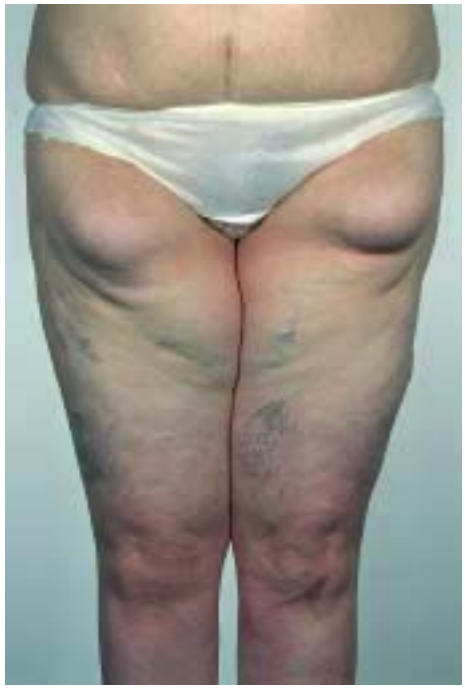
Problems associated with insulin injections

Many patients develop some blurring of vision soon after starting insulin, which makes reading difficult. This is due to a change of lens refraction, and it corrects itself within two to three weeks. Patients should be advised that this may occur, both to avoid extreme anxiety which they may experience, and to stop the needless purchase of new glasses. Transient oedema of the feet is not uncommon during the first few weeks of insulin treatment.

Fatty lumps at injection sites are common, and occasionally so large as to be unsightly. Their cause is not known but they sometimes develop if injections are repeatedly given over a very limited area of skin. For this reason it is best to vary the site from day to day. They are rarely troublesome, but once present they tend to persist; the occasional very large fatty tumour may even require surgical removal. Furthermore if insulin is repeatedly injected into a fatty lump, the rate of absorption may be delayed and this may have some adverse effect on blood glucose control. Fat atrophy at injection sites is now very rare.

Red itchy marks at injection sites after starting insulin are also rare, and if they do occur usually disappear spontaneously. If they are very troublesome, adding hydrocortisone to the insulin bottle so that each dose contains about 1 mg eliminates the problem. Insulin allergy causing urticaria still occurs from time to time though it is certainly a very infrequent event: investigation by skin testing and desensitisation may be needed. Abscesses at injection sites are also remarkably rare.

The illustration of Elizabeth Evans Hughes is from Bliss M. *The discovery of insulin*. Edinburgh: Paul Harris, 1983. The photographs of insulin pens and insulin pump are published with permission from Eli Lilly, Novo Nordisk, and Medtronic MiniMed Ltd.



Fatty lumps at injection sites

The story of Mrs B-J continued: starting insulin

I was put in a Women's ward where I was given my first dose of insulin. I can remember vividly my parents' first visit and my mother's anxious face as she walked down the ward, with one enormous white chrysanthemum in her hand. She had expected to see me prone and white and half dead, not sitting up and a picture of health.

I had a lesson on how insulin burnt up the sugar and produced energy, so that I could return to my former activities, and before long I was doing my own injections. I stayed in hospital for three weeks. The Sunday before I was discharged, my parents were asked to come to the diabetic kitchen to witness me doing my insulin and explaining what I was eating and how it had to be calculated. My mother did not see the injection, having passed out, and she told Sister Wheeler that she would never understand the diet. Sister replied "Don't worry about it. She knows, so give her what she tells you". Such confidence was well founded, as my mother never got the hang of it, and I used to write out the amount of potato etc. before I went to school, and my mother would weigh it up before serving it.

They bought all the necessary equipment from King's when I was discharged, including a very solid metal syringe case which I used constantly until it became redundant with the advent of U100 insulin, and also a copy of RD Lawrence's book "A Diabetic ABC". *I also had a marvellous pair of German-made scales which would weigh up to 2lbs."*

6 Blood glucose monitoring: optimising diabetic control

There are two important reasons for optimising diabetes control: the first is to eliminate symptoms, and the second is the longer-term aim of aborting the development of diabetic complications. Before embarking on complex programmes, it is essential to have a clear view of the requirements of each individual patient. The malaise associated with poorly controlled diabetes almost always responds to better treatment with considerable improvement in well-being. Occasionally, those whose control has been persistently poor for very long periods may for a time feel less well when blood glucose levels are reduced and consequently are at first reluctant to make the effort to improve control.

The key to achieving optimal diabetes control is regular blood glucose measurement accompanied by a clear understanding of the diurnal profiles so obtained. The optimal insulin regimen can then be devised

Blood glucose measurement

Equipment

- A spring-loaded finger pricking device.
- A blood glucose meter and test strips.
- For some, a blood ketone meter combined with a blood glucose meter and test strips.

Note: The MiniMed Continuous Glucose Monitoring System measures interstitial glucose levels every 10 seconds using a sensor inserted under the skin of the abdominal wall; and Glucowatch worn on the wrist is a new technique which repeatedly measures subcutaneous glucose levels. They are still expensive, and more experience in ascertaining their reliability is needed before recommending them for routine use.

Some blood glucose meters

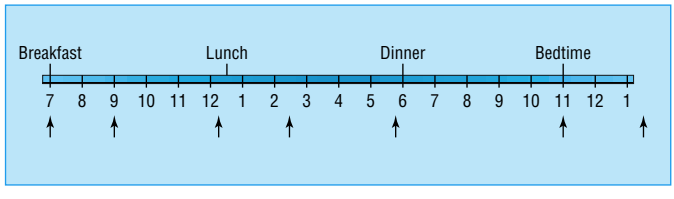
Purpose

- Spot check to detect hypoglycaemia or impending hypoglycaemia.
- Assess control at times of illness.
- Assess blood glucose profile over 24 hours in order to achieve ideal diabetic control.

Apart from the first two indications, isolated blood glucose readings are of little value for optimising control.

Timing and frequency of testing for blood glucose profiles for Type 1 diabetes

- One or two tests should be performed each day as a routine for stable patients.
- Tests should be done at different time points each day to build up a profile over several weeks (see page 27).
- A 12-hour profile can be measured on a single day from time to time, taking recordings before meals (four times a day), one to two hours after meals (three times a day) and at bedtime. Occasionally it helps to record a reading around 3 am.



Times for taking blood

Detecting and eliminating hypoglycaemia (see also chapter 8)
Measurement of blood glucose by patients themselves, or by their relatives, when hypoglycaemia is suspected is the only way of establishing whether or not the blood glucose is actually low. This is of particular value in the assessment of children during periods of bad behaviour, unconsciousness or convulsions. Prediction of hypoglycaemia and therefore prevention is also valuable especially at vulnerable times, notably mid-morning, and at bedtime. Methods of reducing the risks of hypoglycaemia are described in chapters 5 and 8.

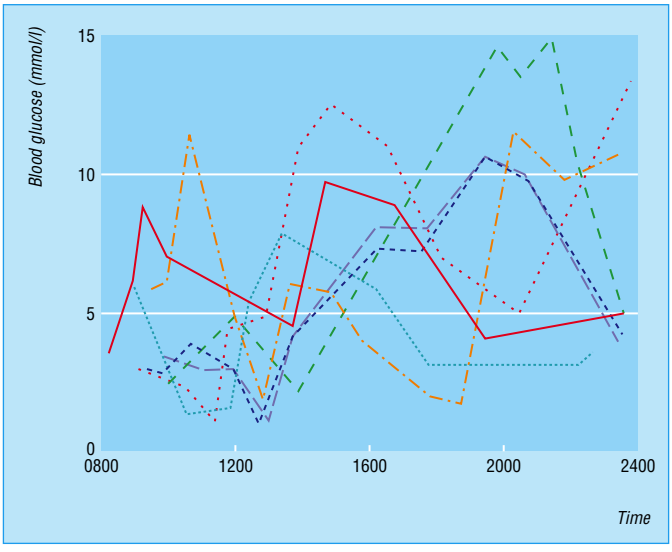
Blood glucose profiles

Assessment by insulin-treated patients of the daily fluctuations of their blood glucose values gives a much greater understanding by patients and doctors of both diabetic control and the effects of different insulin preparations. Indeed, home blood glucose measurement provides an important educational exercise for all seriously motivated diabetic patients as well as being the essential tool to achieve tight blood glucose control. Reproducible blood glucose profiles are essential for making rational adjustments to treatment. They can show not only the times of the peaks and troughs of blood glucose concentration but also the duration of action of different insulin preparations in an individual patient. Unfortunately, those whose lifestyles are chaotic also produce chaotic blood glucose profiles.

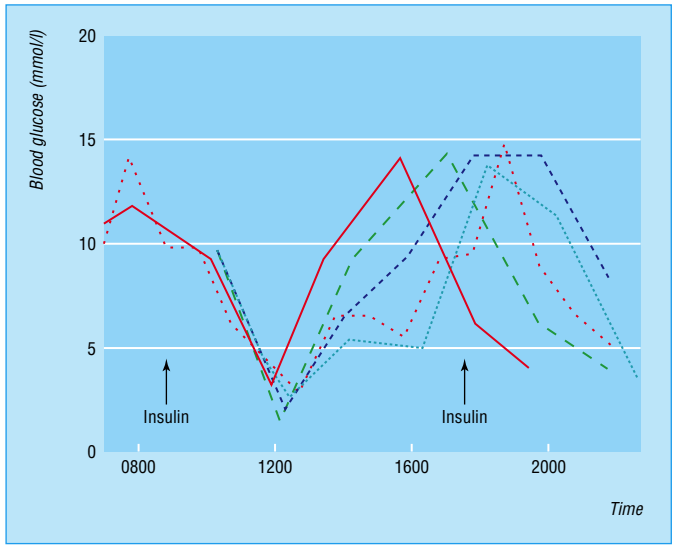
Home blood glucose monitoring by the correct technique, combined with the ability to understand the true significance of the readings, represents on the one hand a very important aspect of diabetes care. On the other hand, obsessional patients who perform tests too often with frequent alterations of insulin dose, cause themselves protracted misery and often disabling hypoglycaemia. While this approach sometimes evolves as a result of the patient's personality, such techniques are all too often encouraged by medical attendants.

The aim is to maintain blood glucose in an acceptable range, usually about 4·5 to 7·7 mmol/l before meals, 6·0-9·0 mmol/l after meals, and >7·0-9·0 mmol/l at bedtime.

Blood glucose monitoring. (A) Inserting the blood glucose strip; (B) loading the finger pricker; (C) pricking the finger; (D) applying the blood to the testing strip; and (E) awaiting result



Chaotic profiles taken by one patient on seven consecutive days



Reproducible profiles taken by one patient on five consecutive days

Occasional excursions outside this range are inevitable, leading to either transient hypoglycaemia or equally transient hyperglycaemia. Thus the occasional high blood glucose readings in an otherwise satisfactory profile can be ignored, although patients are strongly advised to avoid readings below 4·0 mmol/l which if frequent can lead to impaired warning symptoms of hypoglycaemia.

Patients should not respond to isolated high blood glucose readings by taking extra insulin: this causes worsening of blood glucose oscillations rather than an improvement in their blood glucose profile.

Preparing blood glucose profiles

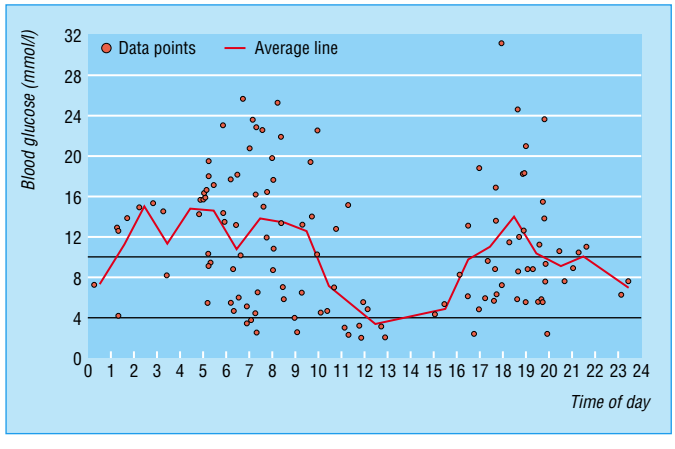
- From patients' record books: visual scanning of records can often detect times of peak and trough readings. Some patients plot their readings graphically which can be very helpful. A few patients present spurious readings which may be difficult to detect.
- Computer-generated profiles: memory records from some blood glucose meters can be downloaded using specially designed computer programs. They can show an excellent visual presentation of readings of any selected time period and may clearly delineate peaks and troughs through 24 hours. Some meters offer this programme on an inbuilt screen.

Blood glucose recording book

Interpretation of blood glucose profiles

Before adjusting insulin treatment, it is essential to understand the causes of fluctuating patterns of the blood glucose profile through 24 hours. The following points are crucial.

- Blood glucose rises as insulin action declines, even when no food is taken, because of hepatic gluconeogenesis. This accounts especially for the rapid increase in blood glucose which occurs in the small hours of the morning *before* breakfast.
- These rapid changes in blood glucose also explain why so many patients record different blood glucose readings each day, since even a half to one hour difference in timing can give a very different result.
- Troughs in the blood glucose profile—that is, representing a tendency to hypoglycaemia—almost always occur around noon and between 3 and 5 am at peak insulin activity, so that measures need to be taken to avoid hypoglycaemia especially preceding these times.



Computer-generated blood glucose profile. Modal day—data points with average line. Patient tends to measure blood glucose at extremes. Target range: high 10.0, low 4.0

- When patients perform three or four isolated blood glucose readings over 24 hours, it is essential for them and their advisers to understand what happens to the blood glucose profile between the single readings. Thus, readings taken at points A, B, C, and D give a very different impression from those taken at points P, Q, R, and S, yet they belong to the same profile.

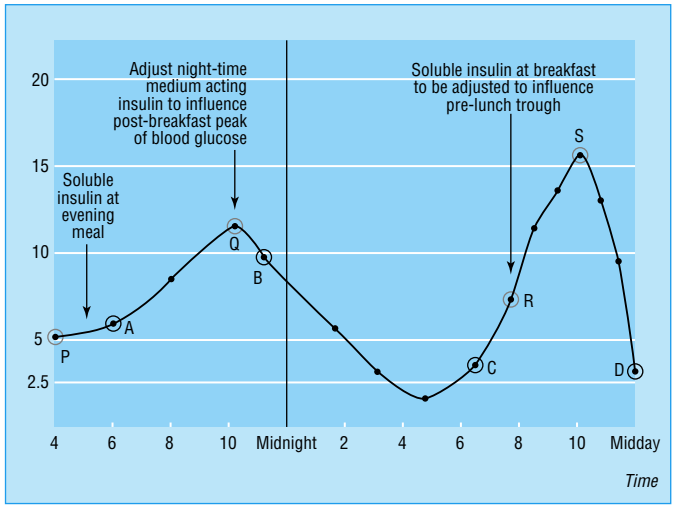
Guidelines for adjusting treatment

- Changes in insulin dose should be made only once or twice weekly except in times of illness when more frequent changes may be needed, or in those following dose adjustment for normal eating (DAFNE) (see page 29).
- Changes in insulin dose at any one time should normally be kept within 10% of the existing daily dose—for example, a change of four units may be made in a patient taking a total daily dose of 40 units.

The following issues need consideration:

- type of insulin (considered in detail on page 19)
- frequency of administration of insulin
- dose of insulin in units
- carbohydrate distribution.

The patterns of the blood glucose profile need to be understood and are much more important than single, randomly taken readings. When consistent daily fluctuations of blood glucose have been shown, treatment should be modified, aiming chiefly to eliminate hypoglycaemic episodes, and thereafter to obtain better control by increasing the blood glucose in the troughs and decreasing it at the peaks. There are several ways of achieving fine glucose control.



Adjustments after assessment of blood glucose profiles. Note the apparent different shape of the profile if readings are taken at points ABCD or PQRS

To increase blood glucose in the troughs

- Eat more carbohydrate at or before the times when blood glucose values are at their lowest, usually mid-morning and at bedtime; the exact amount of extra carbohydrate can be determined only by trial and error
- Reduce the dose of insulin before the trough
- Premeal hypoglycaemia can be ameliorated by substituting short acting insulin with a very short acting insulin analogue

To decrease blood glucose in the peaks

- Reduce by a little the amount of carbohydrate taken at the meals which precede the peaks by two or three hours
- Increase the dose of insulin before the peak

Soluble insulin should be altered to change blood glucose concentrations during the following six hours. Medium acting insulin should be altered to change blood glucose during the following six to 12 hours. The duration of insulin action varies considerably however in individual patients

To decrease fasting hyperglycaemia

- Increase predinner medium (or long) acting insulin.
- If that provokes night-time hypoglycaemia, then split the predinner insulin into two parts, retaining the short acting insulin before dinner and taking medium acting insulin at bedtime; or consider changing the medium acting insulin component to insulin glargine

To lessen nocturnal hypoglycaemia

- Reduce the evening medium acting insulin dose
- Check that the patient is taking their bedtime snack
- Check bedtime blood glucose, taking additional carbohydrate if it is less than 5·0 to 6·0 mmol/l
- Split predinner insulin dose, or consider changing to insulin glargine

For a detailed description of the use of individual insulin types, see chapters 6 and 8

Self assessment of diabetic control in Type 2 diabetes

Home blood glucose monitoring is of value for many (though not necessarily all) patients with Type 2 diabetes, as it is for those with Type 1 diabetes. Those taking oral hypoglycaemic agents or on diet alone have the option to monitor their control by either self blood glucose measurement or regular urine testing. Measurement of the fasting blood glucose two or three times weekly in those on diet alone provides a valuable guide to control, while the addition of some postprandial readings in those taking oral hypoglycaemic agents also provides important information.

Continuous subcutaneous insulin infusion (CSSI)

CSSI was introduced 25 years ago by workers at Guy's Hospital in London, and now the development of more reliable and more sophisticated pumps brings distinct advantages in specific indications to approximately 2 to 5% of those with Type 1 diabetes. A small improvement in overall diabetic control compared with optimised injection regimens can be achieved without necessarily aggravating or indeed actually reducing problems from hypoglycaemia. CSSI is not suitable for those with psychological or psychiatric problems.

Hazards

Experience has virtually eliminated earlier hazards of excessive problems from hypoglycaemia or, because of the very small insulin depot, higher rates of diabetic ketoacidosis. All patients should have a supply of insulin pens or syringes in case of pump failure.

Infusion strategy

- Initially reduce total daily insulin dose by 30%.
- Give half the daily insulin dose as the constant basal pump rate (usually around 1 unit/hour).
- Give half the daily insulin dose divided between the three main meals, giving the insulin boost immediately before the meal.
- The patient is taught to count carbohydrate portions (see page 12) and thereafter will give the bolus doses in direct relation to the amount of carbohydrate consumed (for example, 1 unit for every 10 g of carbohydrate).

During the first few days adjustments need to be made as follows:

- basal rate determined by assessment of fasting and 3 am blood glucose readings
- preprandial boosts are adjusted by assessment of postprandial blood glucose readings.

Note: Specific instructions are given for exercise, and basal rates should be reduced during and after exercise.

Indications for CSSI

- Management of patients with frequent unpredictable hypoglycaemic episodes
- For control of the dawn hyperglycaemic phenomenon when conventional, optimised regimens have failed
- For greater flexibility of lifestyle
- In pregnancy when conventional methods fail
- For patients employed on shift rotas who are not able to achieve glycaemic control on multiple injections

Optimal results are achieved using non-associating, monomeric insulin analogues, for example, lispro insulin (Humalog), or Insulin Aspart (Novo Rapid)

Training of patients is undertaken by the nurse educators and a 24-hour duty scheme is needed to deal with emergencies

Dose adjustment for normal eating (DAFNE)

A more liberal dietary pattern for Type 1 diabetic patients has become possible by using the DAFNE approach, ideal for some people who thus regain considerable freedom while at the same time maintaining good control. It is based on:

- a 5-day structured, group education programme delivered by quality assured diabetes educators
- the educational approach is based on adult educational principles to facilitate new learning
- two injections of medium acting insulin each day (see page 21)
- injections of short acting insulin every time meals are taken
- testing blood glucose before each injection.

This programme enables people to eat more or less what they like when they like, and not to eat if they do not wish to do so. It depends on a quantitative understanding of the carbohydrate values of individual foods, and calculating by trial and error the correct amount of soluble insulin needed for a specified quantity of carbohydrate, developing an insulin/carbohydrate ratio for each individual patient.

DAFNE has been used in continental Europe for many years: the approach is popular and gives considerable benefits to some patients. Good diabetic control can be achieved without any increase in hypoglycaemia and at the same time there is improvement in the quality of life.

The photographs of blood glucose meters and the finger pricking devices are reproduced with permission from MediSense and Roche Diagnostics.

The story of Mrs B-J continues: attending the clinic

Then I began my regular visits to the clinic, which I think was every three weeks at first. This was held upstairs in the pathology laboratory, with the doctors in a small room off. I was fascinated by all that was going on there; so many bottles, test tubes, Bunsen burners, etc., as the tests were done there for all the hospital. The waiting room was very small and if a dozen patients attended it was a crowd, for there were only about 200 diabetic patients on the hospital's roll.

I remember Dr RD Lawrence vividly. He always looked very smart in black jacket and striped trousers and a black bow. He took a great interest in the children, and I felt encouraged that such a great man was also diabetic.

Over the years the number of patients attending the clinic increased. We often had long waits and were obliged to listen to very gruesome tales from some of the adults. Shortly afterwards, the children's clinic was arranged, and I attended with quite a few others on Saturday mornings.

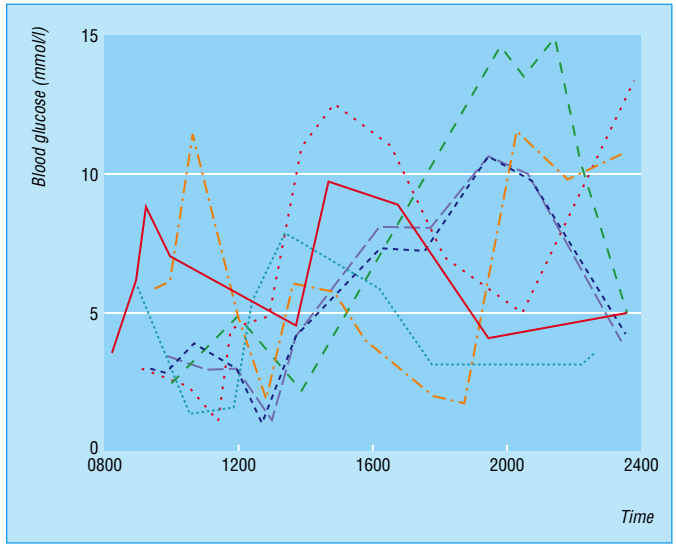
There were "elevenses" laid on, and an ice cream could be had as a portion. At first, ice cream was not considered suitable for diabetics even as part of the diet, so I was happy to find this was no longer so, and a Walls twopenny brick was a real treat.

7 The unstable Type 1 diabetic patient

Blood glucose concentrations inevitably oscillate considerably over 24 hours in many Type 1 diabetic patients. If these swings are used as a definition of instability then such patients might be classified as unstable. Indeed the ardent desire of some doctors to "stabilise" these patients sometimes leads the patient to undertake innumerable blood tests, to keep obsessional records, and to make themselves thoroughly miserable. The failure to succeed leads to recriminations, admissions to hospital, and absence from work. This form of physician-induced, unstable diabetes is made worse by the inappropriate use of home blood glucose monitoring. It needs considerable patience to unravel the effects of such advice, but a more relaxed approach, together with fewer tests, can have a remarkably beneficial effect.

Very unstable diabetes (sometimes described as "brittle") disrupts the lives of a small group of insulin treated diabetic patients, with repeated admissions to hospital due either to hypoglycaemia or ketoacidosis. Homelife, school, and work are totally disrupted. With very few exceptions, this is probably not a special type of diabetes; it most commonly occurs in teenage girls, it is almost always temporary, and problems appear to vanish as life itself stabilises with employment or marriage.

Management of disruptive diabetes demands time and patience; the doctor must identify any technical errors, recommend the best possible diabetic treatment, search for intercurrent illness, and seek social or psychological problems which might cause the patient to manipulate his or her diabetes. Some elderly patients also experience serious problems from violent swings of blood glucose. Loss of support at home following separation or bereavement can be added to the specific problems already described.



Chaotic blood glucose profiles in one patient over seven consecutive days

Disruptive diabetes has several causes, ranging from simple technical errors to gross deceptions of great ingenuity

Solving technical problems

After all the technical issues have been checked, the dose and type of insulin should be adjusted to the best possible regimen (ideal insulin regimens have been described elsewhere). Some obsessional patients respond well to a reduction of the number of daily injections. A few unstable diabetic patients may benefit from continuous subcutaneous insulin infusion, which may also alleviate unpleasant hypoglycaemic episodes.

If recurrent hypoglycaemic episodes are the chief problem then careful education is needed to eliminate them (see chapter 8); careful attention needs to be given not only to the dose of insulin but also to the timing and amount of food, the effects of exercise, and the judicious use of home measurement of blood glucose. Sometimes excessive amounts of insulin, especially soluble insulin, may cause severe hypoglycaemia. Improvement results either from reducing the dose or changing the insulin regimen.

In a few women menstruation regularly causes severe upset of diabetes; control usually deteriorates in the premenstrual phase, causing ketoacidosis at times, followed by an increase in insulin requirement and sometimes troublesome hypoglycaemia. A carefully planned campaign of insulin adjustment usually overcomes this problem.

Identifying technical problems

- The technique of injecting insulin should be meticulously checked
- Injection sites should be inspected
- Equipment needs to be scrutinised
- Sometimes, especially in elderly patients, reduced visual acuity makes measurement of the insulin dose extremely inaccurate
- The brand of insulin itself should be checked
- Techniques of blood glucose testing must be observed and checked with laboratory results
- Adequate understanding of diet should be verified

Above all, patients need encouragement and restoration of self-confidence together with the reassurance that they are neither physically nor mentally abnormal. The telephone number of the doctor or nurse offers added security. If at all possible unstable patients should not be admitted to hospital. If all these measures fail, however, and life is still disrupted by diabetes, then an admission is after all required.

Admission to hospital

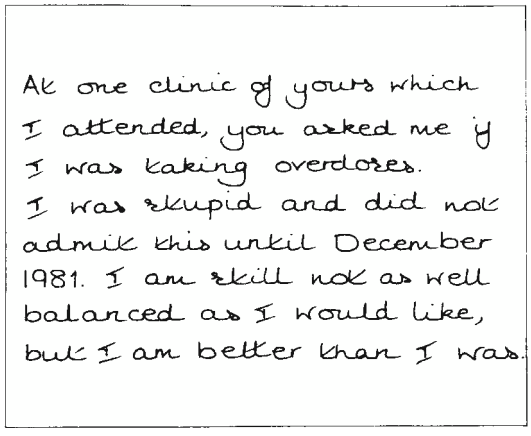
In hospital the nursing staff take over the administration of insulin completely—both the procedure of drawing up the insulin and giving the injections. If some measure of stability is then achieved the patient's equipment is returned for self injection: if chaos resumes it seems likely that the patient is either incompetent or cheating.

If diabetes continues to cause disruption even when the nursing staff are giving insulin injections, some form of manipulation should be suspected. Some patients use great ingenuity; insulin may just be concealed in a locker, but it has also been found inside transistor radios, in the false bottoms of jewellery boxes, and taped outside hospital lavatory windows.

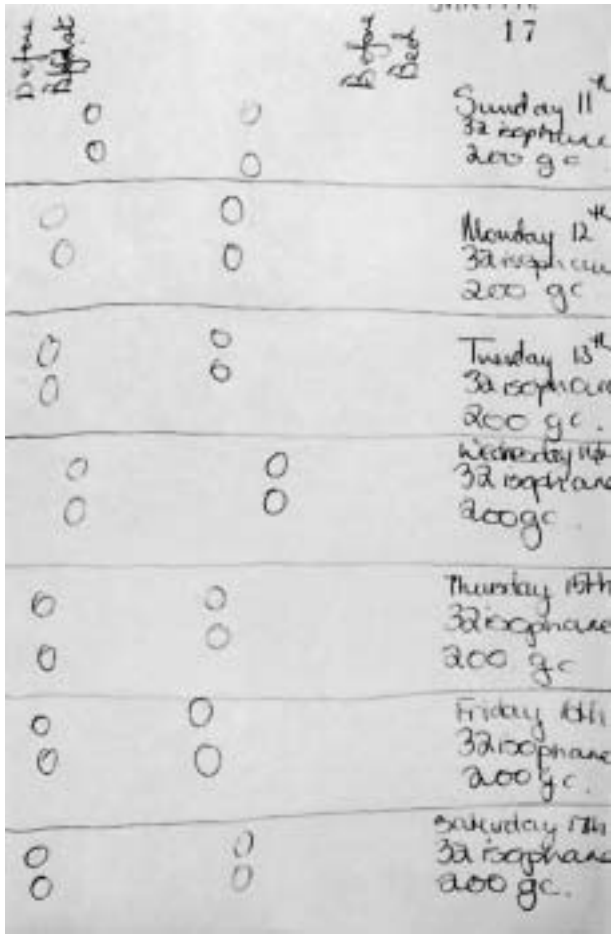
Manipulation should be suspected in patients whose lives are totally disrupted by their diabetes. A careful history may reveal slips which give the vital clue. For instance, one teenager developed profound hypoglycaemia two days after apparently "stopping insulin"; another, whose life was spent in and out of hospital with hypoglycaemia or ketoacidosis, claimed to be perfectly stable in-between, presenting a whole volume of negative urine tests. Even constant insulin infusion does not necessarily solve the problem, especially when the patient replaces the insulin in the syringe with water. When there is strong evidence of manipulation, try hinting at the possibility to the patient or their parents but without accusation. The technique is sometimes successful and gratitude considerable.

Emotional, social, or psychiatric causes underlie disruptive diabetes and the desire to manipulate the situation to cause widespread havoc among families. Teenage defiance is a common cause. Quiet support of families at these difficult times helps to overcome what is almost always a temporary phase. Careful enquiry should establish whether there is family strife, and family counselling is often valuable. Psychiatric advice should only be sought if there is evidence of a psychiatric disorder. It is a huge advantage to work closely with a sympathetic liaison psychiatrist who understands the specific problems of diabetes and is unthreatening to patients; otherwise, confrontation with a psychiatrist may provoke even more aggression. Nonetheless, a few patients remain incapable of independent existence, and then community care strategies are essential.

Various disorders, especially infections and some endocrine disorders, may alter the insulin requirements, although they rarely cause the type of instability already described.


At one clinic of yours which I attended, you asked me if I was taking overdoses. I was stupid and did not admit this until December 1981. I am still not as well balanced as I would like, but I am better than I was.

Letter from a patient



Falsified urine chart from a very unstable teenage patient

8 Hypoglycaemia

Hypoglycaemia is the major hazard of insulin treatment, and problems have increased in the drive to achieve "tight control". Patients may experience the symptoms of hypoglycaemia when the blood concentration is less than 3·0 mmol/l. However, individual susceptibility varies considerably and it is interesting that some patients whose control has been persistently very poor for long periods appear to experience hypoglycaemic symptoms at levels a little above this. The risks of hazard from hypoglycaemia are small in most patients, but because they exist at all, patients taking insulin are barred from certain occupations such as driving trains or buses. All patients taking insulin whose diabetes is reasonably well controlled will experience hypoglycaemia at some stage. At its mildest, it is no more than a slight inconvenience, but at its severest, when unconsciousness can occur, it is both a hazard and an embarrassment. Furthermore, manipulative patients can use hypoglycaemia to threaten family and friends. This sword of Damocles is ever present once insulin treatment has started, and the need to use measures to avoid it requires constant, indeed lifelong, vigilance. Hypoglycaemia occurs infrequently in patients taking oral hypoglycaemics.

Hypoglycaemia ("hypo" "insulin reaction")

This is when the blood sugar goes too low in diabetics taking insulin

Symptoms are sweating, shaking, tingling round the mouth, hazy eyesight or seeing double, slow thinking, in children naughtiness

Causes are late meal, too little carbohydrate, extra exercise, too much insulin.

Cure is to take carboydrate—preferably three dextrosol tablets, glucose, sugar (two large lumps), barley sugar, Lucozade followed by a small snack

Symptoms will soon wear off

If in doubt about an attack, take sugar

Always carry some form of SUGAR with you

Symptoms

Most patients experience the early warning symptoms of hypoglycaemia and can take sugar before more serious symptoms develop. These warning symptoms are well known and are described in the box. Tremulousness and sweating are by far the commonest symptoms, while circumoral paraesthesiae is the most specific. Many patients have highly individual symptoms of hypoglycaemia which range from quite inexplicable sensations to peripheral paraesthesiae.
In three patients carpal tunnel compression resulted in tingling fingers when they were hypoglycaemic, representing their sole warning. Neuroglyopenic symptoms and diminished cognitive function follow if corrective action is not taken, with progressive confusion and eventually unconsciousness and occasionally convulsions. There is a prolonged debate as to whether recurrent hypoglycaemia causes long-term intellectual decline; the evidence in general is unconvincing although major and recurrent episodes in childhood may have an adverse effect in this regard.

Patients who become unconscious from hypoglycaemia need urgent treatment. Brain damage and death do not normally occur because the blood glucose concentration tends to increase spontaneously as the effect of the insulin wears off and the normal counter-regulatory responses become effective. Many diabetics, especially children, need reassurance that they will not die in their sleep. Nevertheless, a very small number of otherwise unexplained deaths at night have been reported in Type 1 diabetic patients (described as the "dead in bed" syndrome) and no precise cause has ever been established. Deaths from prolonged hypoglycaemia are most likely to occur after insulin overdoses, as a result either of a suicide or murder attempt, but even in these circumstances most patients recover.

Symptoms of hypoglycaemia

• **Early warning**	Shaking, trembling Sweating Pins and needles in lips and tongue Hunger Palpitations Headache (occasionally)
Neuroglycopenia	
• **Mild**	Double vision Difficulty in concentrating Slurring of speech
• **More advanced**	Confusion Change of behaviour Truculence Naughtiness in children
• **Unconsciousness**	Restlessness with sweating Epileptic fits, especially in children Hemiplegia, especially in older people (but rare)

Observation of a hypoglycaemic attack

"When she is having a hypo she gives the impression of being drunk. The change in her behaviour is sudden and very noticeable. She slurs her words and appears drowsy. There is lots and lots of yawning. If she is still in the state of which she can walk, she will bump in to things and knock things over and be generally clumsy. She will hardly be aware of where she is or who she is talking to. She rambles"

Diminished awareness of hypoglycaemia

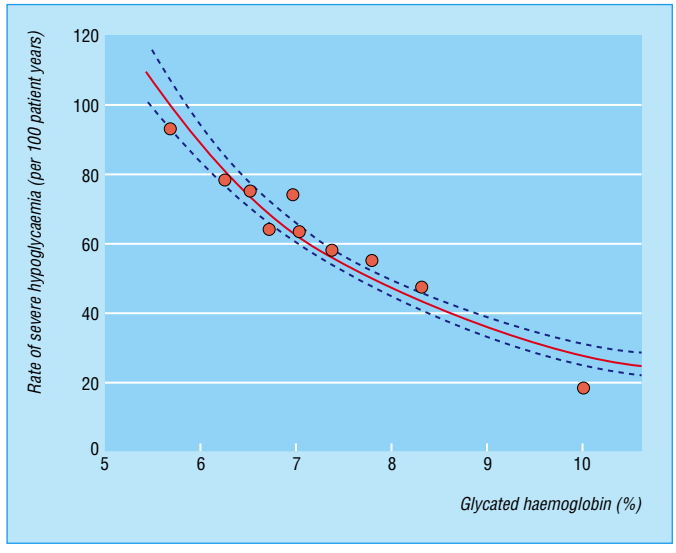
This is the problem which all insulin treated patients dread, and at some stage it affects up to one quarter of Type 1 diabetic patients. It occurs when patients do not experience the early warning symptoms and directly develop diminished cognitive function which prevents them from taking the required preventive action. In this situation, help is required from a third party. This commonly occurs in the home when friends and relations observe the person to be slow-witted with a vacant expression and perspiring face. They may be taciturn, truculent or even obstructive, sometimes refusing to take sugar when advised, although many learn to accept this advice. This state of cognitive impairment can persist for some considerable time, long enough for abnormal behaviour to be noticed during driving, even for several miles; shoppers in the High Street may be unaware that they are shoplifting. If corrective action is not taken, the more serious state of unconsciousness already described can occur.

Night-time hypoglycaemia is very common, usually occurring between 3 and 6 am. The blood glucose concentration often falls below the hypoglycaemic threshold; levels as low as 1·0 mmol/l are not rare, and are known to cause electroencephalogram abnormalities even in the absence of symptoms. Many people become very restless when hypoglycaemic; this is recognised most frequently by the spouse who takes the necessary remedial action. Profound sweating is common, sometimes necessitating a change of nightclothes or bedclothes and may be the only manifestation that hypoglycaemia has occurred. Convulsions are not rare, and some patients wake in the morning with a bitten tongue as the only indication that this may have occured.

Recurrent hypoglycaemia is the principal underlying cause leading to diminished awareness of hypoglycaemia, with dangerous impairment of cognitive function its chief manifestation. It is therefore most likely to occur in those who are most tightly controlled, and was manifest in the famous Diabetes Control and Complications Trial (DCCT) in which severe hypoglycaemia occurred three times more often in the tightly controlled group of patients. This cause outweighs all others, although the use of β adrenergic blockers has the same effect in a small number of patients. Autonomic neuropathy is not normally the cause of diminished warning, and the contentious role of human insulin in this regard has led some patients to change back to animal insulins, though scientific evidence of harm is still lacking. Diminished warning also increases with lengthening duration of diabetes.

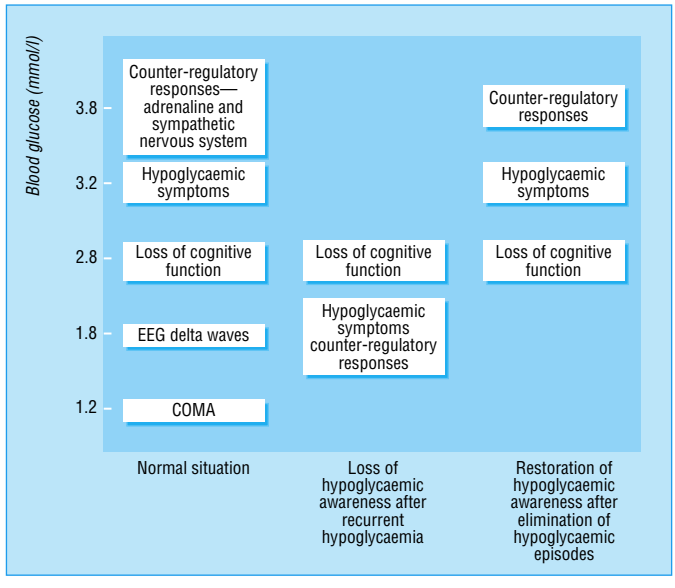
Why does recurrent hypoglycaemia beget loss of warning? The answer probably lies in the readjustment of the threshold of sensitivity of a glucose sensor in the hypothalamic region of the central nervous system. This alters the hierarchy of responses to hypoglycaemia. Thus, the hypoglycaemic symptoms and counter regulatory responses, instead of occurring at a blood glucose just above 3·0 mmol/l, develop at a lower level, rather less than 2·0 mmol/l. Because in either case the loss of cognitive function occurs around 2·8 mmol/l it is clear that, in cases of diminished hypoglycaemic awareness reduced cognitive function develops before hypoglycaemic warning symptoms.

Recent research has also shown that by eliminating recurrent hypoglycaemia it is possible to restore the normal sequence of events when blood glucose falls, thus also restoring adequate warning. It is therefore necessary to eliminate as far as possible all hypoglycaemic episodes, even those occurring



Relationship between glycated haemoglobin and hypoglycaemia rates shown in the DCCT study

Avoiding hypoglycaemia can restore proper warning symptoms. Patients should try to avoid blood glucose levels < 4·0 mmol.



Loss of warning of hypoglycaemia (EEG=electroencephalogram)

quietly at night. This can often be achieved simply by reducing the insulin dose, ensuring adequate carbohydrate intake and to some extent relaxing overall diabetic control in the interest of safety. It is much more difficult, yet possible, to eliminate hypoglycaemia and retain optimal control of diabetes. The acquired skill of the diabetes team and the co-operation of patients is needed if this is to be done, and the time and resources needed are considerable. The introduction of programmes such as blood glucose awareness training (BGAT), in which patients are taught how to recognise the most subtle symptoms of early hypoglycaemia, can be very effective in reducing serious hypoglycaemia and thus helping to restore adequate warning.

A questionnaire which helps physicians to assess whether a patient has diminished awareness of hypoglycaemia is shown in Appendix 1.

Causes of hypoglycaemia

In every patient taking insulin the blood glucose concentration shows peaks and troughs, which can be most clearly shown by home measurements of blood glucose. Since the lowest blood glucose concentrations occur at different times in each patient, it is a great advantage if individual patients know when their own troughs are likely to occur. The commonest times are before lunch and during the night. Some patients in their constant fear of developing diabetic complications drive their blood glucose levels ever lower with disastrous consequences in terms of hypoglycaemia.

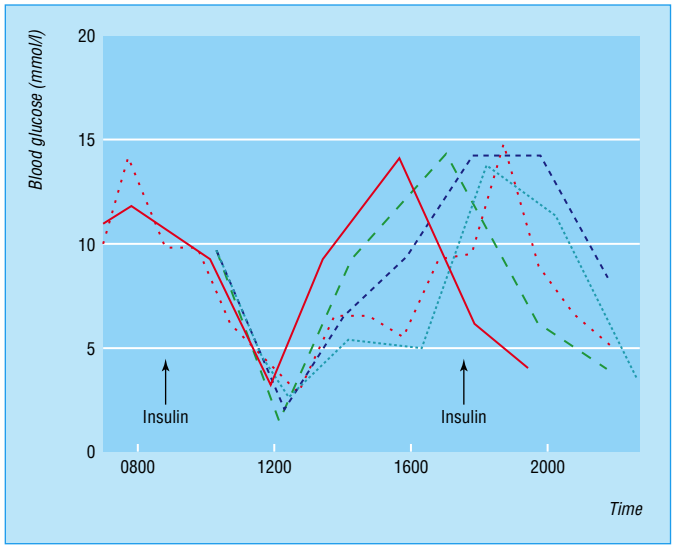
Severe physical activity, such as swimming very long distances, is a powerful stimulus of hypoglycaemia, and as much as 40 to 50 g additional carbohydrate may be needed to prevent it. Hypoglycaemia in these situations is sometimes delayed for several hours. Several well-known sportsmen and women with diabetes show considerable ingenuity and perseverance in the way in which they cope with their diabetes during international competitions, by individual attention to food and insulin intake, carefully timed blood glucose monitoring, and ready availability of sugary fluids such as Lucozade at exactly the right moment.

Hypoglycaemia is particularly likely to occur shortly after stabilisation of new patients, as their insulin requirements may decline considerably; their insulin dose should therefore always be reduced before they leave hospital.

Hypoglycaemia is also troublesome when insulin requirements insidiously decrease during the evolution of such conditions as Addison's disease, hypopituitarism, and malabsorption syndromes.

Treatment and prevention of hypoglycaemia

Much of the skill required to manage insulin treated diabetic patients is therefore devoted to achieving adequate control of diabetes, yet avoiding hypoglycaemia. There are quite straightforward measures which many patients neglect: they must therefore at all times carry a supply of glucose both on their person and in their cars, and take 10 to 20 g at the first warning symptoms, preferably followed by a carbohydrate snack. The late RD Lawrence always demanded that his patients should demonstrate that they were carrying their sugar supply with them. This can take the form of sugar lumps, sweets (non-diabetic), sugar gel or dextrose tablets.



Reproducible blood glucose profiles in one patient during five consecutive days, showing the times of blood glucose "troughs" when hypoglycaemia is most likely to occur

Events likely to provoke hypoglycaemic attacks

- Insufficient carbohydrate in meals
- Delayed meals
- Increased physical activity
- Errors of insulin dosage
- Erratic insulin absorption from areas of fat hypertrophy at injection sites

Items containing 10 g of carbohydrate

• Milk	200 ml (⅓ pint)
• Lucozade	60 ml (4 tablespoons)
• Ribena	15 ml (1 tablespoon)
• Coca Cola	90 ml
• Sugar	2 teaspoons
• Sugar lumps (small)	3
• Dextrosol tablets	3

They should take ample carbohydrate at times when blood glucose troughs occur, notably mid-morning and bedtime, and they must take appropriate amounts of additional carbohydrate before and during vigorous exercise. Careful blood glucose monitoring plays a crucial part in avoiding hypoglycaemic episodes, and helps to restore warning of hypoglycaemia. Patients should try to avoid blood glucose levels below 4·0 mmol/l. Appropriate insulin regimens that need to be devised for individual patients are described in chapters 5 and 6.

Glucagon

Glucagon is a hormone produced by the A-cells of the pancreatic islets. It raises the blood glucose by mobilising the glycogen stores in the liver (and therefore will not work after prolonged starvation). It is given in a 1 mg dose by injection most conveniently intramuscularly. It can also be used subcutaneously or intravenously and is effective in five to 10 minutes. It is of great value for bystanders of severely hypoglycaemic patients who are unable to take oral glucose, and can be injected by family members, nurses or doctors. It is valuable in relieving stress in a home where a diabetic patient, often a child, is prone to recurrent disabling attacks of hypoglycaemia.

Unconsciousness

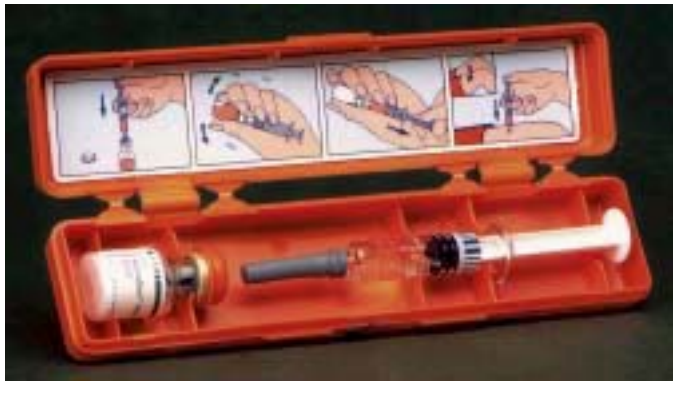
In cases where the patient has lapsed into severe unconsciousness, treatment in hospital is urgently needed. The unconscious patient should be placed in the recovery position, and the airway maintained. Blood should be taken for blood glucose analysis and the sample should be kept in case the patient fails to respond to treatment since the possibility always exists that the coma has another cause. Intravenous glucose is given using 50 ml of 20% glucose solution. The more concentrated 50% solution is highly irritant and should no longer be used. The response is usually immediate but if not, a further dose should be given after five to 10 minutes followed by an infusion of 10% glucose. Once consciousness is restored and a history can be taken, the patient should be fed with longer acting carbohydrate to prevent recurrence. If recovery does not occur rapidly, blood glucose measurement should be repeated and another cause for the coma must be sought. If hypoglycaemia has been profound, cerebral oedema can occur and may require treatment with dexamethasone or mannitol. After recovery appropriate adjustment must be made to the diabetic treatment in order to avoid further episodes, and the patient should be carefully reviewed in the diabetic clinic.

Hypoglycaemia due to oral hypoglycaemics

This may occur during treatment with sulphonylureas and similar agents (but not with metformin) especially in some confused elderly patients who either inadvertently take additional tablets or omit their meals. It is treated in the same way as described above. These patients usually require admission to hospital for continuous glucose infusion to avoid relapse into hypoglycaemia, which often occurs until the drug has been cleared from the circulation.

Dr Charles Fletcher's account of hypoglycaemia

My main problem has always been hypoglycaemia. At first I was nearly always aware of it by day and woke at night, because of the adrenaline response. But, particularly in the past 20 years, it gradually became more difficult. I may now feel normal and do ordinary tasks quite easily with blood sugar as low as 2·5 mmol/l (45 mg/100 ml). Sometimes diplopia, dysphasia, weariness, or inability to think may lead me to do a blood sugar. But I often become too muddled to know what is wrong, and I have had to thank my wife, my children, and many generations of housemen, registrars, and secretaries for spotting these low levels on many occasions. Before I retired 50% glucose was always available with syringe in a drawer in my desk. I became quite used to a quiet registrar's voice in outpatients (and elsewhere) saying, "I think, sir, a little extravenous glucose might help". Lucozade has been invaluable. I always have it available in the car, in the office, and at home. It is acceptably free from sugariness, it saves me chewing and choking on dry glucose tablets, and it is rapidly absorbed. My wife finds it much easier to get me to drink this than to take any other form of sugar when I am severely hypoglycaemic and refuse to acknowledge it. I have made it a rule, which I now keep, even when semi-comatose, that if my wife—or anyone else—tells me to take sugar I do so however sure I may be that I'm not hypoglycaemic. They have only been wrong on rare occasions. I am very sensitive to exercise, but for some reason I find it difficult always to suck prophylactic sweets on country walks or when digging or mowing in the garden.



Glucagon injection kit

Conclusions

Any serious hypoglycaemic episodes can to some extent be regarded as a failure of the doctor, the patient or the treatment regimen itself. It should provoke a serious inquiry to establish the cause and to discover if it is likely to recur. The opportunity for the necessary education should be taken, and patients should be encouraged to carry a diabetic identification card. Finally, the professional attending the patient, whether doctor or nurse, has a duty to inform people who have had an episode of severe hypoglycaemia comprising diminished cognitive function resulting from diminished awareness of hypoglycaemia to stop driving and inform the Driver and Vehicle Licensing Agency. They should also avoid any other potentially dangerous activity. The endeavour to avoid hypoglycaemia needs to be maintained, and patients need considerable support to this end at almost every diabetic consultation throughout life.

The figure showing the relationship between glycated haemoglobin and hypoglycaemia rates is adapted from Donnelly R, et al *BMJ* 2000;320:1062-6. The photograph of the glucacon injection kit is with permission from Novo Nordisk. Dr Charles Fletcher's account of hypoglycaemia is from *BMJ* 1980;280:1115-16.

The story of Mrs B-J continued: hypoglycaemia

When I was 12 I was sent to The Old Palace School in Croydon. Never having done any gymnastics, I was surprised at what was expected of me at the new school. Rope climbing, parallel bars, and marching up and down the long hall where kings had been entertained by archbishops, I was soon very hypo and staggering about like a drunk. This disrupted the session so much that finally I was banned from gym and all other sports. My fellow diabetic, Barbara, had also been banned, but it did not worry me too much. But it was a bitter blow when I was told I would not be allowed to take part in the school pageant "in case I was ill"! My mother went up to see Sister and assured her I would have plenty of sugar etc., but to no avail. Gentle nuns can be very obstinate and quite hardhearted at times.

9 Diabetic ketoacidosis and management of diabetes during surgery

Ketoacidosis

Ketoacidosis results from a lack of insulin. In practice it is usually due to:

- stopping insulin or reducing the dose either in error or deliberately
- resistance to insulin during infections or other intercurrent illness
- the unrecognised onset of Type 1 diabetes.

The clinical onset of ketoacidosis occurs over hours or days. Symptoms of uncontrolled diabetes are always present. Vomiting in Type 1 diabetic patients is always serious. Patients usually consult their doctors during the preceding days, but the presence of uncontrolled diabetes is frequently overlooked. Diabetic control should always be assessed if a diabetic patient becomes unwell for any reason. Many cases of ketoacidosis could be prevented.

Preventing ketoacidosis: sick day rules

During any illness or infection the blood glucose concentration tends to increase and diabetic control deteriorates. Most patients then need a larger dose of insulin than usual, and some who normally take tablets may need insulin just during the illness. The increased need for insulin occurs even when the appetite declines or vomiting begins.

Every insulin treated patient should understand that insulin should never be stopped. Stopping or even reducing insulin during the course of an illness often leads to diabetic ketoacidosis.

When a diabetic person is ill the normal insulin dose should be continued, carbohydrate taken in some palatable fluid form, and the blood tested regularly—four times a day if necessary. If blood glucose readings greater than 15 mmol/l are obtained the dose of insulin should be increased. Additional doses of insulin (about 8 units) may also be given at noon or bedtime when control is very poor. It is preferable to make these adjustments with short acting (soluble) insulin if this is available. If vomiting continues without remission for more than a few hours, admission to hospital for treatment with intravenous fluids and insulin is advisable to prevent ketoacidosis.

Assessment of blood or urine ketones during illness is helpful. Using the new blood ketone meters, readings of 1·0-3·0 mmol/l taken in conjunction with the blood glucose reading usually indicate the need for additional insulin; readings should be repeated within two to four hours. If they persist or increase above 3·0 mmol/l, specialist advice is required from the hospital clinic staff. Ketonuria can be detected using Ketostix, which are readily available.

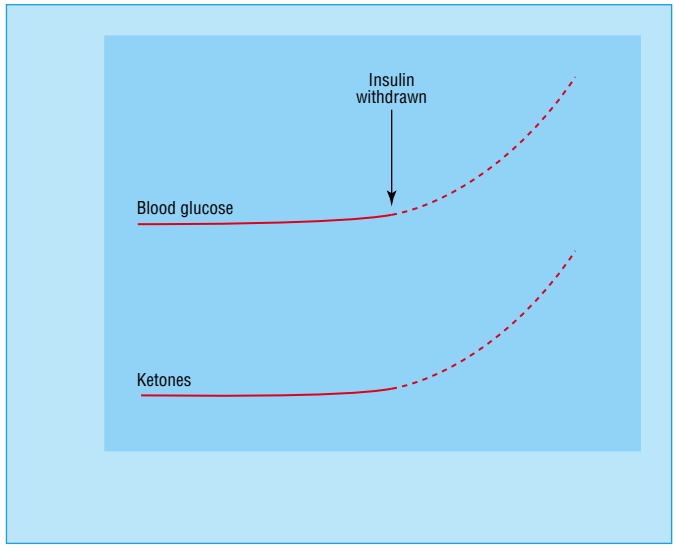
Recognising ketoacidosis

Dehydration is the most obvious clinical feature of patients with ketoacidosis. They are also drowsy, but rarely unconscious—"diabetic coma" is an inappropriate description; they are often overbreathing, but not usually breathless; their breath smells of acetone (though many people cannot smell this); and many also have the gastric splash. In more severe cases patients are

Causes of diabetic ketoacidosis

• Omission or reduction of insulin dose	27%
• Previously unknown diabetes	22%
• Infection	17%
• Miscellaneous	16%
• No cause found	18%

King's College Hospital 1968



Ketoacidosis

Prevention of ketoacidosis

Insulin should never be stopped

Insulin dose during illness or infection

During illness or infection your blood glucose level may rise, causing you to feel dry, thirsty and pass too much urine. The blood glucose is also likely to increase

You MUST continue to take your normal insulin dose NEVER stop it. You may need an increased dose if your blood tests are bad. If you are vomiting, consult your doctor or the diabetic clinic at once. If you are unable to eat, take your carbohydrate portions in liquid form—for example, milk, Lucozade, Ribena

Test your blood twice a day or even more frequently

If you continue to feel unwell, consult your doctor.

Features of ketoacidosis

- Drowsiness
- Dehydration
- Overbreathing
- Acetone on breath
- Hypotension
- Gastric splash

hypothermic (even in the presence of infection) and hypotensive. Hyperosmolar non-ketotic (HONK) patients are usually grossly dehydrated but without overbreathing or the smell of acetone. Inexperienced clinicians often have difficulty in recognising patients with this condition, especially when they seem deceptively well.

Tests for ketoacidosis

- Blood glucose
- Serum potassium and sodium
- Acid-base status
- Urea, creatinine
- Plasma or urine ketones
- Blood count
- Blood culture (when indicated)

Diagnosis

The diagnosis of ketoacidosis is confirmed by laboratory tests.

- **Blood glucose** concentrations may range from slightly increased to extreme hyperglycaemia. The blood glucose concentration itself does not usually indicate the severity of the illness, although most patients are seriously unwell when it is greater than 30 mmol/l.
- **Blood acid-base status** pH ranges from 6·9 to normal. The bicarbonate level is depressed.
- **Plasma ketones** are easily detectable with a ketone meter and exceed 3·0 mmol/l.
- **Urine test** shows heavy glycosuria and ketonuria.
- **Electrolytes**: the serum potassium concentration is either normal or raised, and very rarely low. This measurement is vital, and life-saving treatment is needed to maintain potassium values in the normal range. The sodium concentration is normal or reduced, and urea and creatinine concentrations are often raised through dehydration.
- **Blood count**: if a blood count is performed the white cell count is often spuriously raised to 15-20 $\times 10^9$/l even in the absence of infection.

  Serum amylase is sometimes moderately elevated in patients with diabetic ketoacidosis: it is of salivary origin and need not be indicative of pancreatitis

Blood glucose and ketone meter

Treatment

Patients should be treated in an area where there they can be observed regularly, preferably by staff familiar with managing this condition, in a high dependency area, or if very ill in intensive care.

- Insert a nasogastric tube if consciousness is impaired. Do not allow any fluids by mouth; if patients are thirsty they may suck ice.
- Give intravenous fluids. The regimen needs to be modified according to age, weight, and the presence of cardiac disease. In seriously ill patients and all those with cardiac disease a catheter for measuring central venous pressure is essential. A suitable regimen for most patients is shown in the box; 0·9% saline is used.
- The fluid should be changed to 10% dextrose once the blood glucose concentration has fallen to less than 10 mmol/l. The rate of infusion is determined by individual need but at this stage should probably be about one litre every eight hours.
- Start intravenous soluble insulin immediately. If there is any delay in obtaining intravenous access. Soluble insulin (20 units) can be given immediately intramuscularly.

Treatment of ketoacidosis

Physiological saline:*	1 l in first half hour	-½ h
	1 l over next hour	½-1½ h
	1 l over next hour	1½-2½ h
	1 l over next 2 hours	2½-4½ h
	1 l over next 3 hours	4½-7½ h
	1 l over next 4 hours	7½-11½ h
Total:	6 l	11½ h

*Change to 10% dextrose when blood glucose is less than 10 mmol/l

Insulin treatment

Intravenous insulin: soluble insulin is diluted in 0·9% saline in a syringe, at the concentration of 1 unit/ml. It is given by infusion pump at 6 units/h (0·1 units/kg/h for children) until the blood glucose concentration is less than 10 mmol/l. Blood glucose should fall at a rate of about 5·0 mmol/l/h, and plasma ketones should fall at the same time. When the blood glucose is less than 10 mmol/l, the dose may be reduced to 3 units/h. Higher infusion rates are rarely needed; when they are needed

in insulin resistant patients the rate should be doubled or quadrupled, etc. If the patient is not responding, medical staff should check the equipment for pump failure, blockage or leakage. The insulin infusion is continued until the patient is well enough to eat. The changeover to subcutaneous insulin should be made before breakfast. Preprandial subcutaneous soluble insulin is then given and intravenous insulin discontinued after the meal. Intravenous insulin should not be stopped before subcutaneous insulin has been given (see below).

Intramuscular insulin is used only when an infusion pump is not available. Soluble insulin 20 units is given as a loading dose, than 6 units every hour until blood glucose is less than 10 mmol/l, then continued at two hourly intervals. As with intravenous insulin, higher doses are rarely needed.

Potassium and sodium bicarbonate

Potassium chloride administration should usually start at about the second hour, preferably not before the serum potassium concentration is known. It should be withheld in exceptional cases of oliguria or anuria, or if the serum potassium value remains above 5·0 mmol/l. After the second hour, or earlier if the initial serum potassium value is normal or less than 4·0 mmol/l, 20 mmol potassium chloride should be added to each litre of saline. If the serum potassium value falls below 3·5 mmol/l, 40 mmol should be used in each litre. The exact amount should be determined by serial serum potassium measurements—every two hours at first, then every four hours—and serum potassium maintained between 4·0 and 5·0 mmol/l. An electrocardiographic monitor should be set up; however, there is no substitute for serial potassium measurements.

Sodium bicarbonate is not normally beneficial and is not given unless the blood pH value is less than 7·0 or the patient is shocked. If it is needed, aliquots of sodium bicarbonate (500 ml of 1·26%) with added potassium chloride (15 mmol) should be given. This can be repeated if there is no response within one hour and if the patient's condition remains serious.
Note: Never use sodium bicarbonate 8·4% concentration.

Treatment of the underlying condition

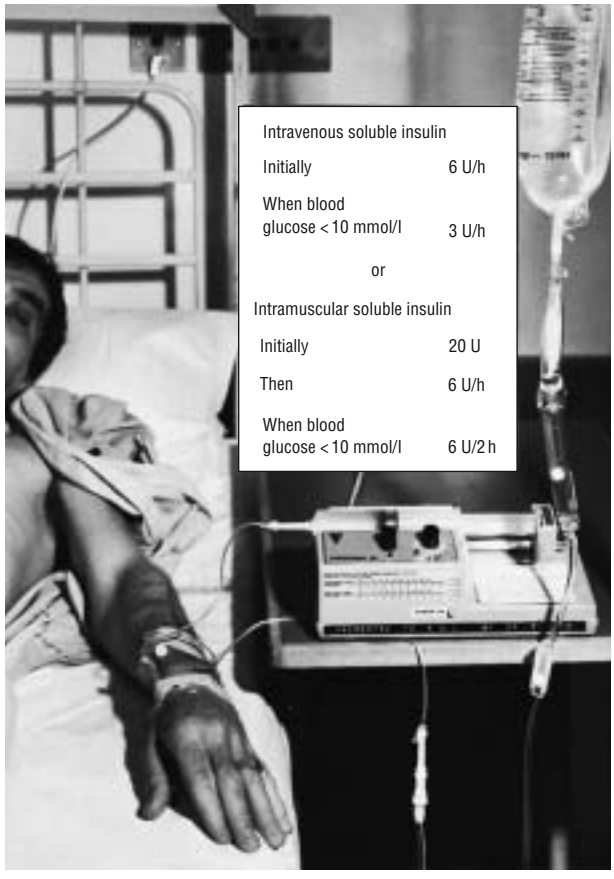
Underlying disease should be sought, especially respiratory or urinary infections, which may not be obvious at the onset. Blood culture, culture of the midstream specimen of urine, and chest radiography are performed. There is no need to give antibiotics routinely. Abdominal pain may occur, especially in young patients with severe ketoacidosis. It is vital to discover whether there is indeed an intra-abdominal cause needing attention if the pain does not resolve rapidly.

HONK patients

Blood glucose in these patients can be extremely high without ketosis or acidosis. Management is the same as that for ketoacidosis, except that 0·45% saline is given if the serum sodium value is greater than 150 mmol/l, and a lower rate of insulin infusion (3 units/h) is often sufficient. In shocked and dehydrated patients prophylactic, low dose, subcutaneous heparin is considered.

Lactic acidosis

These patients are profoundly ill and the cause of the acidosis must be sought and rigorously treated. They are often very insulin resistant due to serious intercurrent illness, and need large amounts of sodium bicarbonate. Absence of a raised plasma ketone level excludes ketoacidosis as the cause of the metabolic acidosis. Metformin induced lactic acidosis should be borne in mind.



Giving intravenous insulin

Potassium chloride administration

Serum value (mmol/l)	< 3·5	3·5-4·0	4·0-5·0	> 5·0
Administer	40 mmol/l	30 mmol/l	20 mmol/l	0

Patients who develop HONK are often elderly or West Indian, and they often turn out to have Type 2 diabetes

Management of insulin treated diabetes during surgery

The chief principle of diabetic management through any crisis in which patients cannot eat or drink for any reason is to continue insulin administration. The best method is to give the insulin by continuous intravenous infusion either by infusion pump or directly from the drip bag.

For operations in which a patient is likely to be maintained on a drip for more than 12 hours a regimen is needed which can be continued for an indefinite period. Again there are two methods of administering the insulin: a variable rate infusion using a pump, or if this is not available, a glucose insulin-infusion.

Note:

- The rate of intravenous infusion must depend on the clinical state of the patient with regard to the volume depletion, cardiac failure, age, etc.
- Potassium replacement is required.
- If the blood glucose is persistently above 10 mmol/l the infusion should be changed to 0·9% saline.
- Blood glucose should be monitored every one to two hours during surgery and regularly postoperatively.
- Try to maintain the blood glucose concentration in a safe range—6·0-12 mmol/l.
- Regular (at least daily) electrolyte measurements are required.

After recovery: changing to subcutaneous insulin

Once the patient starts to eat and drink conversion back to subcutaneous insulin injections is undertaken as follows.

- Always change to subcutaneous insulin before breakfast and never in the evening so that adequate supervision can be assured.
- Stop the insulin pump 30 minutes after the first subcutaneous insulin injection.
- Insulin regimen and dose: if the previous regimen is known then this should be given; if the patient is still in bed or unwell the total dose may need to be 10 to 20% more than usual. If the patient was not previously taking insulin, predicting the requirement is not easy and the amount needs adjustment from day to day. Initially use insulin 30-40 units daily in divided doses given four times daily.

Patients with hyperglycaemia often relapse after conversion back to subcutaneous insulin. When this happens there are three possible approaches.

- Give additional doses of soluble insulin at any of the four injection times (before meals or bedtime).
- Add an intravenous insulin infusion temporarily while continuing the subcutaneous regimen until the blood glucose concentration is satisfactory.
- Revert completely to the intravenous regimen, especially if the patient is unwell.

Surgery in Type 2 diabetes

Management of diabetic patients treated with diet or oral hypoglycaemic agents is more straightforward, so long as the diabetes is well controlled.

If the random blood glucose value is less than 12 mmol/l:

- omit the tablet on the day of surgery
- check the blood glucose concentration before and soon after the operation; if the blood glucose value is more than 12 mmol/l start soluble insulin.

If the diabetes is poorly controlled (random blood glucose greater than 12 mmol/l) the patient should be started on

Dextrose drip and variable rate insulin infusion

(1) Give normal insulin on the night before the operation
(2) Early on the day of operation start an infusion of 10% dextrose, add 20 mmol potassium chloride to each litre, and run at a *constant* rate appropriate to the patient's fluid requirement, usually 100 ml/h
(3) Make up a solution of soluble insulin 1 unit/ml saline in a syringe and infuse intravenously by a line piggybacked to the intravenous drip by using a syringe pump. The infusion rate should normally be as shown in regimen 1, but in resistant cases use regimen 2 or 3

Blood glucose	Soluble insulin infusion rate		
	Regimen 1	Regimen 2	Regimen 3
< 4 mmol/l	0·5 unit/h	1 unit/h	2 unit/h
4-10 mmol/l	2 unit/h	4 unit/h	8 unit/h
10-15 mmol/l	4 unit/h	8 unit/h	16 unit/h
15-20 mmol/l	6 unit/h	12 unit/h	24 unit/h
> 20 mmol/l		Review	

Blood glucose is measured preoperatively and then two hourly until stable, then six hourly

Regimen 1 is satisfactory for most cases; very severely ill patients, shocked patients, and those receiving steroids, salbutamol, or dopexamine infusions may need higher dose infusions, such as regimens 2 or 3, occasionally even more.

Do not stop the insulin infusion since intravenous insulin lasts for only a few minutes

Only if the patient becomes frankly hypoglycaemic (blood glucose < 2·0 mmol/l) should insulin be stopped for up to 30 minutes

Glucose-insulin infusion

(1) Give normal insulin on the night before the operation
(2) Begin an infusion of 10% dextrose containing 20 mmol/l potassium chloride and soluble insulin 15 units/l. Run it at a rate appropriate to the patient's fluid requirements, usually 100 ml/h. Adjust insulin dose as follows

Blood glucose	Soluble insulin infusion
< 4 mmol/l	15 unit/l
4-10 mmol/l	30 unit/l
10-15 mmol/l	40 unit/l
15-20 mmol/l	60 unit/l
> 20 mmol/l	Review

Blood glucose is measured two hourly until stable, then six hourly

Surgery in Type 2 diabetes

- Omit usual treatment
- Use insulin if diabetic control deteriorates
- Maintain blood glucose chart

insulin before the operation, using one of the regimens described on the previous page.

Management of insulin treated diabetes during day surgery
Patients with insulin treated diabetes requiring an anaesthetic for relatively minor operations or investigative procedures (for example, barium radiological examinations, cystoscopy, endoscopy, etc.) can be treated as day cases without hospital admission provided that:

- the procedure is undertaken in the morning first on the list (if the procedure is performed first on an afternoon list, a light breakfast is taken after half the normal insulin dose, followed by regular blood glucose monitoring)
- the procedure does not exceed approximately one hour in duration
- the patient will be able to eat and drink within one hour of the procedure
- the patient is able to self-monitor blood glucose and adjust insulin appropriately.

The blood glucose should be rechecked before discharge. If significant problems with diabetes control persist, then hospital admission may be required after all.

Instructions for management of insulin treated diabetes in day surgery

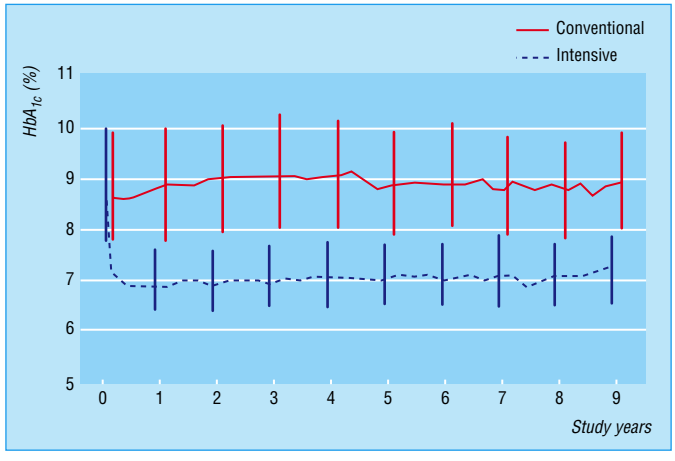
- The procedure should be performed after an overnight fast
- Insulin is taken on the previous evening as usual
- Insulin and breakfast are omitted on the morning of the procedure
- The blood glucose must be measured before leaving home and again before the procedure is commenced.
- If the blood glucose is less than 6·0 mmol/l a 10% dextrose drip is needed
- Within one hour after completion of the procedure, the normal morning insulin dose should be administered followed by appropriate food and fluid giving the equivalent amount of carbohydrate to the usual breakfast

10 Diabetic complications: cause and prevention

Introduction

Patients with long-standing diabetes may develop complications affecting the eyes, kidneys or nerves (microvascular complications) or major arteries. The major arteries are affected in people with diabetes, causing a substantial increase in both in coronary artery disease and strokes as well as peripheral vascular disease. The greatest risk of large vessel disease occurs in those diabetic patients who develop proteinuria or microalbuminuria, which are associated with widespread vascular damage. These complications are often discovered at presentation in Type 2 diabetic patients who must have had diabetes for many years before it has been diagnosed. Issues concerning macrovascular complications are described in chapter 17.

During the last two decades, there has been a considerable increase in understanding the mechanisms underlying the development of the long-term diabetic microvascular complications (retinopathy, nephropathy, and neuropathy) and macrovascular disease, accompanied by major developments in preventing them. The United Kingdom Prospective Diabetes Survey (UKPDS) in particular demonstrated quantitatively the long-term harmful effects of hyperglycaemia and hypertension in the development of both microvascular and macrovascular complications in Type 2 diabetes. Both UKPDS and the Diabetes Complications and Control Trial (DCCT) of Type 1 diabetes demonstrated the benefits of optimal control.



HbA_{1c} values during 9 years in the DCCT study of Type 1 diabetes, showing steady levels in intensively and conventionally controlled groups

Causes and prevention of complications

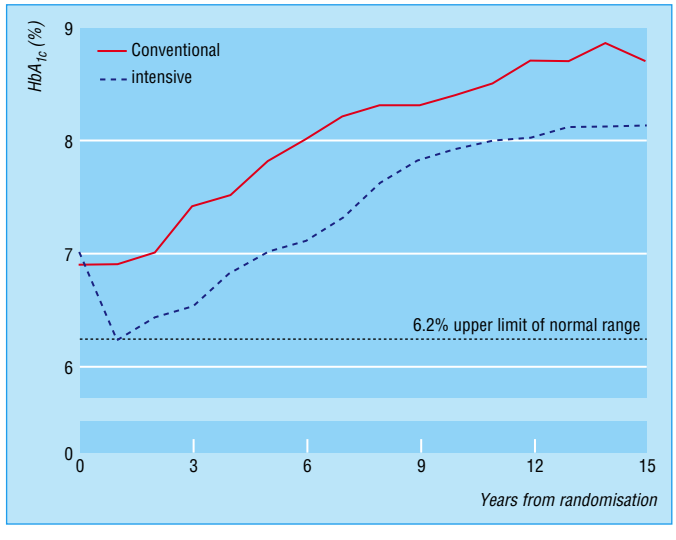
Major advances in recent years have resulted in an actual decrease of some complications, notably nephropathy. Primary prevention of diabetic complications, together with retardation of their progression, is now possible, chiefly by tight control of the diabetes and of hypertension, together with reduction of other "risk factors" detailed in chapter 17. Even when the complications are established, their progression leading to serious damage can be delayed.

Although many attempts have been made to develop specific pharmacological agents to alter the course of diabetic complications, and although many trials are in progress at the present time, none have proved unequivocally successful and none are licensed. There is at present intense interest in and optimism for the use of protein kinase-C inhibitors.

Two major studies

DCCT: a multicentre study of 1441 Type 1 diabetic patients in the United States examining the effects of tight control on the development of microvascular complications, terminated after nine years because of highly significant benefits reported in 1993. The benefits on the microvascular complications were considerable.

UKPDS: a multicentre study of 5102 Type 2 diabetic patients co-ordinated from Oxford, assessed both the harmful effects of persistent hyperglycaemia and hypertension on the development of microvascular and macrovascular complications, and also demonstrated the



HbA_{1c} values during 15 years of the UKPDS study of Type 2 diabetes showing progressive deterioration of both intensively and conventionally controlled groups

benefits of 10 years of better, compared with less satisfactory, control of both glycaemia and blood pressure reported in 1998. Benefits were achieved regardless of the drugs used to reach the required standards of either blood glucose or blood pressure control.

The long-term effects of treatment in the two studies are shown in the two figures demonstrating the stable control in Type 1 diabetes (DCCT) compared with the deteriorating control in Type 2 diabetes as the disease progresses (UKPDS).

Persistent hyperglycaemia

Over many years this is the principal underlying cause of the microvascular complications of diabetes. It is also an independent risk factor for the development of macrovascular coronary artery disease and cataract formation. The UKPDS showed precisely the increasing hazard in relation to continuously rising HbA_{1c} levels, without any specific threshold point, and then demonstrated the benefits of tight control. Once complications are established additional factors, notably hypertension, may accelerate their progression (for further details see chapters on specific complications).

For every 1% increase in HbA_{1c}:

- microvascular complications increased by 37%
- any end point (micro and macrovascular) related to diabetes increased by 21%
- deaths related to diabetes increased by 21%.

(Microvascular complications are here defined as retinopathy requiring photocoagulation, vitreous haemorrhage, and fatal or non-fatal renal failure.)

The progression of neuropathy assessed in a group of Type 1 diabetic patients in a prospective 14-year study conducted in Dusseldorf has also shown clearly that the decline of numerous measurements of nerve function occurs almost exclusively in those with poor glycaemic control.

The effect of better blood glucose control on the microvascular complications was as follows:

- reduction of microvascular complications (chiefly the need for photocoagulation) by 25%
- reduction of any diabetes end point by 12%
- reduction of any diabetes related death by 10%.

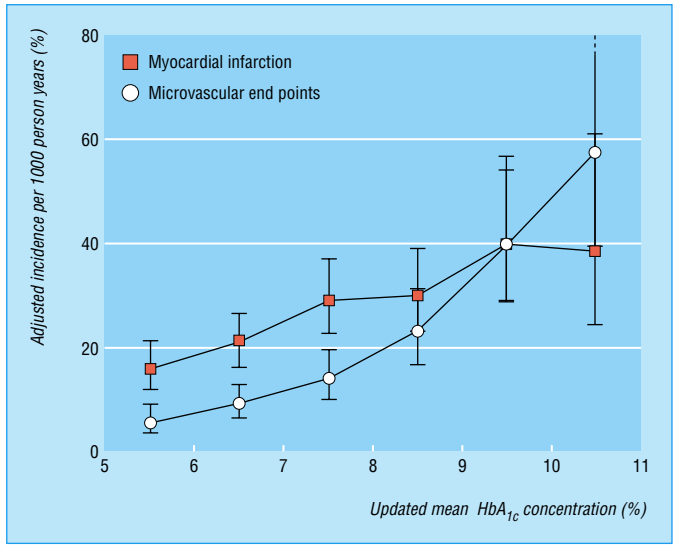
Glycaemic control was also shown to reduce the evolution of microalbuminuria after nine years, and the loss of vibration perception after 15 years of the study. Tight blood glucose control had a non-significant effect on reduction of myocardial infarction, and none on diabetes related mortality.

The DCCT (Type 1 diabetes) demonstrated that primary prevention and retardation of progression of diabetic complications can be achieved over a decade if tight diabetic control is achieved. Retinopathy, nephropathy, and neuropathy were reduced by 35-70% if HbA_{1c} was maintained around 7%. Maintaining tight control requires optimisation of insulin regimen and diet (see chapters 5 and 6), careful blood glucose monitoring, and substantial professional support.

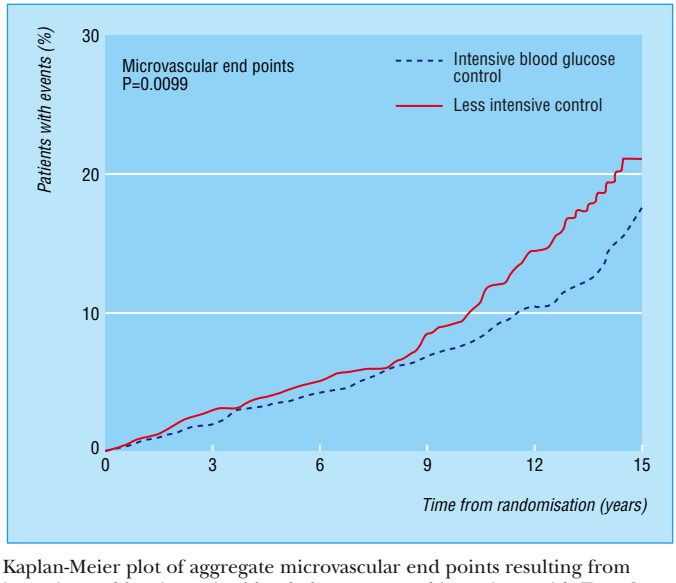
Five years after termination of the DCCT, the EPIC study showed that, despite lapse of the earlier tight blood glucose control, the benefits with regard to amelioration of complications persisted.

Hypertension

This is the principal underlying risk factor for the development of coronary artery disease leading to myocardial infarction, and increases the risk of strokes and heart failure as well.



Incidence rates and 95% confidence intervals for myocardial infarction and microvascular complications by category of updated mean HbA_{1c} values, adjusted for age, sex, and ethnic group, expressed for white men aged 50-54 years at diagnosis and with mean duration of Type 2 diabetes of 10 years (UKPDS)



Kaplan-Meier plot of aggregate microvascular end points resulting from intensive and less intensive blood glucose control in patients with Type 2 diabetes in UKPDS. Microvascular disease here includes renal failure, death from renal failure, retinal photocoagulation, or vitreous haemorrhage

It also exacerbates the progression of retinopathy, the evolution of proteinuria, and probably the deterioration of nerve function as well.

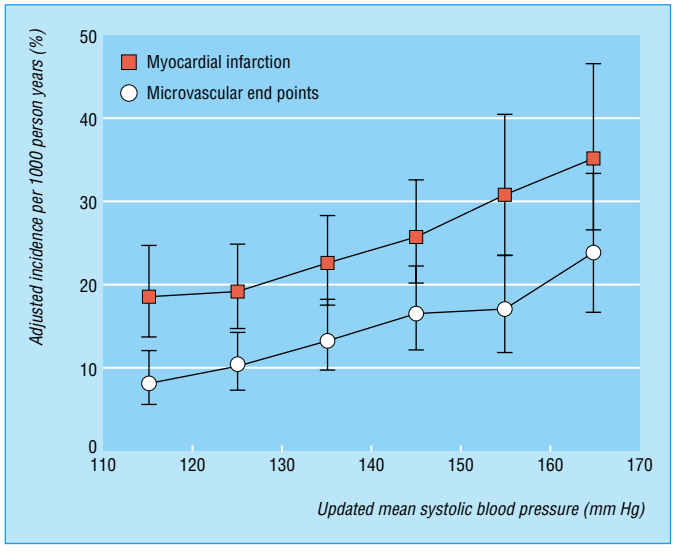
The UKPDS (Type 2 diabetes) has shown that for every 10 mm Hg increase in systolic blood pressure:

- any complication related to diabetes is increased by 12%
- deaths related to diabetes are increased by 15%
- myocardial infarction is increased by 11%
- microvascular complications are increased by 13%.

By achieving a mean blood pressure of 144/82, representing a reduction of systolic blood pressure of 10 mm Hg compared with the less intensively treated group, microvascular end points (chiefly the need for photocoagulation) were reduced by 37%, and risk of vision declining by three lines on the Snellen chart was reduced by 47%, chiefly by protection from the development of macular disease.

Better control of blood pressure also resulted in a 32% reduction in deaths related to diabetes, and a 44% reduction in strokes; there was a non-significant reduction in myocardial infarction.

Further details on the benefits of good blood pressure control in general and on established nephropathy in particular are described in chapters 16 and 17.



Incidence rates (95% confidence interval) of myocardial infarction and microvascular end points by category of updated mean systolic blood pressure, adjusted for age, sex, and ethnic group expressed for white men aged 50-54 years at diagnosis and mean duration of Type 2 diabetes of 10 years (UKPDS)

Smoking

This exacerbates all the complications of diabetes, both microvascular and macrovascular.

Dyslipidaemias

These increase the propensity to macrovascular disease: targets for control are described in chapter 17.

The presence of the above factors in combination additively increases the risks of developing complications.

The benefits of controlling glycaemia required persistently good control over a decade; benefits of successful blood pressure control were witnessed after approximately 4 to 5 years

Targets for control and reduction of risk factors

Blood glucose

The facility for patients to measure their own blood glucose empowers them to achieve optimal control by their own interventions. The aims are as follows:

- Type 1 diabetes: achieve preprandial blood glucose readings mainly in the range 4·5-7·7 mmol/l, postprandial readings in the range 6·0-9·0 mmol/1, and 7·0-9·0 mmol/1 at bedtime, and preferably never below 4·0 mmol/l to avoid blunting of hypoglycaemic awareness.
- Type 2 diabetes: fasting < 5·5 mmol/l;
  postprandial < 9·0 mmol/l.

Glycated haemoglobin

Aim for an HbA_{1c} < 6·5% (normal value 4·0-6·0%) as an ideal, since values > 7% are increasingly associated with development of all microvascular and macrovascular complications, and reduction of HbA_{1c} has been shown to diminish microvascular complications substantially (see below). Values up to 8% are acceptable in those who cannot readily achieve the ideal (and there are many). When HbA_{1c} values exceed 9%, additional education and counselling should be attempted although even then patients may not succeed, and some show no inclination to do so.

Blood pressure

Targets for control are described in chapter 17.

Targets for glycaemic control suggested by the European Diabetes Policy Group

	Low risk	Arterial risk	Microvascular risk
HbA_{1c} %	≤ 6·5	> 6·5	> 7·5
Venous plasma glucose			
Fasting/preprandial			
mmol/l	≤ 6·0	> 6·0	≥ 7·0
mg/dl	< 110	≥ 110	≥ 126
*Self-monitored blood glucose**			
Fasting/preprandial			
mmol/l	≤ 5·5	> 5·5	> 6·0
mg/dl	< 100	≥ 100	≥ 110
Postprandial or peak			
mmol/l	< 7·5	≥ 7·5	> 9·0
mg/dl	< 135	≥ 135	> 160

*Fasting capillary blood glucose is about 1·0 mmol/l (18 mg/dl) lower than venous plasma blood glucose. Postprandial capillary blood glucose is about the same as venous plasma blood glucose

Weight
Body mass index < 25 is ideal; 27 acceptable; greater than 30 represents obesity.

Lipids
Targets for control are described in chapter 17.

Smoking
Aim: to stop smoking.

Complications screening programme

Detection of the earliest signs of diabetic complications is an essential requirement of diabetes care leading to early preventive and treatment strategies which can abort progression of some of the most serious consequences.

Screening is ideally performed as a structured service undertaken by nurses and technicians outside the process of professional consultation, which should be informed by printed results from the screening programme. Screening should be performed at onset and then annually, from the onset of diabetes in all diabetic patients. Complications in Type 1 diabetes, however, are unlikely to develop during the first five years after diagnosis, so that the complete annual screening protocol can be deferred for a short time. The screening programme can be performed wherever appropriate facilities exist. Once complications are present and established, more frequent screening or treatment, or both may be needed.

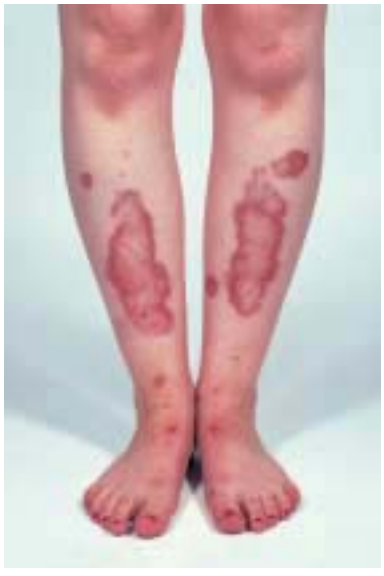
Eye screening requires specialist equipment and is often undertaken as a community responsibility, and there are strong representations that there should be a national screening programme. Detection and prevention of foot problems linked to delivery of adequate community podiatry services is also crucial and highly effective in preventing serious foot disorders.

The annual complications screening programme
This comprises:

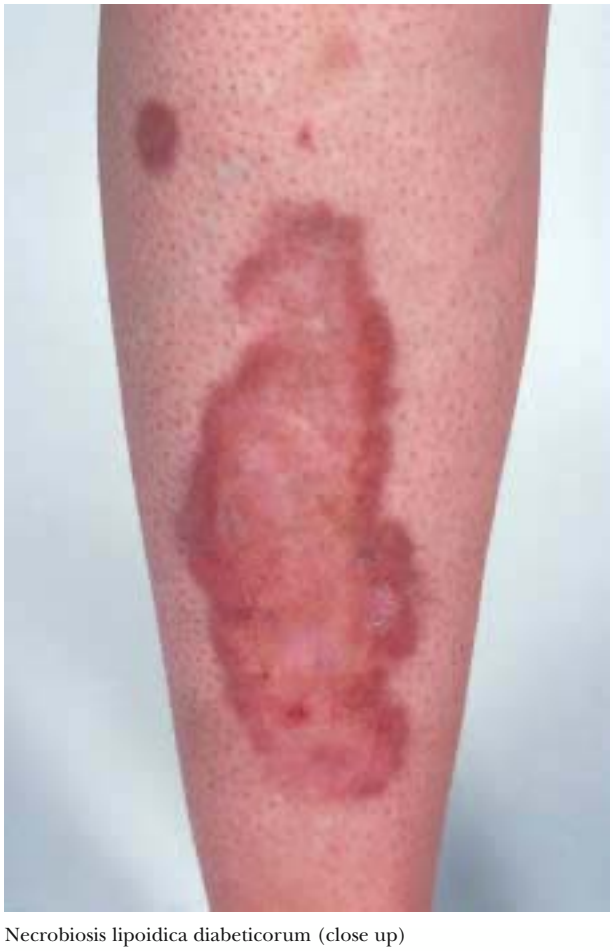
- weight (height): body mass index
- blood pressure
- eye examination (visual acuity, fundoscopy, and photography)
- foot examination:
  check for deformities, abrasions and ulcers
  sensation (monofilament tests, and other sensory modalities if available, see pages 52 and 63)
  palpate foot pulses
- blood tests: HbA_{1c}; lipid profile; creatinine
- urine tests: strip tests for proteinuria or microalbuminuria (if either of these are positive, total 24 hour proteinuria or the albumin creatinine ratio (ACR) should be measured, preferably on an early morning urine sample)
- assessment of smoking status.

Other complications

Necrobiosis lipoidica diabeticorum
Necrobiosis is an uncommon and unsightly blemish of the skin which chiefly affects diabetic women. It is unrelated to microvascular complications. The shin is the most common site. The lesions show rather atrophic skin at the centre with obviously dilated capillaries (telangiectasis) and a slightly raised pinkish rim; ulceration sometimes occurs. The lesions are indolent and rarely resolve. There is no effective treatment although steroid applications and even injection have been attempted.



Necrobiosis lipoidica diabeticorum



Necrobiosis lipoidica diabeticorum (close up)

Cheiroarthropathy

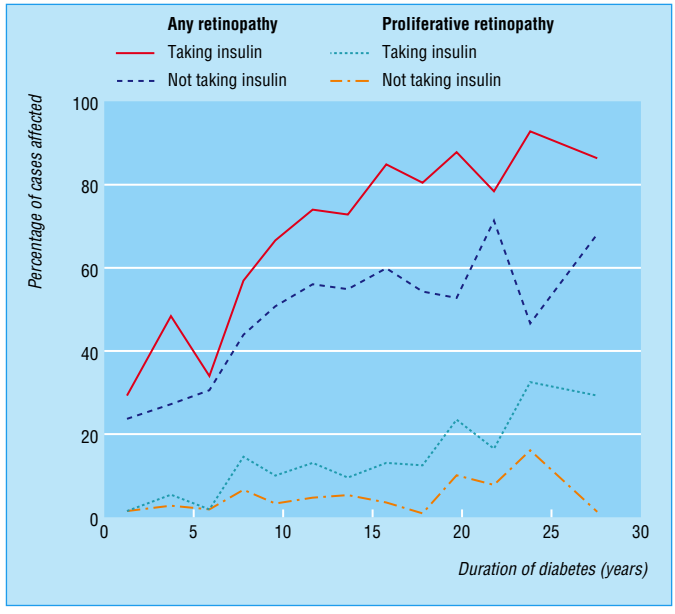
The development of tight, waxy skin, probably as a result of glucose related alteration of collagen structure, leads to some limitation of joint mobility. A relatively common yet symptomless consequence of these skin changes is the development of some fixed curvature of the fingers which may typically be seen in some patients with long-term diabetes. Those affected are unable to place the palm of the hand on a flat surface. The characteristic appearance is shown in the illustration on the right.

The figure showing the HbA_{1c} values during 9 years in the DCCT study of Type 1 diabetes is adapted from DCCT investigators. *New Engl J Med* 1993;329:977-86. The HbA_{1c} values during 15 years of UKPDS study of Type 2 diabetes and the Kaplan-Meier plot of aggregate microvascular end points are adapted from Diabetes Control and Complications Research Group. *New Engl J Med* 1993;329:977-86. The figures showing incidence rates and 95% confidence interval for myocardial infarction and microvascular complication by category of updated mean HbA_{1c} values and mean systolic blood pressure are adapted from UKPDS. *BMJ* 2000;321:405-17. The table showing targets for glycaemic control is adapted from European Diabetes Policy Group *Diabetic Med* 1999;16:716-30.

Cheiroarthropathy

11 Retinopathy

Blindness is one of the most feared complications of diabetes, but also one of the most preventable. Diabetes is the commonest cause of blindness in people aged 30 to 69 years. Twenty years after the onset of diabetes almost all patients with Type 1 diabetes and over 60% of patients with Type 2 diabetes will have some degree of retinopathy, and even at the time of diagnosis of Type 2 diabetes, approximately one-quarter of patients already have established background retinopathy. As treatment is now available to prevent blindness in the majority of cases, it is essential to identify patients with retinopathy before their vision is affected.



Frequency of retinopathy (any degree) and proliferative retinopathy by duration of diabetes in people receiving or not receiving insulin, and who were diagnosed to have diabetes at or after 30 years of age

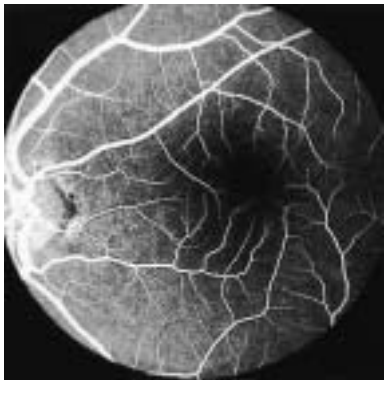
Classification of retinopathy

Diabetic retinopathy is due to microangiopathy affecting the retinal precapillary arterioles, capillaries, and venules. Damage is caused by both microvascular leakage due to break down of the inner blood-retinal barrier and microvascular occlusion. These two pathological mechanisms can be distinguished from each other by fluorescein angiography, which is the "gold standard" for assessing diabetic retinopathy.

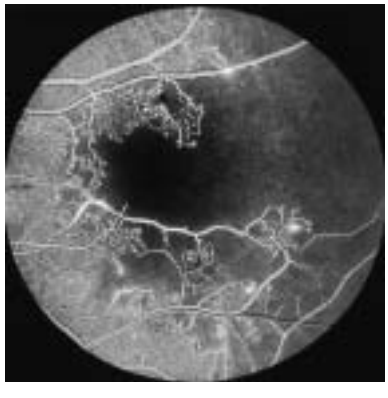
Background retinopathy

Microaneurysms are small saccular pouches possibly caused by local distension of capillary walls. They are often the first clinically detectable sign of retinopathy, and appear as small red dots commonly temporal to the macula.

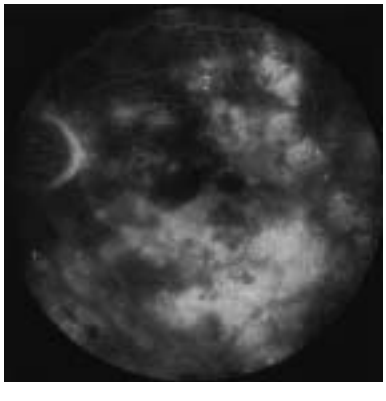
Haemorrhages may occur within the compact middle layers of the retina, and appear as "dot" or "blot" haemorrhages, or



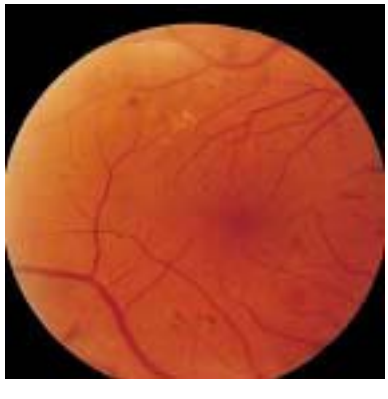
Fluorescein angiogram showing normal eye



Fluorescein angiogram showing capillary closure



Fluorescein angiogram showing capillary leakage with macular oedema



Microaneurysms, haemorrhages, and exudates

rarely, in the superficial nerve fibre layer where they appear as "flame shaped" haemorrhages (the latter better recognised as related to severe hypertension).

Hard exudates are yellow lipid deposits with relatively discrete margins. They commonly occur at the edges of microvascular leakage, and may form a "circinate" pattern around a leaking microaneurysm. They may coalesce to form extensive sheets of exudate. Vision is affected when hard exudates encroach upon the macula.

Retinal oedema is due to microvascular leakage and indicates breakdown of the inner blood-retinal barrier. It appears clinically as greyish areas of retinal thickening, and may assume a petal-shaped cystoid appearance at the macula, where it may cause marked visual deterioration.

Clinically significant macular oedema (CSMO) requires treatment. It is defined as any one of the following:

- retinal oedema within 500 μm (one-third of a disc diameter) of the fovea
- hard exudates within 500 μm of the fovea, if associated with adjacent retinal thickening
- retinal oedema that is one disc diameter (1500 μm) or larger, any part of which is within one disc diameter of the fovea.

Twenty percent of eyes with untreated CSMO will suffer significant visual loss in two years compared with 8% of treated eyes.

Pre-proliferative retinopathy

Retinal ischaemia due to microvascular occlusion may lead to neovascular proliferation. Signs of ischaemia include:

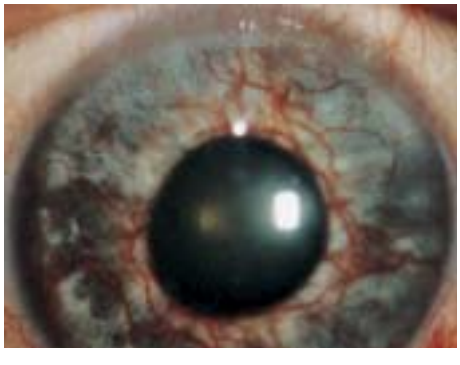
- cotton wool spots, which appear as white patches with rather feathery margins and represent nerve fibre layer microinfarcts; they become highly significant when there are more than five
- large dark "blot" haemorrhages
- venous beading and looping
- intraretinal microvascular abnormalities (IRMA).

Proliferative retinopathy

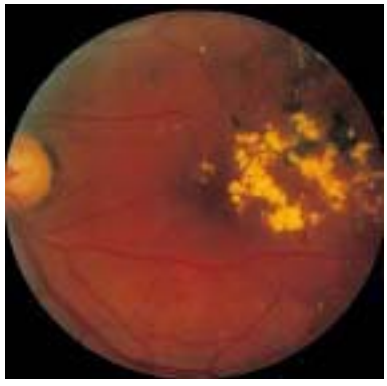
New vessel formation may occur at the optic disc (NVD) or elsewhere on the retina (NVE). Disc new vessels are particularly threatening to vision and if allowed to progress commonly lead to vitreous haemorrhage. If untreated, 26% of eyes with "high-risk" and neovascular proliferation on the disc will progress to severe visual loss within two years. With laser treatment this is reduced to 11%.

Advanced eye disease

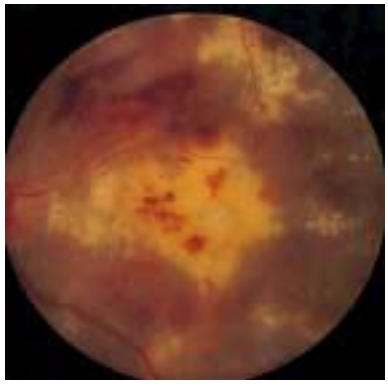
In advanced proliferative diabetic retinopathy, progressive fibrovascular proliferation leads to blindness due to vitreous haemorrhage and traction retinal detachment. Rubeosis iridis and neovascular glaucoma occur when new vessels form on the



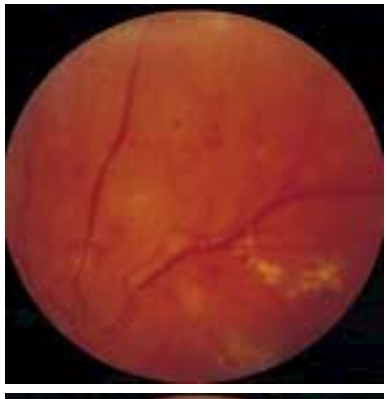
Rubeosis iridis



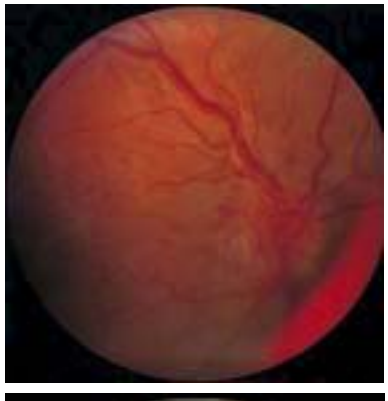
Exudative maculopathy



Exudative maculopathy



Pre-proliferative retinopathy with venous bleeding, cotton wool spots and some hard exudates



Disc new vessels (NVD)

New vessels elsewhere (NVE)

iris and in the anterior chamber drainage angle, leading to a most painful blind eye which occasionally requires enucleation.

Blindness in diabetic patients

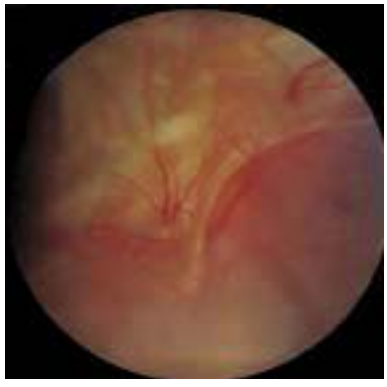
Vision-threatening retinopathy is usually due mainly to neovascularisation in Type 1 diabetes and maculopathy in Type 2 diabetes. In North America, 3·6% of patients with Type 1 diabetes and 1·6% of patients with Type 2 diabetes are legally blind. In England and Wales about 1000 diabetic patients are registered as blind or partially sighted each year, with diabetic retinopathy being the commonest cause of blindness in the working population.

Vitreous haemorrhage occurs suddenly and painlessly. The blood usually clears over the following weeks, but the underlying proliferative retinopathy causes repeated haemorrhages and progressive visual loss in most cases if it is not treated. Retinal detachment resulting from contracting fibrous bonds sometimes causes blindness.

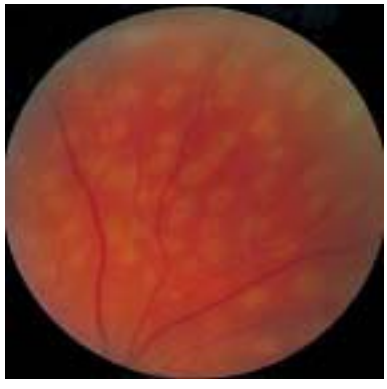
Maculopathy: Macular disease has three causes in diabetic patients—exudative maculopathy, retinal oedema, and ischaemia. Deterioration of vision in these situations is often insidious, it can to some extent be prevented by appropriate laser treatment, but once vision has been lost it cannot be restored. Ischaemic maculopathy due to loss of perifoveal capillaries may cause severe visual loss and is very difficult to treat.

Cataract: Lens opacities or cataract develop earlier in diabetic patients and often progress more rapidly.

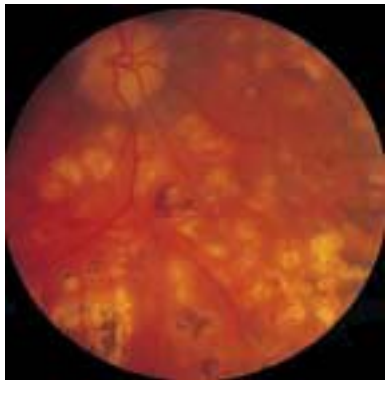
Primary open-angle glaucoma has an increased prevalence in diabetic patients compared with the general population.



Advanced eye diseae—Retinitis proliferans



Recent argon laser photocoagulation



Longstanding photocoagulation scars

Prevention of blindness

The presence of retinopathy must be actively sought by physicians because, if detected early enough, blindness can be prevented in many cases by treatment with laser photocoagulation. The indications for laser treatment are:

- NVD or NVE; advanced pre-proliferative changes
- clinically significant macular oedema as defined above
- encroachment of hard exudates towards the fovea.

Chronic vitreous haemorrhage which precludes a view of the retina can be treated by vitrectomy and endolaser. Tractional retinal detachment can be managed by vitrectomy with the use of heavy liquids and silicone oil. Restoration of visual acuity can be impressive, but is dependent on the underlying condition of the retina itself.

Clinical examination of the eyes and screening

(For further details on screening see page 45)

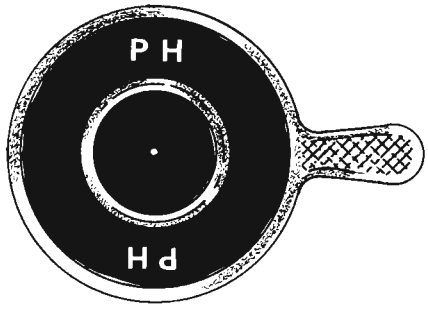
Visual acuity and retinal examination should be performed annually on all diabetic patients after 12 years of age, or more often if advancing changes are observed. Vision-threatening retinopathy rarely occurs in Type 1 diabetes in the first five years after diagnosis or before puberty. However, more than one-quarter of Type 2 diabetic patients have been found to have retinopathy at diagnosis, and screening should start immediately.

Visual acuity should be checked annually, or more often if significant retinopathy is present or if it has changed

unexpectedly. This should be done with patients wearing their spectacles or through a "pinhole" if they are not.

Retinal examination. Routine fundal examination should be performed on all diabetic patients, using fundoscopy or retinal photography or preferably both. The pupils should be dilated and the fundus examined in a darkened room. Tropicamide 1% (Mydriacil) eye drops are recommended as they have a short duration of action of just two to three hours. There is no reason to avoid pupillary dilatation in patients being treated for chronic open-angle glaucoma, although those on treatment for closed-angle glaucoma must not undergo pupillary dilatation.

Once background retinopathy is present the patients should be examined every six to 12 months or more often if there is any change of visual acuity, and referred to an ophthalmologist when indicated (see box). Pregnant patients require more frequent follow up as retinopathy may progress rapidly during pregnancy (see page 78).



Pin hole

Digital camera used for retinal photography

Screening methods

Conventional examination, using an ophthalmoscope in a darkened room with the pupil dilated is a minimum requirement. Observers must be well trained, but even consultant ophthalmologists do not achieve the required 80% sensitivity.

Retinal photography through dilated pupils. The preferred method now uses digital photography which yields suitable images which can be electronically stored, making them easily available for consultation, review, and teaching. Conventional colour photographs also provide good images, whereas the quality of Polaroid photographs is less than ideal.

It would be ideal to provide both conventional funduscopic and photographic screening procedures, and there is already some evidence that the combined screening procedure reduces the failure rate. A national screening programme has been proposed and has already been adopted in Wales.

Indications for referral to an ophthalmologist

- Reduced visual acuity from any cause
- Presence of proliferative or pre-proliferative changes
- Presence of clinically significant macular oedema
- Presence of hard exudates near the macula
- Presence of any form of progressing or extensive diabetic retinopathy especially when the lesions are near the macula

The blind diabetic patient

Blind registration is available for those patients with visual acuity of less than 3/60 in their better eye or gross field defects, affording some financial help and social service support. Patients with a visual acuity of less than 6/60 in their better eye are eligible for registration as partially sighted. They must be registered by an ophthalmologist using the BD8 form. Printing in braille is valuable but many diabetic patients have impaired fine sensation in their fingertips, making it difficult for them to read it. Insulin "pens", in which palpable clicks correspond to units of insulin are valuable for blind patients.

The figure showing frequency of retinopathy is adapted from Pickup JC, Williams G, eds. *Textbook of diabetes*, 2nd ed, Oxford: Blackwell Scientific Publications, 1997.

12 Peripheral neuropathies

Diabetic neuropathies constitute a diverse group of conditions. The commonest is a diffuse polyneuropathy which damages distal peripheral nerves (chiefly affecting the feet), together with the autonomic nervous system. The dying back of axons is associated with segmental demyelination. Polyneuropathy is a classic diabetic complication developing mainly in those with poor diabetic control, progressing (albeit at very variable rates) as the duration of diabetes lengthens and often, but not always, associated with other long-term diabetic complications. In contrast, mononeuropathies and acute painful neuropathies run a well-defined course from the relatively acute onset to almost complete recovery in six to 18 months. These reversible neuropathies, which may be the reason for initial presentation of diabetes, can occur after any duration of diabetes, are commoner in Type 2 diabetic men, and are not necessarily associated with other diabetic complications.

Pressure neuropathies are commoner in those with diabetes and include carpal tunnel syndrome (median nerve), ulnar neuropathy, and rarely foot drop (lateral popliteal nerve).

Diabetic neuropathies

Progress
- Diffuse polyneuropathy
  - Symmetrical sensory neuropathy
  - Autonomic neuropathy

Recover
- Mononeuropathies
  - Proximal motor (femoral) neuropathy
  - Radiculopathies (especially truncal)
  - Cranial nerve palsies
- Acute painful neuropathies

Symmetrical sensory neuropathy

Diffuse neuropathy affects peripheral nerves symmetrically, chiefly those of the feet and legs. It is almost always sensory, though motor involvement causing weakness, and wasting does occur rarely. Peripheral neuropathy is common in long-standing diabetic patients, but in Type 2 diabetic patients it may already be present at the onset of diabetes. Progression of neuropathy is reduced by good control of diabetes over many years. The potential of pharmacological agents to alter the course of neuropathy has been extensively studied, but so far none of the drugs investigated has demonstrated convincing clinically significant benefit.

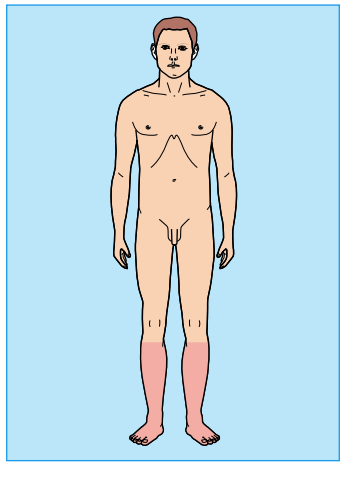
Neuropathy is usually symptomless and therefore a hazard to the unwary patient, who is at risk of foot injury and infection. In more advanced neuropathies the patient is aware of sensory loss; numbness (and in some, a sensation of coldness) may progressively worsen until there is almost complete anaesthesia below the knee associated with proprioceptive loss which makes patients feel quite unsafe, but this is not common. Paraesthesiae are quite often described by patients: they range from a persistent minor inconvenience to a source of considerable discomfort and even pain needing medication to alleviate the symptoms. Management of painful neuropathy is described on page 57.

Neurological examination almost always reveals absent ankle reflexes, and only rarely absent knee reflexes. Diminished light-touch and vibration perception are common and by careful examination can be shown to occur at an almost identical level in both legs, hence the description of "stocking" neuropathy. Some patients demonstrate a highly selective form of sensory impairment with gross loss of pain and thermal sensation (accompanied by severely abnormal autonomic function tests) while light-touch and vibration perception remain almost intact. Such a dissociated sensory loss can cause some confusion in the clinical assessment of neuropathy. Impairment of joint position sense is extremely unusual except in the most advanced cases. Further details of clinical sensory assessment are described on page 63.

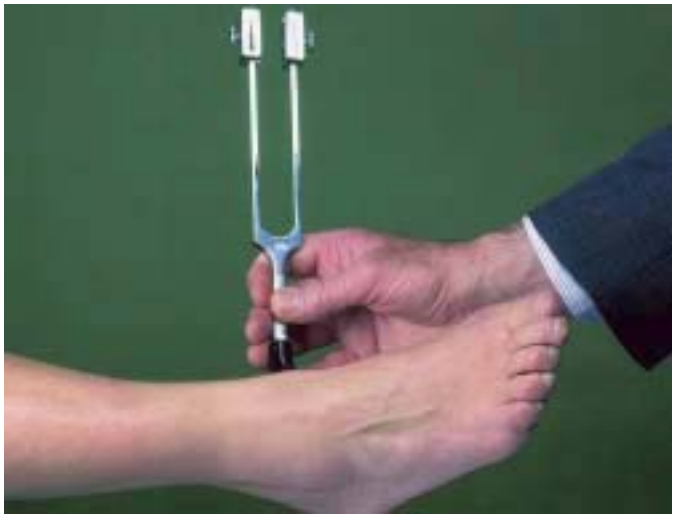
Diabetic neuropathies

	Recover	Progress
	Mononeuropathies/ painful neuropathies	Sensory/autonomic neuropathy
Onset	acute	gradual
Duration of diabetes mellitus	any	long standing
Other complications	none	often
Sex	M > F	M = F*
Diabetes mellitus	Type 2 > Type 1	Type 1 = Type 2*

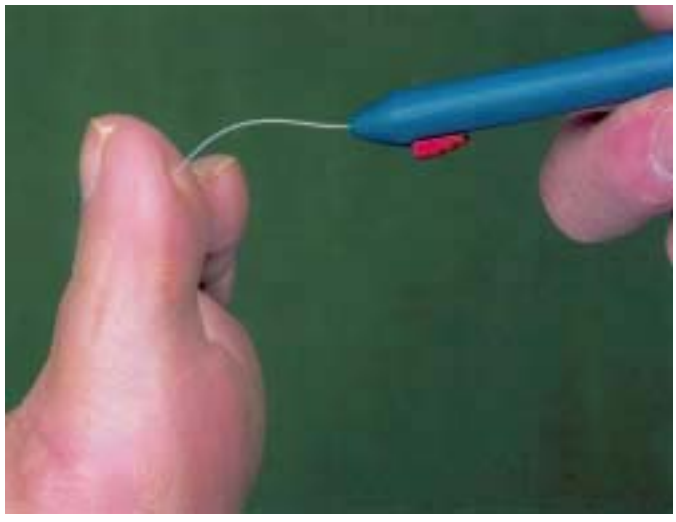
*Symptomatic autonomic neuropathy occurs chiefly in female Type 1 diabetic patients



Symmetrical sensory neuropathy



Testing vibration perception at the medial malleolus with a Rydell Seifer quantitative tuning fork



Testing light-touch sensation with a monofilament

Neuropathy and the hands

Diabetic neuropathy rarely causes symptoms in the hands, and when it does the disease is already advanced in the feet and legs. Numbness and clumsiness of the fingers are thus very unusual and more likely to be due to some other neurological disorder. Impairment of sensation is, however, enough to prevent blind diabetics from reading braille. Paraesthesiae and numbness in the fingers, especially at night, are usually due to carpal tunnel syndrome, which is commoner than in non-diabetics. It is easily and effectively relieved by minor surgery performed under local anaesthetic without admission to hospital.

Interosseous muscle wasting, especially of the first dorsal interosseous, is often seen. It is usually due to ulnar nerve compression at the elbow, and typical sensory defects in the fourth and fifth fingers are detectable. It causes little disability and there is no satisfactory treatment. Patients are advised not to lean on their elbows, thereby avoiding further damage to the ulnar nerve.

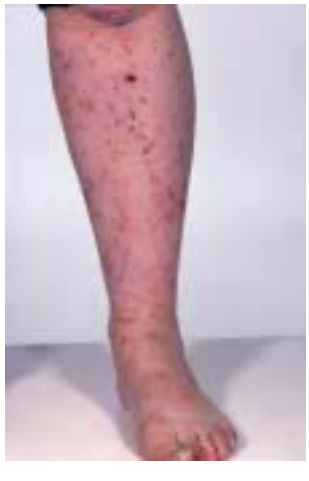
Neuropathy and the feet

Reduced sensation in the feet may result in unnoticed trauma from ill-fitting shoes, nails or stones when walking barefoot, or burns from hot water bottles or sitting too close to a fire. Self-inflicted wounds from crude attempts at chiropody are dangerous because they often become infected. Proprietary corn cures which contain salicylic acid can cause ulceration, sepsis and necrosis, and should never be used. Diabetic foot disorders are described in chapter 15.

Features of neuropathy of the hands

- Inability to read braille
- Carpal tunnel compression
- Ulnar nerve compression

Interosseous muscle wasting



Cat scratches on a neuropathic, insensitive leg

13 Autonomic neuropathy

Diffuse damage to both parasympathetic and sympathetic nerves, probably developing in that order, is common in diabetic patients with diffuse peripheral neuropathy. Fortunately the disabling symptoms which result are not common, and even when they do occur some of them, especially diarrhoea, vomiting, and postural hypotension, are curiously intermittent.

Clinical features of autonomic neuropathy

Gastrointestinal
- Diarrhoea
- Gastroparesis

Cardiovascular
- Postural hypotension
- Persistent tachycardia
- High foot blood flow
- Vascular medial calcification

Genitourinary
- Erectile dysfunction
- Neurogenic bladder

Sweating
- Gustatory sweating
- Dry feet

Respiratory
- Depressed cough reflex
- Respiratory arrests
- ? Deaths from respiratory arrests

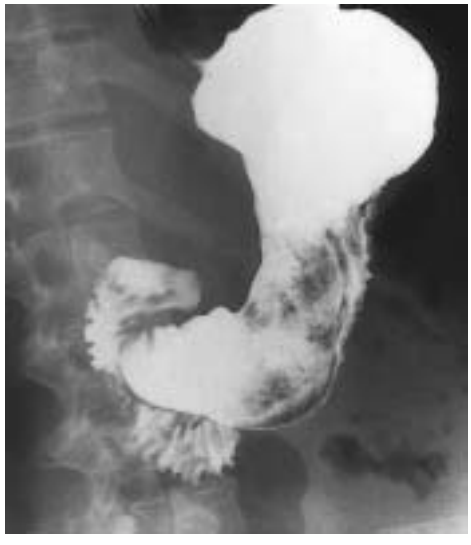
Gastrointestinal system

Diarrhoea

This is a catastrophic watery diarrhoea with severe nocturnal exacerbations and faecal incontinence, preceded momentarily by characteristic abdominal rumblings. Malabsorption does not normally occur. The symptoms are intermittent, with normal bowel actions in between, and sometimes even constipation. These features persist for months or years, rarely disappearing altogether. The diagnosis is made, firstly, by excluding other causes of diarrhoea such as coeliac disease or pancreatic malfunction, and secondly, by establishing the presence of peripheral and autonomic neuropathy. The diarrhoea may be treated with any antidiarrhoeal agent, the best of which is codeine phosphate. Tetracycline in two or three doses of 250 mg has a dramatic effect in about half of patients; it should only be used at the onset of an attack. Some authorities suggest the use of tetracycline or metronidazole for two to three weeks but long-term antibiotics are not indicated. Clonidine can be tried but is of little value; octreotide can be effective but side effects are common.

Gastroparesis

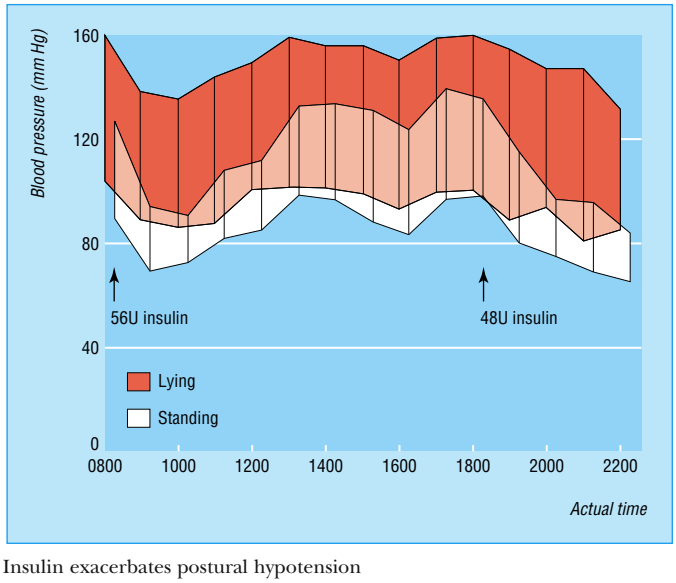
Diminished gastric motility and delayed stomach emptying sometimes occur in diabetic patients with autonomic neuropathy, but rarely cause symptoms. Intermittent vomiting may occur, and in exceptional cases it is intractable. The diagnosis is established by the presence of a gastric splash, impaired gastric emptying on investigation by isotopic techniques, and screening during barium studies; endoscopy and other investigations are needed to exclude other gastric disorders. Any antiemetic can be useful, including metoclopramide or domperidone. Erythromycin acting as a motilin agonist has been used but is only of value when used intravenously. In the very rare cases of intractable vomiting, percutaneous endoscopic jejunostomy can help, and even more rarely radical surgery by two-thirds gastrectomy with Roux-en-Y loop may be required and can succeed.



Radiograph showing food retention caused by gastroparesis

Cardiovascular system: postural hypotension

Postural hypotension is defined by a fall in systolic blood pressure on standing of more than 20 mm Hg. Development of symptoms depends both on the actual fall of blood pressure (usually greater than 30 mm Hg) and on the actual systolic blood pressure when standing, which becomes impossible if it is less than 70 mm Hg. When assessing postural hypotension, the blood pressure should be taken with the patient lying down,



Insulin exacerbates postural hypotension

and the standing reading taken over approximately three minutes during which the time blood pressure continues to fall.

Treatment is only needed if symptoms occur and if they are troublesome, which is rare. Patients should stop drugs which might aggravate hypotension (notably tranquillisers, antidepressants, and diuretics), sleep with the head of the bed raised, and wear full length elastic stockings. The best results are obtained from measures which increase plasma volume, namely a high salt intake and fludrocortisone (increasing the dose slowly from 0·1 to 0·4 mg). Treatment failures are common, and the oedema which results from treatment may be disagreeable. Successful treatment has been reported with indomethacin; a combination of fludrocortisone, flurbiprofen, and ephedrine may help. Midodrine is an α agonist which is also of value and available on a named patient basis.

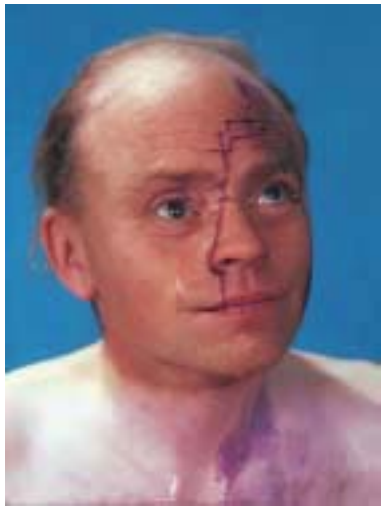
Treatment of hypotension

- Stop drugs that may aggravate hypotension
- Sleep with head of bed raised
- Wear full length elastic stockings
- Appropriate medication

Gustatory sweating

Facial sweating (including scalp, neck, and shoulders) which occurs while eating tasty food, notably cheese, is a common symptom of autonomic neuropathy. Once present, it seems to persist indefinitely, although amelioration after renal transplantation occurs for no known reason.

When it becomes a severe embarrassment with sweat rolling down the face and chest at every meal, it can be effectively treated with an anticholinergic agent, namely propantheline bromide, although side effects are common, or by a topical application of glypyrronium powder. The cream should be applied on alternate days to the areas affected by sweating, avoiding contact with the mouth, nose, and eyes. The area should not be washed for four hours after application. Systemic absorption is low and the only contraindication is narrow-angle glaucoma, as there is the possibility of accidental direct instillation into the eye. Although recommended to be given on alternate days, many patients prefer to use it only on social occasions.



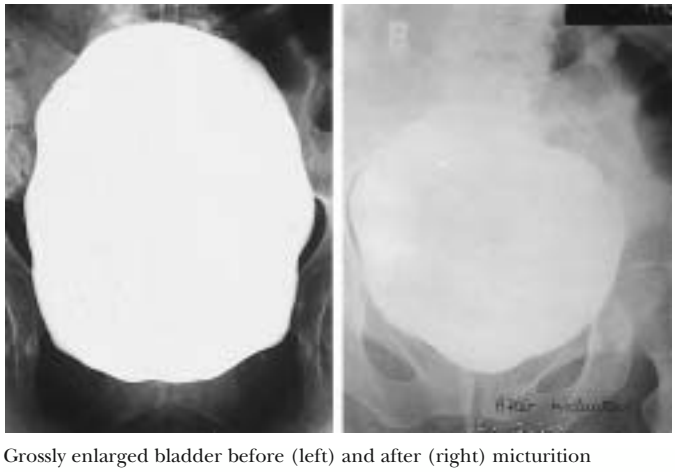
Gustatory sweating. The sweating is highlighted by starch-iodide powder

Respiratory arrests

Transient respiratory arrest occurs sometimes if susceptible neuropathic patients are given any agent which depresses respiration, notably anaesthetics or powerful analgesics such as morphine and its derivatives. These patients must be monitored carefully even during minor surgery. Rare unexplained deaths in patients with established autonomic neuropathy may be due to respiratory arrest.

Neurogenic bladder

Urinary retention is a serious and usually late complication of autonomic neuropathy. Apart from the discomfort, intractable urinary infections may develop. The diagnosis of bladder retention is now simple using ultrasound techniques. Cystoscopy may be needed to exclude other causes of bladder neck obstruction. Treatment is now by self catheterisation two or three times daily.



Grossly enlarged bladder before (left) and after (right) micturition

Erectile dysfunction

Erectile dysfunction is a common problem, occurring more often in those with diabetes than in others, and is due to neuropathy or peripheral vascular disease, or both. It is also frequently has psychogenic causes and some drugs can also be responsible. It is important that appropriate advice and treatment is sought from trained counsellors.

Autonomic neuropathy causes erectile impotence, which is permanent and irreversible. Retrograde ejaculation can also occur. The onset of organic impotence in neuropathy is always gradual, progressing over months or even years. Erectile ability fails first, and ejaculation declines later. Nocturnal erections are absent in these patients, whereas they are often retained in psychogenic impotence. It is often difficult to distinguish between organic and psychogenic erectile dysfunction in diabetic patients. The presence of peripheral and autonomic neuropathy makes an organic cause more likely, especially when other autonomic symptoms are present. After a full clinical examination, patients complaining of erectile dysfunction have free testosterone and serum prolactin levels measured in order to exclude other disorders, but levels are usually normal in patients with erectile dysfunction due to diabetes. There is no cure for autonomic impotence. Hormone treatment with testosterone is useless because it serves only to increase libido without improving erectile ability. In many cases, careful explanation of the cause to the affected couples will allay their fears and anxieties.

Features of erectile dysfunction

Organic
- Gradual onset
- Permanent
- Absent nocturnal erection
- Ejaculation often retained

Psychogenic
- Sudden onset
- Intermittent
- Nocturnal erections occur
- Penile tumescence tests give normal results

All patients complaining of erectile dysfunction should undergo a full clinical examination including examination of the external genitalia

Treatment of erectile dysfunction

Diabetes centres should now offer a specialist service for advice on management of patients with erectile dysfunction. This should include psychosexual therapy which can succeed in 50% to 80% of patients who are well motivated and is also of value in conjunction with specific treatments.

Several effective specific treatments are now available: oral sildenafil is generally the first choice, and sublingual apomorphine serves as a second choice.

Summary of treatments of erectile dysfunction
- Oral sildenafil
- Sublingual apomorphine
- Prostaglandin preparations
- Vacuum devices
- Penile prostheses

Oral sildenafil

This can be successful in almost two-thirds of the diabetic patients treated, which is rather less than in non-diabetic people where even higher rates of success have been reported. It is taken half to one hour before sexual activity (initial dose 50 mg; subsequently 50 to 100 mg according to response; not to be used more than once in 24 hours). Sexual stimulation and foreplay are necessary for it to be effective. It is contraindicated in those taking nitrates, those whose blood pressure is less than 90/50 mm Hg, after recent stroke or myocardial infarction, or in other situations where sexual activity is inadvisable. There are several potential side effects which are rarely troublesome (see *BNF*).

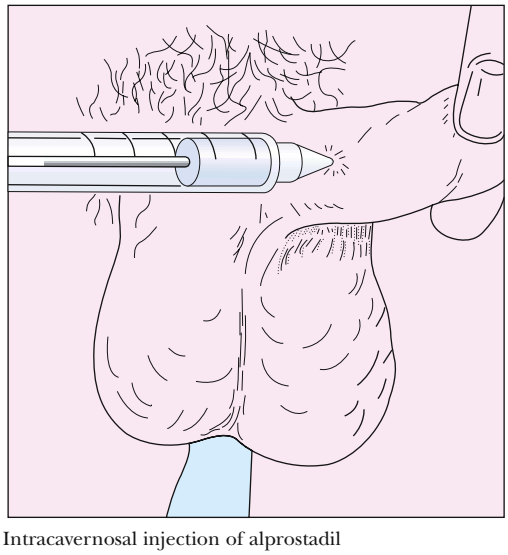
Sublingual apomorphine

This is rapidly absorbed and acts as a dopamine agonist. It is effective within 10 to 20 minutes, requiring sexual stimulation at the same time. The dose range is 2 to 3 mg. It is effective in approximately 50% of diabetic patients.

Prostaglandin preparations

Transurethral alprostadil can provide erections adequate for intercourse. An applicator for direct urethral application is provided.

Intracavernosal injection: alprostadil is now used less than previously but can be effective; modern injection systems have made its use acceptable. The side effects and contraindications are described in the *BNF*.



Intracavernosal injection of alprostadil

Vacuum devices

An external cylinder is fitted over the penis, enabling air to be pumped out, resulting in penile engorgement which is sustained by application of a ring fitted to the base of the penis. This technique is suitable for a wide range of patients and is

successful in most cases of erectile dysfunction, although when the problem is of long standing several attempts may be necessary before a successful erection is achieved.

Penile prostheses
Semirigid, malleable prostheses can be surgically inserted and are particularly valuable for younger patients with confirmed and permanent neuropathic impotence. Ejaculation in some patients is retained.

Diagnosis of autonomic neuropathy

Gustatory sweating is the only symptom which is almost pathognomic of diabetic autonomic neuropathy. Peripheral neuropathy (at least absent ankle jerks) must be present before the diagnosis can be made. A resting tachycardia, postural hypotension, or a gastric splash may be present.

Bedside cardiovascular tests for autonomic neuropathy are now well established: their most important role is probably in the exclusion of autonomic neuropathy. Normal and abnormal values are shown the table.

The loss of heart rate variability during deep breathing due to vagal impairment is the most reliable and simplest test of autonomic neuropathy. It is best assessed using a cardiotachograph during deep respirations (six breaths per minute) taking average readings during six breaths; it can also be performed using an ordinary electrocardiograph during a single deep breath (five seconds in, five seconds out). The heart rate difference (maximum rate during inspiration minus minimum rate during expiration) in those under 55 years old is always greater than ten. Heart rate increase on standing up should be assessed, and there should normally be an overshoot as well.

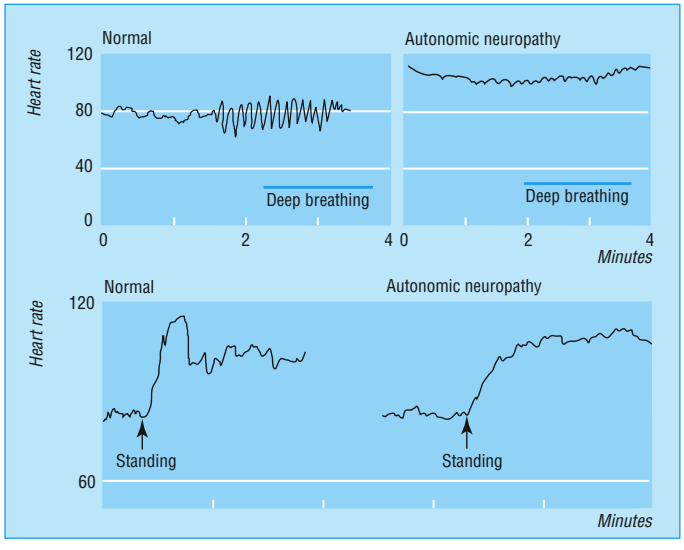
The Valsalva manoeuvre can be included among the tests: a mercury sphygmomanometer is used, the patient blowing hard into the empty barrel of a 20 ml syringe to maintain the mercury column at 40 mm Hg for 10 seconds. Maximum heart rate during blowing, followed by minimum heart rate after cessation, are recorded. There should be a bradycardia after cessation of blowing; the ratio of maximum : minimum heart rate is normally greater than 1·21 and clearly abnormal when less than 1·10. The Valsalva test should not be performed in those with proliferative retinopathy. Many other sophisticated tests need special equipment.

The figure showing the intracavernosal injection of alprostadil is from Tomlinson J, ed. *ABC of Sexual Health*. London: BMJ Publishing Group, 1999.

Normal values for autonomic function tests*

	Normal	Abnormal
Heart rate variation (deep breathing) (beats/min)	>15	<10
Increase in heart rate on standing (at 15 seconds) (beats/min)	>15	<12
Heart rate on standing 30:15 ratio	>1·04	<1·00
Valsalva ratio	>1·21	<1·20
Postural systolic pressure fall at 2 min	<10 mm Hg	>30 mm Hg

*These test results decline with age. The figures apply generally in those less than 60 years old.



Heart rate changes in a normal subject (left) and a patient with autonomic neuropathy (right) showing loss of heart rate variation in autonomic neuropathy during deep breathing, at six breaths a minute (top), and loss of "overshoot" cardiac acceleration on standing (bottom)

14 Mononeuropathies and acute painful neuropathies

The rapid onset, severity, and eventual resolution of mononeuropathies contrast sharply with the long-term nature and irreversibility of diffuse peripheral neuropathy. The two forms of neuropathy occur quite independently of each other. Mononeuropathies are more frequently seen in Type 2 diabetic men and may even occur as the presenting symptom of diabetes.

Acute painful neuropathies

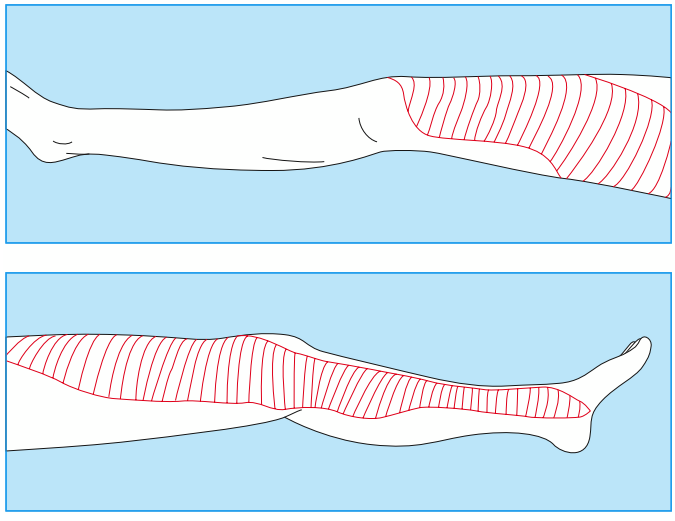
Acute painful neuropathies begin relatively acutely at any stage of diabetes, sometimes paradoxically eight to 12 weeks after starting insulin, or as the presentation of Type 2 diabetes. The acute and persistent pain can be disabling. Distribution of the pain is radicular over the territory of several adjacent nerve roots, affecting either the legs or the abdominal wall (the latter very rarely accompanied by muscle bulging from motor weakness). The thighs are affected in patients with femoral neuropathy. Both feet and legs can be affected symmetrically in a "stocking" distribution. Patients usually recover from these neuropathies in a period of six to 18 months. These neuropathies occur independently of peripheral sensory or autonomic neuropathy.

The pain causes exceptional distress because it is protracted and unremitting. Constant burning sensations, paraesthesiae or shooting pains occur, but the most characteristic symptom is a cutaneous hypersensitivity (allodynia) leading to acute discomfort on contact with clothing and bedclothes. The pain leads to insomnia and depression, and is sometimes accompanied by catastrophic weight loss. Patients are so distressed that they may seek several opinions on their condition, and often believe that they must have a malignant disease.

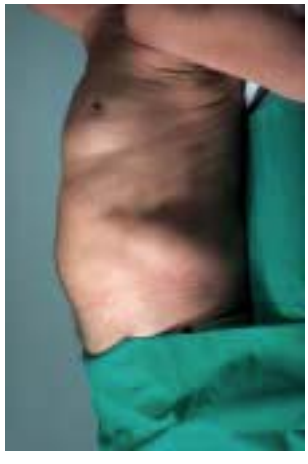
Treatment

This is difficult, but above all, the promise that the symptoms always eventually remit may sustain patients during the wretched months of their illness. It sometimes helps for them to meet a patient who has already recovered from neuropathic pain. Diabetic control should be optimal, and insulin should be given if necessary. Initially, simple analgesics such as paracetamol taken regularly should be tried. Tricyclic antidepressants have a specific effect in the management of neuropathic pain and are valuable in this condition: a useful combination is a preparation containing a phenothiazine (fluphenazine) with nortryptiline (Motival). Gabapentin is effective and carbamazepine may help. Capsaicin cream may have a small effect after itself causing some initial discomfort. Topiramate may be used for limited periods in those with severe and protracted pain. Drugs of addiction should be avoided although just occasionally and for a short period an opioid derivative can be used at bedtime to help distressed patients to sleep.

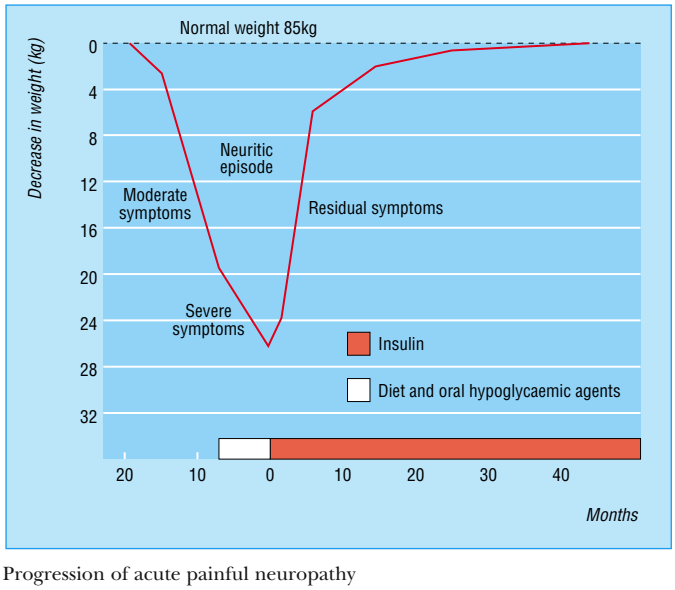
Application of Opsite (a thin adhesive film) can help alleviate contact discomfort. Electrical nerve stimulators applied to the site of pain may help, and patients can then take an active part in their treatment.



Diabetic radiculopathy area of exquisite contact sensitivity can easily and reproducibly be traced with a finger—shown in red



Bulging of the abdominal wall in a patient with truncal radiculopathy



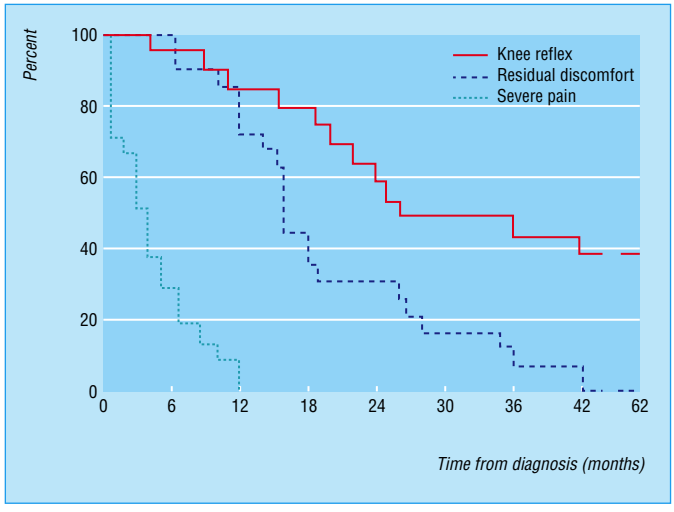
Progression of acute painful neuropathy

Femoral neuropathy
(proximal motor neuropathy or diabetic amyotrophy)

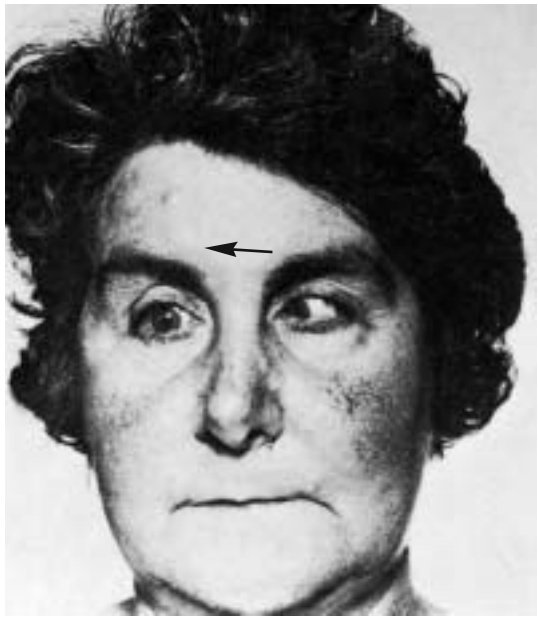
Pain with or without wasting in one or both thighs is the cardinal feature of this disagreeable condition. The quality of the pain is similar to that in painful peripheral neuropathy described above, and management is along similar lines. The knee jerk is absent while the ankle jerk is often retained. Sensation in the thigh may be altered or impaired. In some cases, motor weakness in the thigh is profound, causing falls. Other neurological disorders must be considered and excluded. Full recovery within about one year is the rule. Active physiotherapy may be needed to strengthen the wasted thigh muscles and restore mobility.

Cranial nerve palsies
Third and sixth nerve palsies presenting with diplopia of sudden onset are characteristic. Pain behind the eye occurs sometimes in third nerve palsies; the pupil is usually spared, and ptosis does not normally occur. Full examination and careful follow up are needed, but extensive investigation is not normally required. Complete recovery occurs spontaneously in about three months.

The first illustration is adapted from Bloom A, Ireland J. *A colour atlas of diabetes,* 2nd ed. London: Wolfe Publishing Ltd, 1992. The figure showing recovery from femoral neuropathy is adapted from Coppack SW, et al. *Quart J Med* 1991;79:307–13. The photograph of cranial nerve palsy is reproduced from Spillane, JD. *An atlas of clinical neurology.* Oxford: Oxford University Press.



Recovery from femoral neuropathy



Cranial nerve palsy

15 The diabetic foot

Foot ulceration, sepsis, and amputation are universally known and feared by almost every person on hearing that they have diabetes. Yet at the same time, these are potentially the most preventable of all diabetic complications by the simplest techniques of education and care; and if lesions do occur, the majority can be cured by immediate and energetic treatment, for which good provision must be made.

Diabetic foot problems

Neuropathic foot
- Painless
- Calluses, ulcers, sepsis, osteomyelitis
- Charcot joints, oedema, good pulses

Critically ischaemic foot
- Painful, pink, cold, no pulses

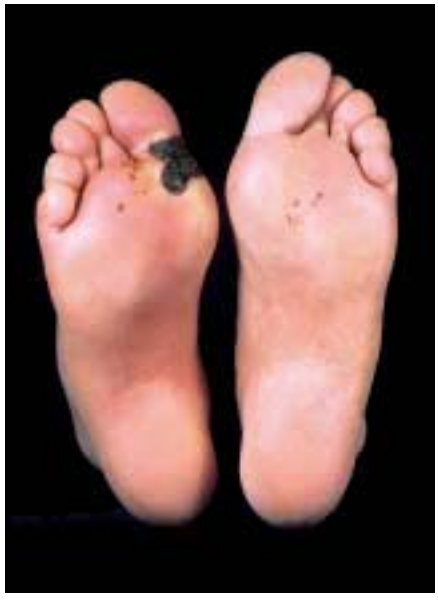
Diabetic foot disorders

Neuropathy and ischaemia are the principal disorders underlying foot problems. Whenever a patient presents with an active lesion it is essential to decide at an early stage whether the foot problem is:

(a) neuropathic with an intact circulation
(b) ischaemic with (usually) or without neuropathy (neuro-ischaemic foot)
(c) critically ischaemic needing very urgent attention.

A combination of ulceration and sepsis in an ischaemic foot carries a higher risk of gangrene, and early arterial assessment and management are key to avoiding major amputation.

Men of low socioeconomic class are most prone, and Asian patients least liable, to diabetic foot disorders.



Corn plaster injury

Precipitating causes of foot ulceration and infection

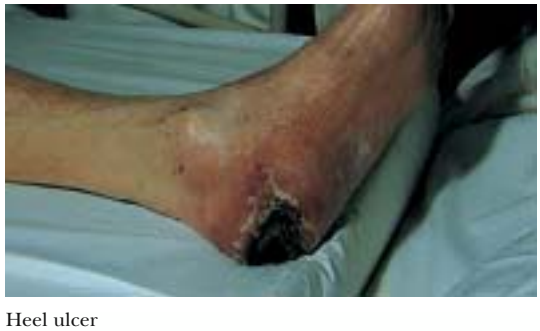
- Friction in ill fitting or new shoes
- Untreated callus
- Self treated callus
- Foot injuries (for example, unnoticed trauma in shoes or when walking barefoot)
- Burns (for example, excessively hot bath, hot water bottle, hot radiators, hot sand on holiday)
- Corn plaster
- Nail infections (paronychia)
- Artifactual: rarely, self-inflicted foot lesions are described and occasionally failure to heal is due to this cause
- Heel friction in patients confined to bed: it is essential that all patients confined to bed should have their heels elevated to avoid the friction which regularly causes heel blisters and sepsis, needing weeks or months of treatment, and sometimes requiring major amputation with consequent and very serious medicolegal implications.

Note: Most of the above are avoidable.

Foot deformities predisposing to ulceration
- Callus
- Clawed toes
- Bunions
- Pes cavus
- Hallux rigidus
- Hammer toe
- Charcot foot
- Deformities from previous trauma or surgery
- Nail deformities
- Oedema

Staging the diabetic foot

1 Normal
2 High risk
3 Ulcerated
4 Cellulitic
5 Necrotic
6 Major amputation



Heel ulcer

The neuropathic foot

Ulcers develop on the tips of the toes and on the plantar surfaces of the metatarsal heads and are often preceded by callus formation. If the callus is not removed then haemorrhage and tissue necrosis occur below the plaque of callus, leading to ulceration. Ulcers can be secondarily infected by staphylococci, streptococci, gram negative organisms, and anaerobic bacteria, which can quickly lead to cellulitis, abscess formation, and osteomyelitis. Sepsis complicating apical toe ulcers can lead to in situ thrombosis of the digital arteries, resulting in gangrene of the toe. The foot is invariably warm, with intact, often bounding pulses.

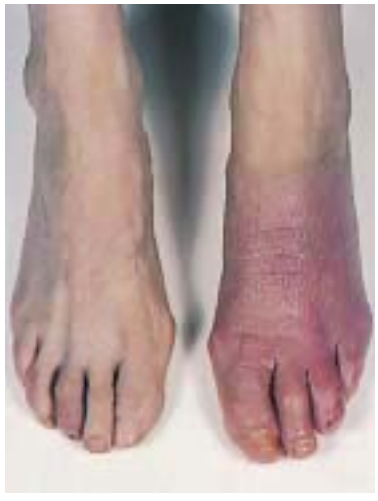
Clinical features

Neuropathic foot
- Warm with intact pulses
- Diminished sensation; callus
- Ulceration (usually on tips of toes and plantar surfaces under metatarsal heads)
- Sepsis
- Local necrosis
- Oedema
- Charcot joints

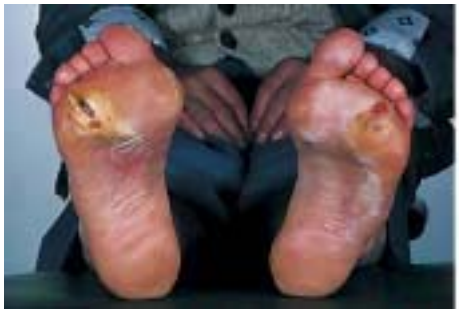
Ischaemic (neuro-ischaemic)
- Pulseless, not warm
- Usually diminished sensation
- Ulceration (often on margins of foot, tips of toes, heels)
- Sepsis
- Necrosisor gangrene
- Critical ischaemia (urgent attention) foot pink, painful, pulseless and often cold

The ischaemic (neuro-ischaemic foot)

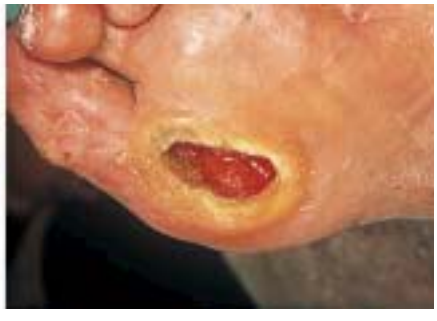
The absence of foot pulses must always alert physicians to the possible presence of ischaemia which requires specific assessment and often treatment as well. Lesions on the margins of the foot and absence of callus are characteristic features. Gangrene may be present as well. It is essential to identify critical ischaemia with its characteristic pink, painful (sometimes extreme and persistent pain during day and night) and pulseless, sometimes cold, foot. The ankle/brachial pressure index assessed by Doppler ultrasonography can give a useful guide to the presence or absence of ischaemia (see page 61).



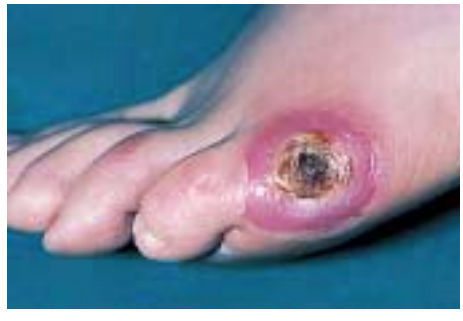
The pink, painful, ischaemic foot



Neuropathic ulcers



Neuropathic ulcer



Ischaemic ulcer

Management

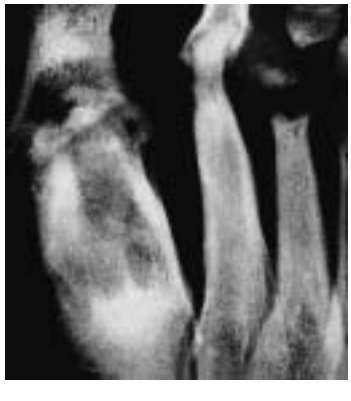
Infected diabetic foot lesions should be treated only by those with sufficient experience and facilities. General practitioners very rarely have such experience and should normally refer patients for specialist care.

Six aspects of patient treatment
- Wound control
- Microbiological control
- Mechanical control
- Vascular management
- Metabolic control
- Education

The ulcerated foot
- Arrange urgent foot ulcer care from the specialist foot care team.
- Expect the team to ensure as a minimum:
  - local wound management, appropriate dressings, and debridement as indicated
  - antibiotic treatment as appropriate
  - investigation and management of vascular insufficiency
  - specialist footwear to distribute foot pressures appropriately
  - good blood glucose control.

Treatment of diabetic foot ulcers

Management of the ulcer falls into three parts: removal of callus, eradication of infection, and reduction of weight bearing forces, often requiring bed rest with the foot elevated. Excess keratin should be pared away with a scalpel blade by the



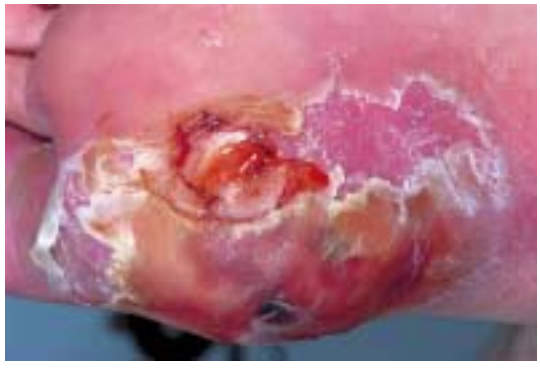
Radiograph showing osteomyelitis

podiatrist to expose the floor of the ulcer and allow efficient drainage of the lesion. A radiograph should be taken to assess the possibility of osteomyelitis whenever a deep penetrating ulcer is present, or when lesions fail to heal or continue to recur.

A bacterial swab should be taken from the floor of the ulcer after the callus has been removed; culture of excised tissue may yield even more reliable information. A superficial ulcer may be treated on an outpatient basis, and oral antibiotics prescribed according to the organisms isolated, until the ulcer has healed. The most likely organisms to infect a superficial ulcer are staphylococci, streptococci and sometimes anaerobes. Thus treatment is started with amoxycillin, flucloxacillin, and metronidazole, and adjusted when results of bacteriological culture are available. Choice and duration of antibiotic administration require considerable expertise and laboratory guidance. The patient should be instructed to carry out daily dressing of the ulcer. A simple non-adherent dressing should be applied after cleaning the ulcer with physiological saline.

Deep indolent ulcers also require local wound care and antibiotics; application of a total contact plaster cast, lightweight scotch cast boot or air cast boots which conform to the contours of the foot, thereby reducing shear forces on the plantar surface, may be used. Great care must be taken, especially with the fitting of plasters, to prevent chafing and subsequent ulcer formation elsewhere on the foot or ankle.

Any foot lesion which has not healed in one month requires further investigation and a different approach.



Gross sepsis leading to abscess formation

Danger signs: urgent treatment needed

- Redness and swelling of a foot which even when neuropathic causes some discomfort and pain; this clinical picture often indicates a developing abscess, and urgent surgery may be needed to save the leg
- Cellulitis, discolouration, and crepitus (gas in soft tissues)
- A pink, painful, pulseless foot even without gangrene indicates critical ischaemia which needs very urgent arterial investigation followed by surgical intervention whenever possible

All the above require immediate hospital admission, urgent treatment, and appropriate investigation

Urgent treatment

1 Bed rest.
2 Intravenous antibiotics. In the first 24 hours before bacteriological cultures are available it is necessary to provide a wide spectrum of antibiotic cover. Thus quadruple therapy may be necessary consisting of amoxycillin, flucloxacillin with metronidazole (to treat anaerobes), and either ceftazidine 1 g three times daily or gentamicin to treat gram negative organisms. This treatment can be adapted when results of bacteriological culture are available. The emergence of multiple resistant *Staphylococcus aureus* (MRSA) is presenting a very serious problem, firstly because it can be responsible for the ravages of sepsis, and secondly because these patients become "lepers", needing isolation while in hospital. Available treatments include intravenous vancomycin and intramuscular teicoplainin.
3 An intravenous insulin pump may be necessary to control the blood glucose.
4 Surgical debridement to drain pus and abscess cavities and to remove all necrotic and infected tissue including devitalised and infected bone resulting from osteomyelitis. Deep tissue swabs should be sent to the laboratory. If necrosis has developed in the digit, a ray amputation to remove the toe and part of its associated metatarsal is necessary and is usually very successful in the neuropathic foot with intact circulation. Skin grafting is occasionally needed and accelerates wound healing.

The ischaemic foot

Sepsis in the presence of ischaemia is a dangerous combination and should be treated urgently as described above. When ischaemia is suspected, or an ulcer does not respond to medical treatment, vascular investigation is required.

- Doppler studies to measure the pressure index (the ankle/brachial ratio of systolic blood pressure):
  - pressure index > 1·2 indicates rigid or calcified vessels or both
  - pressure index > 1 is normal (or calcified)

Radiograph of atheromatous narrowing of the femoral artery before and after balloon dilatation by angioplasty

- pressure index < 0·9 indicates ischaemia present
- pressure index < 0·6 indicates severe ischaemia.

Note: Vascular calcification is common so that spuriously high readings can be obtained. This must be taken into account when the pressure index reading is evaluated.

- Arterial imaging by techniques including duplex scanning, magnetic resonance angiography, and conventional arteriography is performed with a view to angioplasty or arterial reconstruction, or both. Infrapopliteal angioplasty or distal bypass to the tibial or peroneal vessels are now well established procedures and are important for limb salvage in the diabetic foot.
- Amputation of the toe is usually unsuccessful in the neuro-ischaemic foot (in contrast to the neuropathic foot with an intact circulation) unless the foot can be revascularised. If this is not possible, then a dry necrotic toe should be allowed to autoamputate. After attempts to control infection, below knee amputation is indicated in those with rampant progressive infection or extensive tissue destruction.

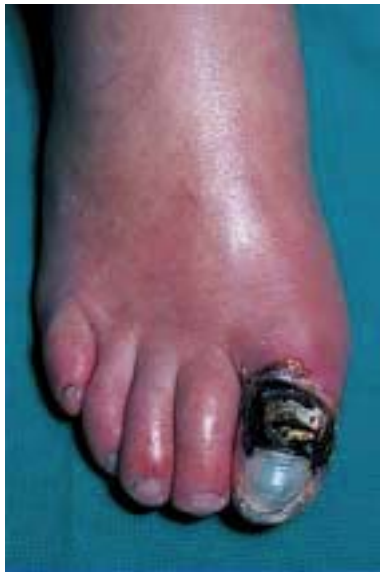
Rest pain in the severely ischaemic limb can be relieved by successful revascularisation, but if that fails, pain relief with opiates may be necessary. Paravertebral lumbar block has been disappointing in promoting healing, but occasionally rest pain is ameliorated. If all these measures fail and pain remains intractable, then below knee amputation may be needed.

Renal protection during arteriography

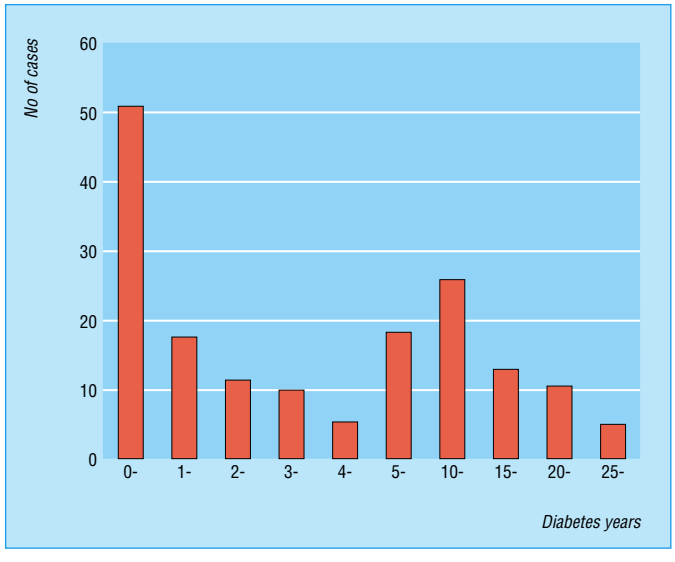
The intravenous dye used during angiography can precipitate acute oliguric renal failure in patients with early renal impairment. The following preventive measures should be taken:

- avoid dehydration
- give intravenous fluids starting four hours before the procedure
- intravenous insulin sliding scale should be used
- monitor urine output
- check creatinine before and on the day after the procedure.

Note: Metformin should be stopped 48 hours before an angiographic procedure and resumed 48 hours after the procedure has been completed.



Critical ischaemia with ischaemic gangrene of great toe



Duration of known diabetes at time of amputation

The neuropathic joint (Charcot's joint)

Loss of pain sensation together with possible rarefaction of the bones of the neuropathic foot may have serious consequences; abnormal mechanical stresses usually prevented by pain may occur, and the susceptible bones are then damaged by relatively minor trauma. Patients present with a hot swollen foot, sometimes aching, and the appearances are often mistaken for infection. Injury may have occurred days or weeks earlier, or may not even have been noticed. Sometimes Charcot changes develop after minor amputations which change the normal weight bearing stresses. Radiographs at this stage are normal, but gross damage appears and develops rapidly during the following weeks, leading to gross deformity of the foot. The destructive process does not continue indefinitely but stops after weeks or months. Bony changes are most often seen at the tarsal-metatarsal region of the foot, but they occur also at the ankle or at the metatarso-phalangeal region. Changes at other sites are rare.

Early diagnosis is essential. The initial presentation of unilateral warmth and swelling in a neuropathic foot is extremely suggestive of a developing Charcot joint. Bone scans are more sensitive indicators of new bone formation than radiography and should be used to confirm the diagnosis. It is

The neuropathic Charcot joint

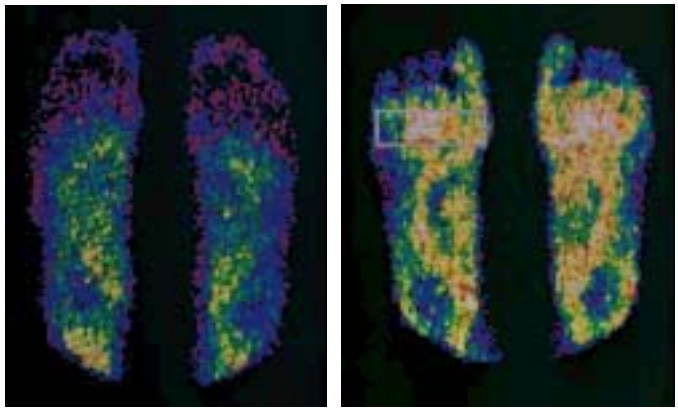
essential to exclude infection as the cause of these changes, where the differential diagnosis can be difficult: a gallium white cell scan and a magnetic resonance imaging scan can help and are appropriate and important investigations.

Management initially comprises rest, ideally bed rest or use of non-weight bearing crutches, until the oedema and local warmth have resolved. Alternatively, the foot can be immobilised in a well moulded total contact plaster which is initially non-weight bearing. Immobilisation is continued until bony repair is complete, usually in two to three months. The use of bisphosphonates in preventing bone damage from occurring in the evolution of the Charcot foot is under investigation and appears promising. In long-term management, special shoes and insoles should be fitted to accommodate deformity and prevent ulceration, which is the major hazard of the Charcot foot.

Neuropathic oedema

Neuropathic oedema consists of swelling of the feet and lower legs associated with severe peripheral neuropathy: it is uncommon. The pathogenesis may be related to vasomotor changes and arterio-venous shunting. Ephedrine 30 mg three times daily has been shown to be useful in reducing peripheral oedema by reduction of blood flow and increase of sodium excretion.

Neuropathic oedema is associated with severe peripheral neuropathy



Isotope bone scans of a normal foot (left) and a neuropathic foot (right) showing high blood flow

Long-term care after wound healing is complete

Appropriate footwear, ongoing podiatry, and regular reassessment are required.

Footwear

Redistribution of weight bearing forces on vulnerable parts of the foot can be achieved by special footwear. Moulded insoles made from substances with energy absorbing properties such as plastozote and microcellular rubber are suitable for long-term redistribution of weight bearing forces. Special shoes to accommodate the shape of the foot, and moulded insoles are often necessary. In cases of severe deformity, shoes may be constructed individually for the patient. However, in most patients, extra depth "stock" shoes will usually suffice. Failure to wear appropriate shoes is commonly a cause of foot ulceration or recurrence in those who have previously had problems.

Footwear

For everyday use, especially when on your feet for long periods

- Wear a lace-up shoe, with plenty of room for the toes, and either flat or low heeled
- Do not wear slip-ons or court shoes, except for special occasions
- Do not wear slippers at home

Examination of the foot: screening and prevention

The foot must be examined routinely at the onset of diabetes and at every annual review thereafter. The key issues in assessment are listed in order of importance.

Essential

- Patient should be aware of the need for foot care.
- Identify the critically ischaemic foot.
- Active lesions should be sought (for example, hidden lesions between the toes) and treated immediately.
- Deformities, callus, skin cracks, and discoloration need to be detected and managed.
- A simple sensory test should be performed for example, a monofilament sensory test under the great toe (inability to detect 10 g or more indicates risk of foot ulceration).
- Examine the pulses (dorsalis pedis and posterior tibial).

Care and preventive measures

- Active lesions should be treated immediately
- Written advice and education on foot care should be provided
- Advice is needed on appropriate shoes to accommodate foot deformities
- Regular podiatry to remove excess callus and provide nail care is essential

Other assesments
- Assess ankle reflex.
- Assess other sensory modalities for example pinprick, vibration perception at the medial malleolus or tip of the great toe.

Advice and education must follow the examination.

Guidelines for foot care

Low current risk foot
(normal sensation, palpable pulses)

- Individual foot care education.

At risk foot
(neuropathy, absent pulses, or other risk factor described above)

- Enhance foot care education.
- Inspect feet every three to six months.
- Advise on appropriate footwear.
- Review need for vascular assessment.
- If previous ulcer, deformity or skin changes manage as high risk.

High risk foot
(Ischaemia deformity, skin changes, or previous ulcer)

- Arrange frequent review (one to three monthly) from foot care team.
- At each review, evaluate:
  intensified foot care education
  specialist footwear and insoles
  skin and nail care according to need.
- Ensure special arrangements for people with disabilities or immobility.

The ulcerated foot
The care pathway is described on page 60.

Conclusions

Many foot problems can be prevented, and all diabetic patients should be aware of the potential problem of foot damage. Every patient should be issued with information containing straightforward safety instructions.

A good podiatrist must be available for diabetic patients. Ill-fitting shoes are the cause of many problems. New shoes should always be broken in by wearing them initially for only short periods. If the foot is in any way misshapen, for example, from bunions, hammer toes, Charcot deformities or as a result of surgery, shoes must be specially made to fit. It is a great advantage if a shoe fitter attends the chiropody clinic; it is possible to make simple shoes fit on the spot (Dru shoes for example) while awaiting delivery of more elaborate fitted shoes made in a workshop.

Close liaison between the podiatrist, orthotist, nurse, physician, and surgeon is vital in the care of the diabetic foot. The diabetic foot clinic is the optimum forum for provision of intensive podiatry, close surveillance and prompt treatment of foot infection, and the provision of specially constructed shoes.

Care of your feet

DO
- Wash feet daily with mild soap and warm water
- Check feet daily
- Seek urgent treatment for any problems
- See a podiatrist regularly
- Wear sensible shoes

DO NOT
- Use corn cures
- Use hot water bottles
- Walk barefoot
- Cut corns or callosities
- Treat foot problems yourself

Danger signs
- Check your feet every morning
- Come to the clinic *immediately* if you notice
  Swelling
  Colour change of a nail, toe, or part of a foot
  Pain or throbbing
  Breaks in the skin, including cracks, blisters, or sores

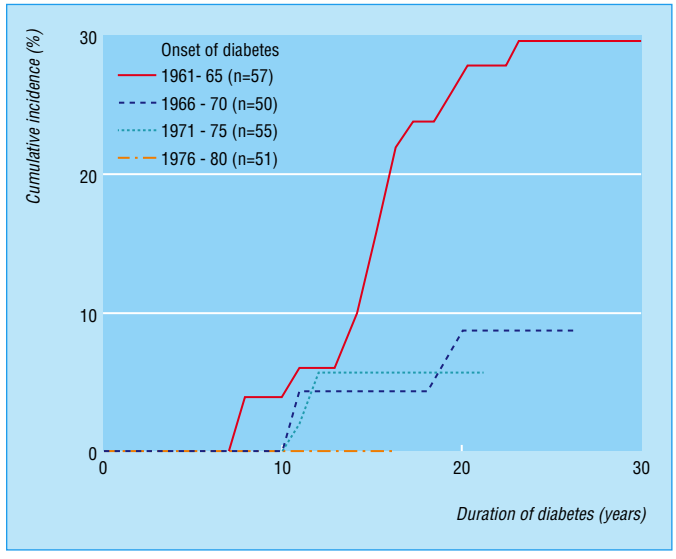
The Dru shoe

The histogram showing duration of diabetes at time of amputation adapted from Malins, J. *Clinical diabetes mellitus.* London: Eyre and Spottiswoode, 1968.

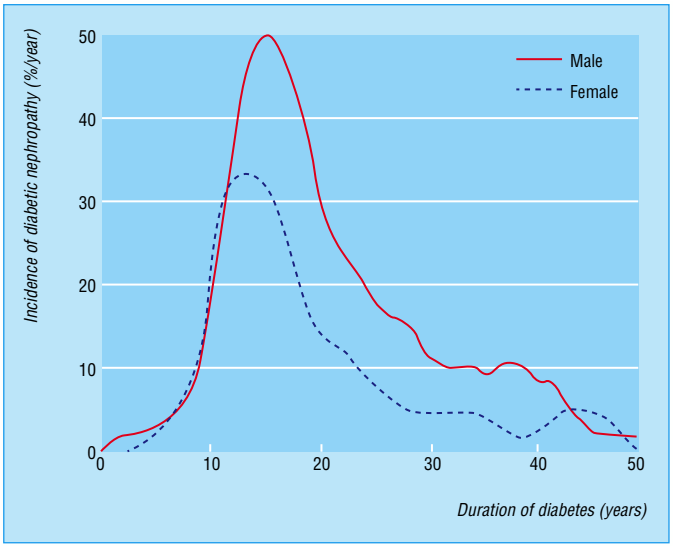
16 Diabetic nephropathy

The development of proteinuria in any diabetic patient is ominous. It is associated with a risk of severe retinopathy and neuropathy, and above all carries a major increased risk in mortality from coronary artery disease, as well as progression to renal failure in some patients. Yet developments in this field to improve the prognosis have been substantial. The overall prevalence of proteinuria in Type 1 diabetes has decreased over half a century from more than 50% of patients down to between 10 and 20%, presumably as a result of better overall diabetic care. Furthermore, at the earliest sign of proteinuria, administration of medication and very tight blood pressure control ameliorate the course of the disease and substantially delay the development of renal failure. And for those who are less fortunate, transplantation and dialysis restore a good quality of life to the majority.

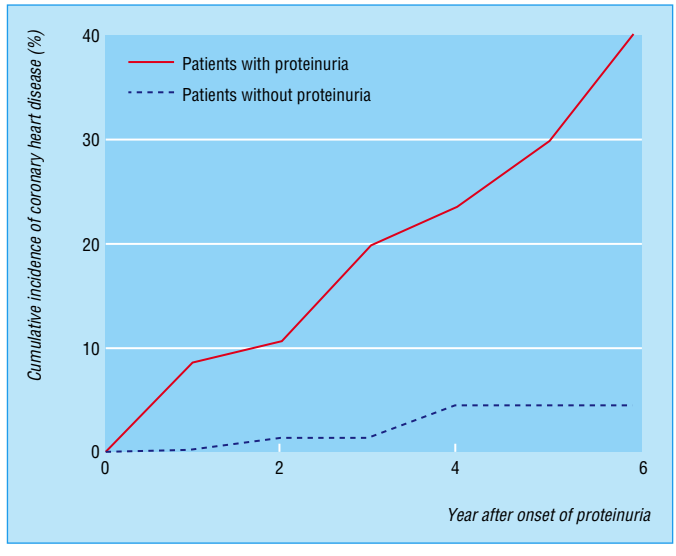
Proteinuria occurs in both Type 1 diabetes and Type 2 diabetes. African-Caribbean and Asian Type 2 diabetic patients have a much higher prevalence of this disease and its associated morbidity.



Cumulative incidence of nephropathy in patients diagnosed with Type 1 diabetes over a 20-year span



Incidence of diabetic nephropathy in Type 1 diabetic patients



Cumulative incidence of coronary heart disease in Type 1 diabetic patients with and without proteinuria

Onset of nephropathy

The earliest detectable development in Type 1 diabetes of small amounts of proteinuria (microalbuminuria) occurs more than five years after diagnosis. Both micro and gross albuminuria are detected at diagnosis in many with Type 2 diabetes, both because diabetes may already be long standing and in some instances because of established hypertension or other renal diseases.

Initiating factors and progression promoters in diabetic nephropathy

Factors leading to initiation and those determining rate of progression of diabetic renal disease are gaining increasing recognition.

Both development and early progression of diabetic nephropathy are most likely to occur following years of poor

Nephropathy—initiating factors and progression promoters

Initiating factors
- Persistently poor diabetic control
- Hypertension in Type 2 diabetes
- Genetic factors

Progression promoters
- Blood pressure
- Proteinuria
- Persistently poor diabetic control
- Dyslipidaemia
- Genetic factors
- Smaller kidneys (or glomeruli)
- Smoking
- High dietary protein

diabetic control; tight control over a decade both delays onset of the disease and slows progression chiefly of its early phase, although there is some effect in established disease as well. There is also a familial propensity to nephropathy in both Type 1 and Type 2 diabetes, although the precise genetic factors responsible have not been identified.

The most powerful factor in slowing progression of the disease is the skilful and timely use of angiotensin converting enzyme (ACE) inhibitors and angiotensin receptor blockers, together with the strict management of hypertension when it occurs

Stages of nephropathy

Early physiological changes

At the onset of Type 1 diabetes there is evidence of hyperfiltration with increased glomerular filtration rate, and large kidneys and glomeruli. These defects can be reversed by meticulous diabetic control; there is however no good evidence that these changes predispose to the subsequent development of nephropathy.

Stages of progression of diabetic nephropathy

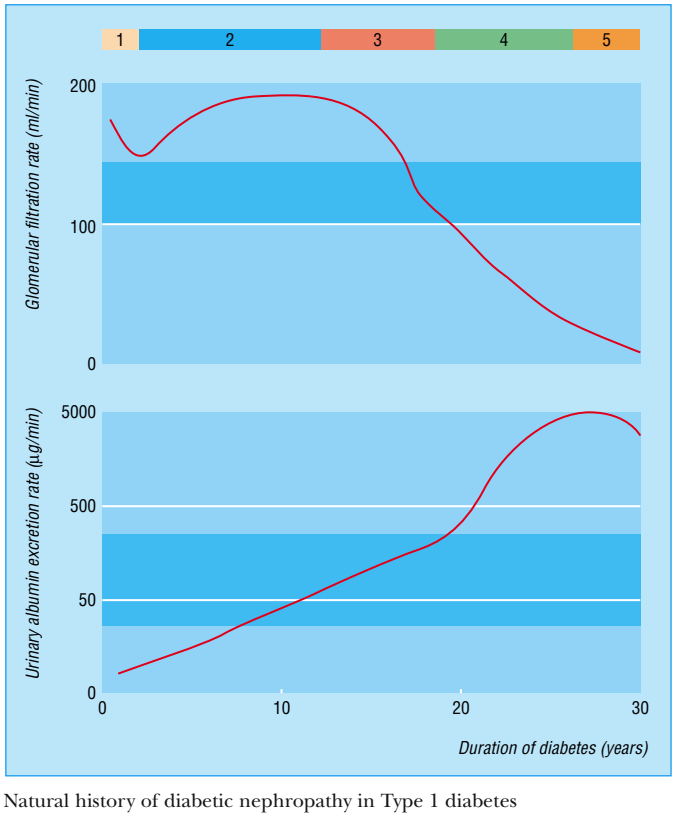
	Normal (I)	Incipient (II)	Persistent (III)	Clinical (IV)	End stage (V)
Albuminuria (mg/24 h)	<20	20–300 (microalbuminuria)	⩾300 (up to 15 g/day)	⩾300 (up to 15 g/day)	⩾300 (can fall)
Glomerular filtration rate (ml/min)	High/normal Hyperfiltration	Normal/high	Normal or decreased	Decreased	Greatly decreased
Serum creatinine (μmol/l)	Normal 60-100	Normal 60-120	High normal 80-120	High 120-400	Very high >400
Blood pressure (mm Hg)	Normal	Slightly increased	Increased	Increased	Increased
Clinical signs	None	None	Anaemia ± oedema, increased blood pressure, may be none	Anaemia ± oedema, increased blood pressure, may be none	Anaemia oedema, increased blood pressure, uraemic symptoms

Course of nephropathy

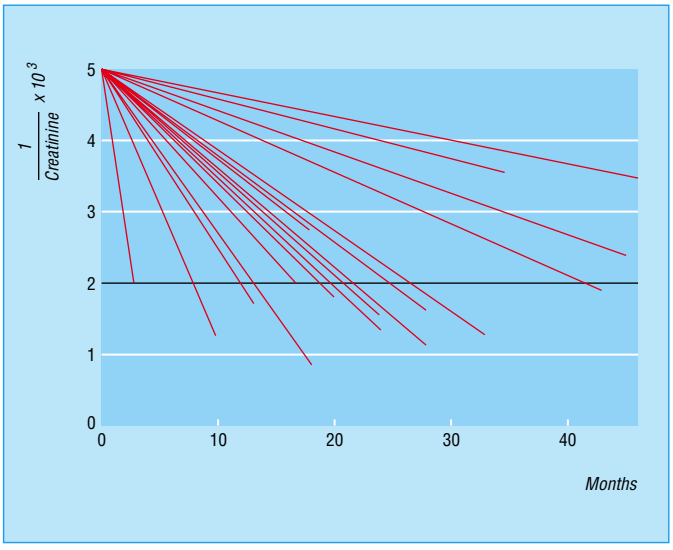
The natural course of diabetic nephropathy is the progression through five stages from normal renal function to end-stage renal failure as shown in the table. Incipient nephropathy, identified by the appearance of microalbuminuria, is the earliest clinical stage and is not associated with significant clinical signs or any changes other than a very small increase in blood pressure. At this stage urine testing by conventional methods will give negative or only trace positive results.

As the disease progresses albuminuria increases until end-stage nephropathy, when it may decrease; there are wide variations in the amount of protein excretion. The glomerular filtration rate shows a progressive decline, which varies considerably between patients and is usefully assessed by calculating the inverse creatinine value (1/serum creatinine concentration) which can be plotted against time and is generally linear. Blood pressure rises progressively. The nephropathy is commonly asymptomatic until it is advanced, when oedema and breathlessness develop. Anaemia often occurs relatively early in the course of the disease before renal failure is established, and much sooner than in non-diabetic renal disease, as a result of diminished erythropoietin production, which can be corrected by injections of erythropoietin, sometimes with considerable clinical benefit.

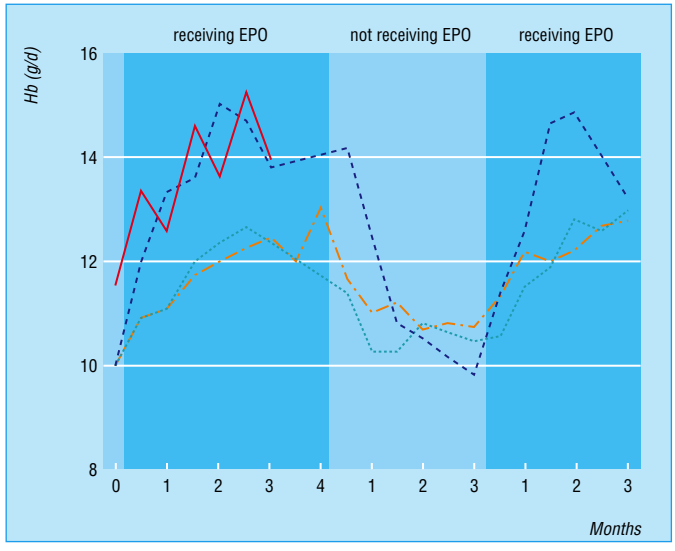
Hyperlipidaemia is common in nephropathy, as are other risk factors for vascular disease, including changes in the concentrations of fibrinogen and other clotting factors.



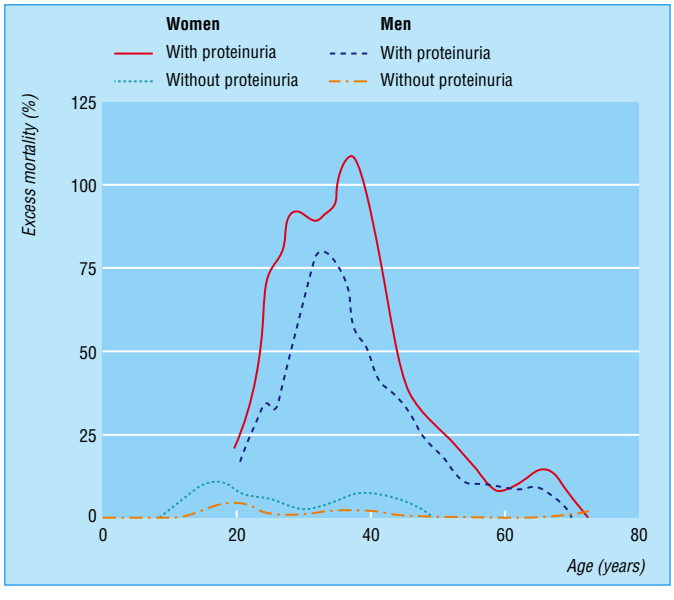
Natural history of diabetic nephropathy in Type 1 diabetes



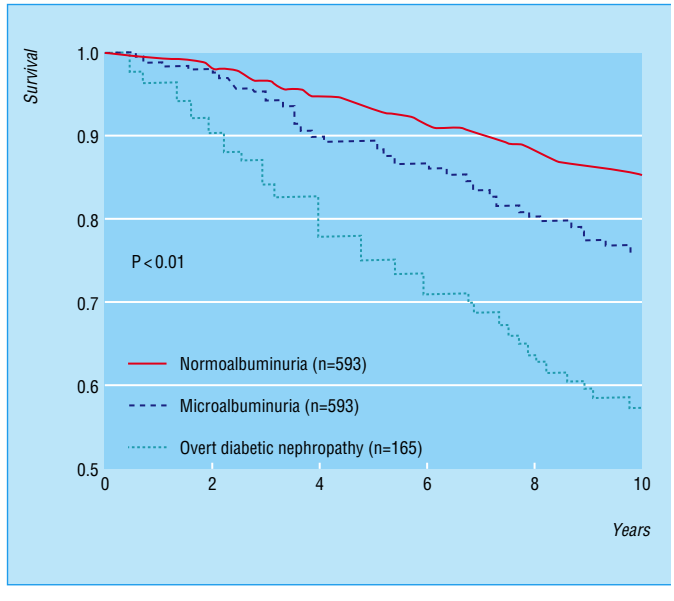
Decline of renal function in 16 Type 1 diabetic patients with nephropathy



Erythropoietin (EPO) deficient anaemia in early diabetic nephropathy treated successfully with EPO injections



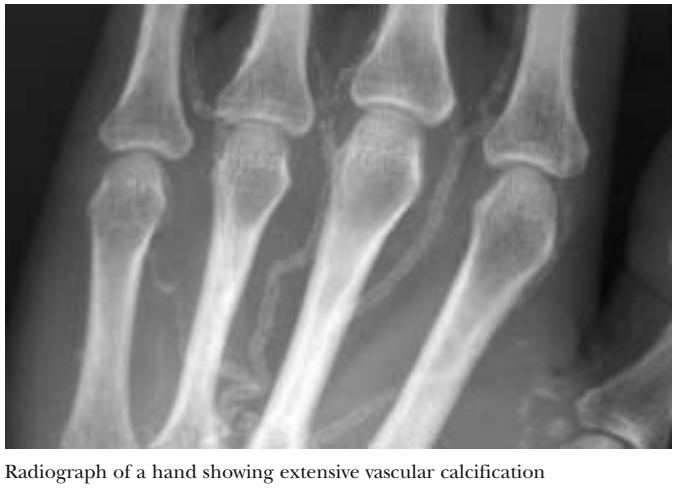
Relative mortality of Type 1 diabetic patients with and without persistent proteinuria



Survival curves of Type 1 diabetic patients with respect to all cause mortality in relation to presence or absence of proteinuria

Accompanying problems

Patients with diabetic nephropathy normally have many other complications, the problems increasing as the stage of nephropathy advances. Almost all have retinopathy, often proliferative, and they have a tenfold increased risk of blindness compared with patients without proteinuria. A very high proportion have coronary artery disease, with an excess risk of death many times higher than in patients without proteinuria. They are also more at risk of peripheral vascular disease (often with vascular calcification, peripheral and autonomic neuropathy, foot ulceration, and amputation.



Radiograph of a hand showing extensive vascular calcification

Primary prevention

Poor diabetic control predisposes to the development of nephropathy, although it is not the only factor, and genetic components also predispose to this disease. Tight control of diabetes early in its course helps to delay development of nephropathy (see chapter 10). Many patients will never acquire proteinuria even after several decades. Blood pressure management at all stages is crucial (see below).

Early detection and diagnosis

Testing urine samples for the presence of protein or microalbuminuria should be routine practice at every clinic visit and is obligatory at the annual review. If proteinuria develops it is important to distinguish the onset of nephropathy from other causes of renal disease.

Microalbuminuria is detected by measurement of the albumin/creatinine ratio or urinary albumin concentration. The test is best performed on the first morning urine sample. If albuminuria is discovered for the first time, confirmation is required within one month if proteinuria is present, and if microalbuminuria is detected it should be confirmed twice within the ensuing months. All laboratory and near patient commercial tests specifically designed for microalbuminuria have satisfactory sensitivity (>80%) and specificity (>90%).

If proteinuria has evolved gradually over several years in the presence of retinopathy, and there are no unusual features such as haematuria, unequal size kidneys, or a history of urinary tract complaints, then extensive investigation is not necessary. In Type 2 diabetes, however, there is a greater chance of non-diabetic renal disease being present, and renal biopsy may be needed, especially if there are atypical features. In both Type 1 diabetes and Type 2 diabetes the absence of retinopathy should make one suspect other causes of renal failure. Indeed, rapid onset of proteinuria in any patient is never due to diabetes and should always be fully investigated, including a biopsy.

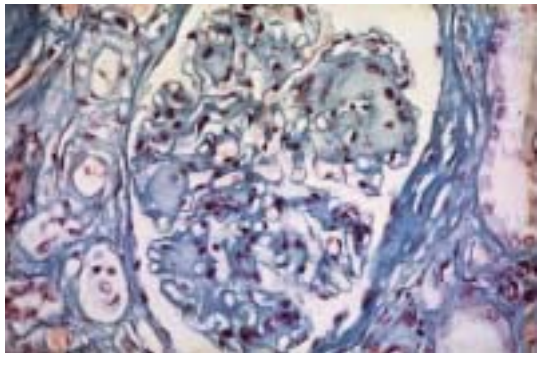
The characteristic pathological lesion is diabetic glomerulosclerosis, including basement membrane thickening, mesangial expansion, and in the later stages, the classical Kimmelstiel-Wilson nodules, together with hyalinisation of efferent and afferent arterioles. Mesangial expansion, which requires an experienced pathologist to interpret, correlates well with renal function.

Diagnostic categories

Microalbuminuria
- Albumin creatinine ratio >2·5 mg/mmol/l (men) >3·5 mg/mmol/l (women)
- Urinary albumin concentration >20 mg/l
- Urinary albumin excretion rate 30-300 mg/24 h, or 20-200 μg/min in an overnight specimen

Proteinuria
- Urinary albumin excretion rate >300 mg/24 h
- Urinary protein excretion >500 mg/24 h (albumin creatinine ration >30 mg/min; albumin concentration >200 mg/l)

Investigations in a patient with proteinuria
- Midstream urine
- 24 hour urine protein
- Renal ultrasonography
- Blood count
- Erythrocyte sedimentation rate
- Antinuclear factor
- Serum complement
- Serum lipids
- (Renal biopsy only as indicated)



Diabetic glomerulosclerosis showing Kimmelstiel-Wilson nodules

Management

- Tight diabetic control delays the onset and slows the progression chiefly in the early phase of the disease, with a smaller effect in fully established disease.
- Established microalbuminuria (—that is three positive samples) is managed with ACE inhibitors:
  (a) in Type 1 diabetes treatment is commenced regardless of blood pressure which should in any case be maintained < 130/80 mm Hg, or 125/75 mm Hg, in younger patients
  (b) in Type 2 diabetes treatment aims to maintain a blood pressure < 130/80 mm Hg.

  ACE inhibitors and angiotensin receptor blockers always reduce microalbuminuria and delay the onset of established proteinuria.
- Established proteinuria (proteinuria > 500 mg/24 h or albuminuria > 300 mg/24 h).

This phase is always managed by vigorous hypotensive treatment now aiming where possible for a blood pressure of <130/80 mm Hg or even less (125/75 mm Hg) in younger patients. This both reduces (or occasionally eliminates) proteinuria and slows progression of the declining glomerular filtration rate, best demonstrated in Type 1 diabetes, and is also protective in some aspects of cardiovascular disease (see below).

- Dyslipidaemia. Detection and management of hyperlipidaemias is important because of the huge increase in cardiovascular disease in those with nephropathy. The idea, however, that reducing lipidaemia slows the progression of the renal disease is unproven.
- Restricting protein intake has a small beneficial effect on the decline in glomerular filtration rate but unless the intake is very high (greater than 1 g/kg/day) restriction is not normally advised.
- Oral hypoglycaemic treatment in Type 2 diabetes must be reviewed: metformin should not be used when renal function is impaired because of the danger of lactic acidosis; also long acting, renally excreted sulphonylureas, notably glibenclamide, should not be used because of the risk of accumulation and hypoglycaemia. Drugs such as gliclazide which are mainly metabolised are preferable (see page 15).
- Anaemia often occurs relatively early in the course of the disease before renal failure is established, and much sooner than in non-diabetic renal disease, as a result of diminished erythropoietin production, which can be corrected by injections of erythropoietin, sometimes with considerable clinical benefit.
- Smoking should be strongly discouraged.
- Regular reassessment and management of all the other diabetic complications to which diabetic nephropathy patients are especially prone is extremely important in those who develop this disease.

Treatment of nephropathy

- Optimal diabetic control
- Antihypertensive treatment and diuretics
- Reduce hyperlipidaemia
- Stop smoking

Choice of antihypertensive agent

ACE inhibitors or angiotensin receptor blockers are the agents of choice for diabetic nephropathy patients at any stage. They probably provide benefits over and above their blood pressure lowering effects in comparison with other agents. However, the overriding need is to maintain good blood pressure control in all those with nephropathy using any hypotensive agent to suit the individual patient, and indeed for Type 2 diabetes, United Kingdom Prospective Diabetes Study (UKPDS) demonstrated no drug preferential benefit. Combinations of drugs are almost always required, especially in overweight Type 2 diabetic individuals when more than two drugs are often needed: the ideal blood pressure of <130/80 mm Hg may be difficult to achieve and some compromise following informed discussion with patients may be necessary. Tight blood pressure control may induce or exacerbate postural hypotension, which should be avoided.

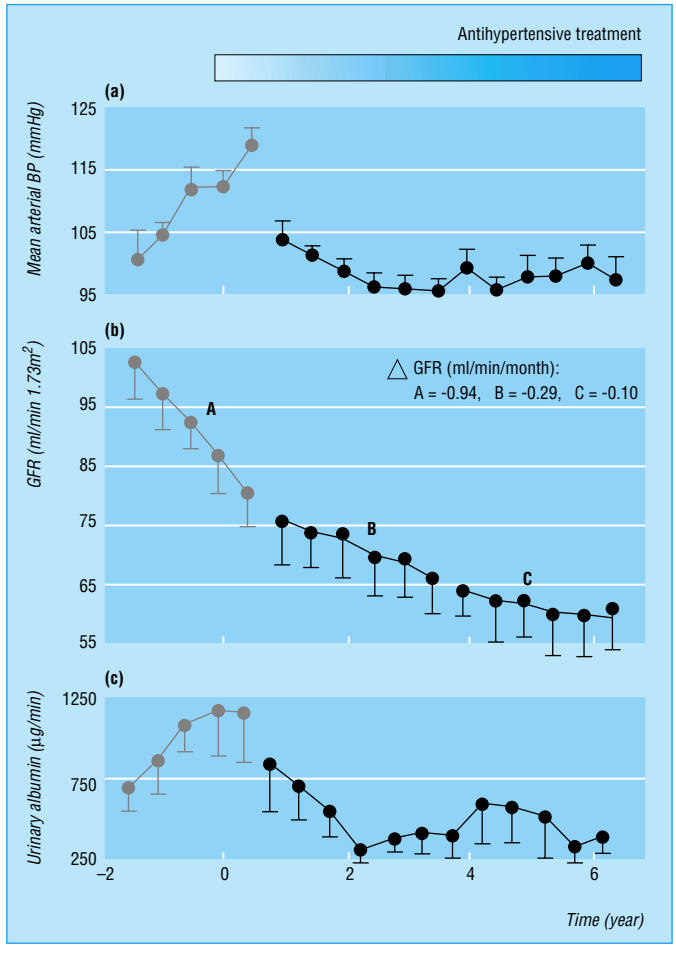
For microalbuminuria

Even in normotensive patients, aim to titrate medication to reduce or eliminate microalbuminuria. Doses often needed are:

- ACE inhibitors*: enalapril 10 mg twice daily
  lisinopril 20 mg daily
  ramipril 2·5 mg daily.
- Angiotensin 2 receptor antagonists*†:
  irbesartan 300 mg daily
  losartan 50 mg daily.

*Only some of the longest established and therefore most studied agents are mentioned here by name. Many others are available and can be found in the *BNF*.

†Angiotensin 2 antagonists do not cause the cough often induced by ACE inhibitors, but are more expensive than the latter.



Antihypertensive treatment reduces the decline of glomerular filtration rate (GFR) in patients with Type 1 diabetes

For hypertension
ACE inhibitors (or angiotensin 2 antagonists) to which may be added:

- calcium channel blockers (amlodipine)
- diuretics
- β blockers
- others.

For African-Caribbeans a calcium channel blocker such as verapamil may be a first choice hypotensive agent.

Creatinine should always be checked within one to two weeks of starting treatment with a ACE inhibitor or angiotensin 2 antagonist in case a rapid rise occurs in patients with renal artery stenosis. In this case, the patient should be changed to another agent, and appropriate investigations undertaken as well. For further information on treatment of hypertension, specific indications and contraindications to individual drugs, see chapter 17.

Patient counselling
Considerable fear is generated when patients are told that they have "kidney damage", with visions of dialysis and transplantation. This is needlessly and carelessly damaging to patient well being; the proper context of the disease must be presented by physicians and nurse educators.

Renal failure
Patients with end-stage renal failure need much time and consideration. Renal support therapy is now available for most of these patients, who should be assessed by a nephrologist once the serum creatinine exceeds 150 μmol/l. Transplantation or dialysis is often necessary at a lower serum creatinine concentration than in non-diabetic people, often around 450 to 550 μmol/l. Kidney transplantation is the treatment of choice if the patient has good cardiac function, and cardiac and peripheral vascular assessments are essential before treatment becomes necessary. Results of renal transplantation are shown in the table. If cardiac function is severely compromised them continuous ambulatory peritoneal dialysis (CAPD) is preferable as long-term treatment. Long-term haemodialysis is used chiefly for those who cannot cope with CAPD for reasons of blindness, lack of manual dexterity or intellectual decline. One year after starting CAPD more than 70% of patients are alive, and even after five years approximately 40% survive.

Pancreas transplantation
Combined kidney and pancreas transplantation is undertaken at some major centres and aims to eliminate diabetes and with it the need for insulin injections. Patients who are liberated from decades of restrictions and self discipline imposed by diabetes experience a joy and happiness rarely witnessed in medical practice. This is the prime reason for offering combined transplantation. Other benefits derive from slowing the redevelopment of glomerulosclerosis in the transplanted kidney, and halting the progression of neuropathy. Retinopathy which by this stage has almost always been laser treated is unaffected. The future potential for islet cell transplantation is briefly described on page 22.

Management of diabetes during transplantation and dialysis

At transplantation
Continuous intravenous insulin infusion is always used; soluble insulin diluted in physiological saline at 1 unit/ml. Infusion

Counselling for patients with nephropathy
Physicians and nurse educators should explain that:

- Not all of those (approximately 20%) with microalbuminuria will go on to develop established nephropathy
- Many of those with microalbuminuria or even established nephropathy will not progress to end-stage renal failure
- Modern treatment properly taken is highly effective in retarding disease progression

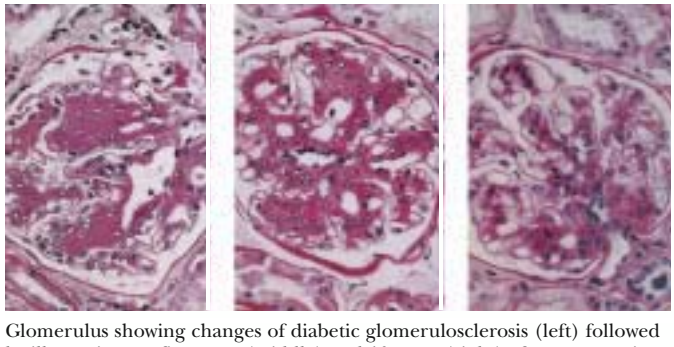
Approximate expected survival after cadaver renal transplantation

	1 year	5 years
Patient survival (diabetic/non-diabetic)	90/95%	80/88%
Graft survival* (diabetic/non-diabetic)	75/85%	55/75%

*Graft survival includes death with a functioning graft

Approximate expected survival rates after pancreas and renal transplantation

	1 year	5 years
Patient survival (diabetic)	95%	80%
Functioning pancreas graft survival	85%	60%
Functioning kidney graft survival	90%	70%



Glomerulus showing changes of diabetic glomerulosclerosis (left) followed by illustrations at five years (middle) and 10 years (right) after pancreatic transplant, showing considerable resolution after 10 years

rates vary considerably, usually in the range 2 to 20 units/h; during high dose steroid treatment the higher infusion rates are often needed. Intravenous insulin infusion is continued until drips have been taken down and the patient can eat: the insulin regimen to be followed thereafter is described in chapter 9. The daily dose is started about 20% above the pre-transplant dose and adjustments thereafter are made by trial and error.

During peritoneal dialysis

With solutions of low glucose content (1·36% glucose) no adjustment to the normal insulin regimen is needed. Dialysates of high glucose content can however severely disrupt diabetes and additional insulin is needed. During CAPD insulin can be administered entirely from the dialysis bags: soluble insulin is added to each bag, initially using the existing total daily dose in divided amounts, often giving less at night. The dose may eventually be quite different from that given subcutaneously. The technique is often satisfactory and excellent control can be achieved without hypoglycaemia. Patients whose technique is poor, and who are thus liable to peritonitis, should not be given intra-peritoneal insulin. The use of dialysis fluids with high glucose content (3·86%), needed to alleviate fluid overload, causes havoc with diabetes control; additional soluble insulin (in a dose appropriately titrated) must be given before administration of the strong glucose concentrations, or possibly more effectively, additional insulin given intra-peritoneally.

Rejection

High doses of steroids always upset diabetes within a few hours. This problem may be anticipated by increasing the first insulin dose after steroids have been given. Intravenous insulin infusion (about 4 units/h) for a few hours as a supplement to the normal subcutaneous insulin is almost always needed until the administration of methylprednisolone has been completed.

Non-diabetic renal disease

Some patients, especially those with Type 2 diabetes, develop unrelated non-diabetic renal disorders. Clues to their presence have been presented above, but renal artery stenosis should be particularly mentioned. This is most common in Type 2 diabetic patients with hypertension and peripheral vascular disease. Full renal assessment including ultrasound examination is essential, although renal arteriography is often required to confirm or exclude the diagnosis. The distinction is vital as ACE inhibitors may provoke considerable deterioration in renal function or even precipitate renal failure.

Urinary tract infections

Urinary tract infections occur in diabetic people with the same frequency as in those without diabetes, but they are sometimes exceptionally severe and may cause the renal papillae to slough, causing necrotising papillitis, or rarely emphysematous cystitis with air in the bladder wall. Infection is particularly troublesome in the rare patient with urinary retention from neurogenic bladder. Diabetic control is easily disturbed by urinary infection, as with any infection, and must be regained quickly, with insulin if necessary, while the infection is treated with antibiotics.

Pyelonephritis with septicaemia is not uncommon in diabetes, with occasional formation of a perinephric abscess. The source of the infection may not be immediately apparent and sometimes patients present in profound shock without an obvious site of infection.

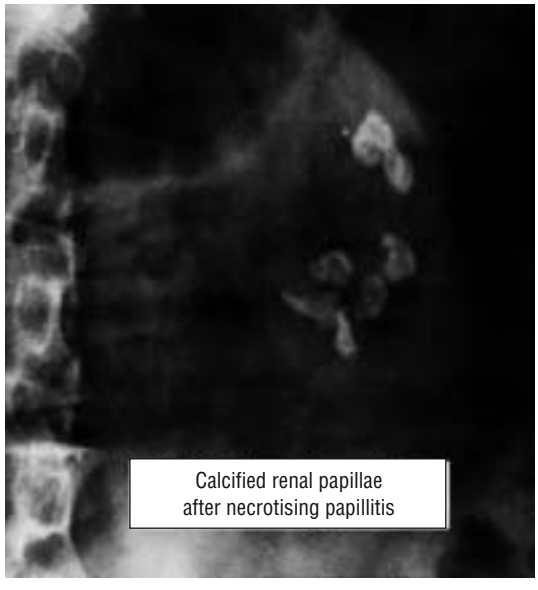
Renal protection during arteriography

The intravenous dye used during angiography can precipitate acute oliguric renal failure in patients with early renal impairment. The following preventive measures should be taken:

- Avoid dehydration
- Give intravenous fluids starting four hours before the procedure
- Intravenous insulin sliding scale should be used
- Monitor urine output
- Check creatinine before and on the day after the procedure

Note: Metformin should be stopped 48 hours before an angiographic procedure and resumed 48 hours after the procedure has been completed

Glomerulonephritis and other renal disorders can occur in diabetic patients and require renal biopsy for diagnosis and treatment in their own right



Radiograph showing calcified renal papillae after necrotising papillitis

The figure showing incidence of diabetic nephropathy is adapted from Anderson, AR. *Diabetologia* 1983;25:496–501. Cumulative incidence of nephropathy in patients diagnosed with diabetes over a 20-year span is adapted from Bojestig M, et al. *New Engl J Med* 1994;330:15-18. The cumulative incidence of coronary heart disease is adapted from Jensen T, et al. *Diabetologia* 1987;30:114-18. The graph showing treatment with erythropoietin is adapted from Watkins, P. *Diab Med* 1999;16:1-7. The figure showing relative mortality of diabetic patients with and without persistent proteinuria is adapted from Borck-Johnsson K, et al. *Diabetologia* 1985;28:590-6. The figure showing survival curves with respect to all cause mortality is adapted from Rossing P, et al. *BMJ* 1996;313:779-84. The figure showing antihypertensive treatment reduces the decline of glomerular filtration rate in patients with Type 1 diabetes is adapted from Parving H-H, et al. *BMJ* 1987;294:1443-7. The slide of glomerulus showing changes of diabetic glomerulosclerosis is reproduced from Fioretto P, et al. *New Engl J Med* 1998;339:69-75. Copyright Massachussetts Medical Society.

17 Cardiovascular disease, hypertension, lipids, and myocardial infarction

Diabetic patients, particularly those with Type 2 diabetes and those with proteinuria, are at very considerable risk of excessive morbidity and mortality from cardiovascular, cerebrovascular and peripheral vascular disease leading to myocardial infarction (MI), strokes and amputations. Much effort must be given to reducing as far as possible the risk factors which are known to predispose to major atheromatous arterial disease. Many effective measures can now be taken, adding considerably to the complexity of treating diabetic patients, especially those with Type 2 diabetes. The difficulties and dangers of polypharmacy are discussed on page 14.

> ***The glycaemic disturbance (of Type 2 diabetes) may be mild, but the rest of the disease is not***
> **George Alberti, 1988**

Coronary artery disease

Cardiovascular disease is substantially increased in diabetes, hyperglycaemia representing an independent risk factor. It is the chief cause of death and this observation strongly influences the management of diabetes by the important focus on reducing the risk factors responsible.

Diabetes more than doubles the risk of cardiovascular disease. In the United Kingdom, 35% of deaths are attributable to cardiovascular causes, compared with about 60% in those with Type 2 diabetes, and 67% of Type 1 diabetic patients alive after 40 years of age. The relative risk is greater for women than for men, so that the sex ratio is equal in those with diabetes, with a loss of the usual male predominance. The development of MI over a period of seven years in middle-aged diabetic patients without known pre-existing coronary heart disease (CHD) is the same as that in non-diabetics with existing CHD. The presence of proteinuria and even microalbuminuria is associated with a particularly large risk of CHD and a high mortality from MI
(see page 65).

Those at especially high risk of developing CHD include

- Smokers
- Hypertensives
- Those with insulin resistance associated with obesity
- Patients of Asian origin
- Those with microalbuminuria
- Those with diabetic nephropathy (macroalbuminuria)
- Those with poor glycaemic control (16% increased risk of MI for every 1% increase in HbA_{1c})
- Those with hyperlipidaemic states

Prevention of cardiovascular disease: effects of lowering blood pressure

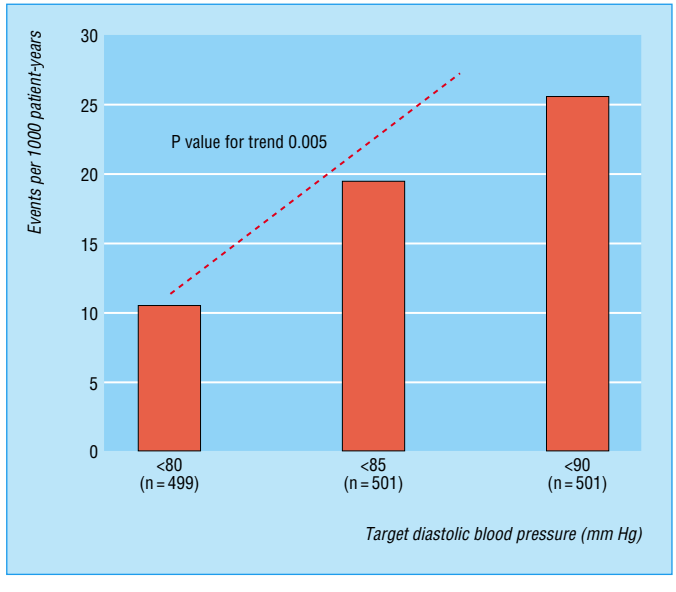
Those with a high risk of cardiovascular disease stand to gain proportionately greater benefit by reduction of risk factors. Two recently published major trials demonstrated exactly what can be achieved.

The United Kingdom Prospective Diabetes Study (UKPDS; published in 1998, see page 42): tight *v* less tight blood pressure control (mean 144/82 mm Hg *v* 154/87 mm Hg). Benefits:

- heart failure reduced by 56%
- strokes reduced by 44%
- combined MI, sudden death, stroke, peripheral vascular disease reduced by 34% (MI alone was reduced non-significantly by 16%).

There were also considerable benefits on the development of retinopathy and proteinuria (see page 43).

Heart Outcomes Prevention Evaluation (HOPE) and MicroHOPE study: this study over 4·5 years comprised 9297 patients overall and included 3577 diabetic patients (98% with Type 2 diabetes). Patients with diabetes and one other risk factor for cardiovascular disease were randomly treated with the



Rates of serious cardiovascular events according to target diastolic blood pressure in 1500 patients with hypertension and Type 2 diabetes

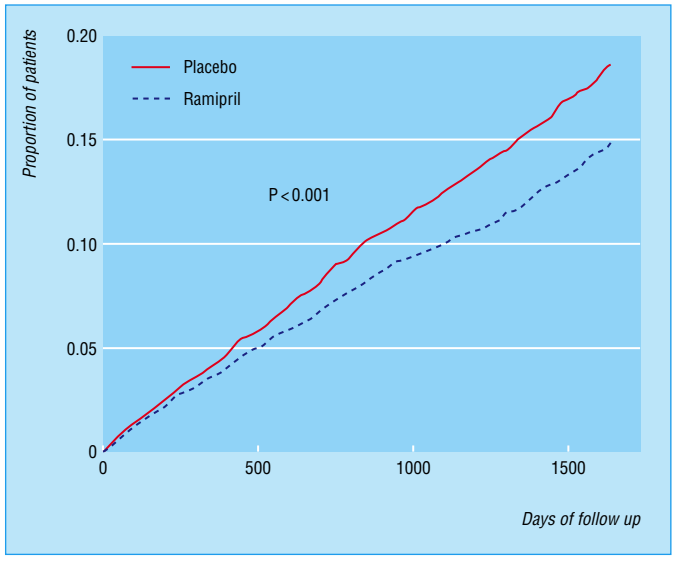
angiotensin converting enzyme (ACE) inhibitor ramipril 10 mg daily or placebo. Systolic blood pressure decreased by approximately 2 to 3 mm Hg and yielded a reduction of combined MI, strokes, and deaths from cardiovascular diseases of 25%.

The demonstrated benefits included:

- MI relative risk reduced by 22%
- stroke relative risk reduced by 33%
- cardiovascular death relative risk reduced by 37%.

It was concluded that ACE inhibitors were the first line treatment for blood pressure control in diabetes. Despite some caveats relating to the handling of the study, this treatment should probably be extended to normotensive patients with high cardiovascular risks.

There have been several other published multicentre trials: the results of the HOPE and other studies are well summarised by Bilous R, HOPE and other recent trials of antihypertensive therapy in Type 2 diabetes; in Amiel S ed. *Horizons in Medicine*, Royal College of Physicians of London, 2002.



Kaplan-Meier estimates of the composite outcome of myocardial infarction, stroke, or death from cardiovascular causes in the ramipril group and the placebo group in the HOPE study

Blood pressure management

Blood pressure should be checked annually or more frequently as indicated. Borderline readings should be re-checked several times before the decision to use medication is taken. Before using pharmacological agents, the measures shown in the box all have some effect on reducing blood pressure:

Blood pressure > 160/100 mm Hg should always be treated in those with and without diabetes aiming for a level of <140/80 mm Hg (audit standard <140/90 mm Hg).

Blood pressure 140-159/90-99 mm Hg should be treated in diabetic patients, most aggressively in those with cardiovascular risk factors, especially if there is evidence of end organ damage; aim for <140/80 mm Hg.

Blood pressure > 140/80 mm Hg should be treated if there is evidence of target organ damage or if the 10-year CHD risk exceeds 15%, aiming for a level <140/80 mm Hg. Blood pressure >130/80 mm Hg should be treated in those with microalbuminuria or macroalbuminuria (see page 68).

Factors that affect blood pressure

- Salt restriction
- Weight reduction or exercise programmes
- Reduction of excessive alcohol intake

First and foremost, the key objective when choosing antihypertensive drugs is to lower blood pressure by any means because most of the benefits relate to the blood pressure achieved rather than the drug used

Choice of antihypertensive drugs

The key objective is to lower blood pressure by any means because most of the benefits relate to the blood pressure achieved rather than the drug used. There are however additional advantages using ACE inhibitors or angiotensin 2 receptor antagonists (see page 68). Indeed, results of the recent losartan intervention for end point reduction (LIFE) study showed additional benefits resulting from the use of the angiotensin 2 receptor antagonist losartan when compared with β blocker atenolol, despite comparable blood pressure reduction. Many, probably most, patients will need more or than one drug to achieve the intended goal. A pragmatic approach to treatment it is often needed.

ACE inhibitors or angiotensin 2 receptor blockers are the first choice in those with microalbuminuria. ACE inhibitors, angiotensin 2 receptor blockers, cardioselective β blockers or thiazide diuretics are reasonable first line treatment in those without microalbuminuria. Long acting dihydropyridine calcium channel antagonists (for example, amlodipine) have an important role in treating hypertension and are second line agents.

Guidelines for choice of antihypertensive drugs

- Those with heart failure should have an ACE inhibitor whenever possible, which can be combined with a diuretic
- Conversion to an angiotensin 2 receptor blocker is helpful if a patient develops cough on an ACE inhibitor
- Addition of an angiotensin 2 receptor blocker to an ACE inhibitor will have an additive benefits on blood pressure but not on microalbuminuria
- Blood pressure in some African and Caribbean patients may respond better to calcium channel antagonists and diuretics than to other agents

Lipids and diabetes

Hyperlipidaemias also commonly exist in those with diabetes and increase still further the risk of ischaemic heart disease, especially in Type 2 diabetes. Detection and control of hyperlipidaemia can effectively reduce MI, coronary deaths and overall mortality. Indeed, even when low density lipoprotein (LDL) cholesterol is normal or even slightly raised in Type 2 diabetes (the major abnormalities being low HDL cholesterol and high triglyceride) the LDL particles may be qualitatively different and more atherogenic.

Screening for dyslipidaemia

This is an essential aspect of the annual review. If the lipid profile is entirely normal, further screening could be postponed for three to five years unless circumstances change. An elevated triglyceride needs confirmation when fasting.

Diabetic control

Optimising diabetic control often improves an abnormal lipid profile in Type 1 diabetes and sometimes in Type 2 diabetes.

Other medications and alcohol

Some drugs and also a high alcohol intake disturb plasma lipids (see table) and this aspect of treatment must be examined and if necessary modified.

Lifestyle measures

The importance of stopping smoking, weight reduction and exercise are described in chapter 3. Advice on low fat diets is also needed. Hypertension must be treated.

Lipid modifying drugs

Statins are the first line of drugs for treating hypercholesterolaemia; and fibrates for treating hypertriglyceridaemia. Statins and fibrates can be used alone or together for treating mixed hyperlipidaemia. Specialist advice should be taken on resistant or complex hyperlipidaemic states.

Targets for treatment

The current targets for cholesterol and other lipid fractions are now the same for primary prevention in diabetes as for secondary prevention in people without diabetes. They are as follows:

- total cholesterol <5·0 mmol/l
- fasting triglyceride <2·0 mmol/l
- LDL cholesterol <3·0 mmol/l
- high density lipoprotein (HDL) cholesterol >1·1 mmol/l.

It is desirable that the ratio (HDL cholesterol)/(total cholesterol − HDL cholesterol) should be >0·25. Alternatively, the total cholesterol/HDL cholesterol should be <3.0.

There is much debate as to whether lipid lowering agents are of any value if started over the age of 75 years.

Heart Protection Study

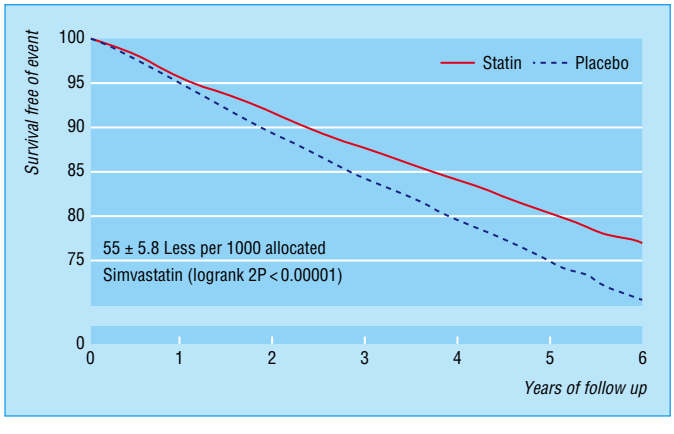
This huge double blind trial of 20 000 people at increased risk of vascular disease examined the benefits of treatment with simvastatin 40 mg daily, and reported its results in March 2002 (see further reading list). The vascular event rate curves began to separate by the end of the first year, and the absolute benefits of treatment were seen to increase over time, becoming very obvious after five to six years. The major results are shown overleaf.

Some causes of secondary hyperlipidaemia

	Main lipid abnormalities
Alcohol abuse	↑ Triglyceride, ↑ HDL
Therapeutic drugs (diuretics, oral contraceptives retinoids, corticosteroids, anabolic steroids, progestogens related to testosterone)	↑ Triglyceride or cholesterol or both, ↓ HDL
Hypothyroidism	↑ Cholesterol
Chronic renal failure	↑ Triglyceride
Nephrotic syndrome	↑ Cholesterol, ± ↑ triglyceride
Cholestasis	↑ Cholesterol
Bulimia	↑ Triglyceride
Anorexia nervosa	↑ Cholesterol
Pregnancy	↑ Triglyceride

Targets for blood lipids control suggested by the European Diabetes Policy Group

	Low risk	At risk	High risk
Serum total cholesterol			
mmol/l	<4·8	4·8-6·0	>6·0
mg/dl	<185	184-230	>230
Serum LDL cholesterol			
mmol/l	<3·0	3·0-4·0	>4·0
mg/dl	<115	115-155	>155
Serum HDL cholesterol			
mmol/l	>1·2	1·0-1·2	<1·0
mg/dl	>46	39-46	<39
Serum triglycerides			
mmol/l	<1·7	1·7-2·2	>2·2
mg/dl	<150	150-200	>200



Simvastatin: vascular event by follow up duration

- reduction in vascular deaths 17%
- reduction in major vascular events 24%
- reduction in strokes 27%.

Treatment benefits were not only obvious among patients who had previously had an MI but also in those with prior cerebrovascular or peripheral vascular disease. Benefits were also demonstrated for diabetic patients without previous CHD. Both sexes benefited equally, and elderly patients over 75 years of age were seen to benefit to the same extent as younger patients.

The implications of the results of this study for diabetic patients are obvious, and serious consideration with regard to offering treatment with a statin drug must be given.

Severe hyperlipidaemic states

Extreme mixed hyperlipidaemias are, on rare occasions, associated with uncontrolled diabetes. The plasma has a milky appearance, and xanthomata appear in the skin as bright yellow papules particularly at the elbows, knees and buttocks. Even the retina assumes the pallor of lipaemia retinalis described as the "peaches and cream" appearance. The condition normally resolves when glycaemic control is achieved, lipid levels often return to normal, and the xanthomata disappear. Most but not all patients have Type 2 diabetes. The condition needs to be carefully monitored, and it is often wise to administer lipid lowering drugs until resolution. Sometimes they need to be continued indefinitely.

Milky plasma of severe hyperlipidiaemia

Myocardial infarction (MI)

The greater risk of CHD in diabetic patients is accompanied by a greater risk of MI: approximately 10% of all MIs occur in diabetic patients. Unfortunately, mortality rates are also about twofold higher. The increased mortality is attributable to left ventricular dysfunction leading to left ventricular failure and cardiogenic shock.

Presentation of MI in those with diabetes is the same as in those without, although a greater proportion lack chest pain. Absence of cardiac pain has been attributed to the cardiac denervation although this concept lacks conviction.

Treatment

Thrombolysis

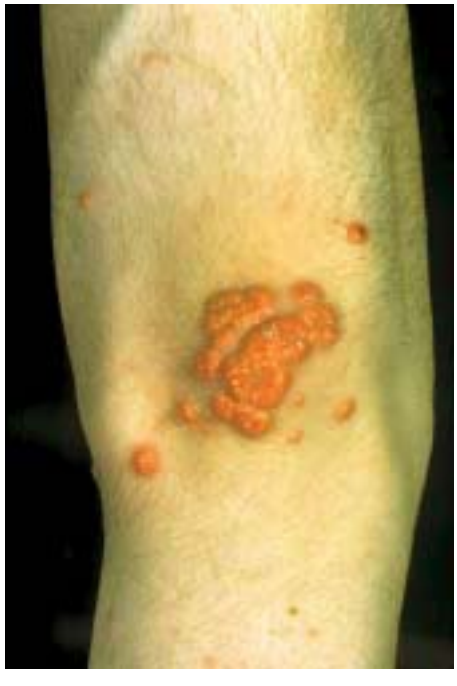
This should be used in diabetic patients for the same indications as for non-diabetics. The risk of vitreous haemorrhage in those with proliferative retinopathy is negligible. The benefits to diabetic patients with MI appears even greater than to those without diabetes.

Aspirin 300 mg

Given at the onset of a MI, enteric coated aspirin 75-150 mg daily should be continued indefinitely thereafter.

Insulin treatment

Commenced at the onset of an MI, insulin treatment leading to optimal glycaemic control confers benefits on reducing mortality after discharge from hospital. The DIGAMI (Diabetes, Insulin, Glucose infusion in Acute Myocardial Infarction) study examined 620 patients with established or newly diagnosed diabetes (plasma glucose exceeding 11·1 mmol/l): an insulin and glucose infusion was started at presentation, followed by a multiple insulin injection regimen for at least three months. Twelve-month mortality was reduced by 30%. Confirmation of these results is required, but at present this study presents the best evidence available for the management of diabetes after MI.



Xanthomata on the knee from severe hyperlipidaemia

All diabetic patients, newly diagnosed and with established disease should be treated with insulin, and in many, this should be continued indefinitely. Some discretion needs to be used among elderly patients, or others who cannot easily or comfortably use insulin, in whom other simple measures can also achieve very good diabetic control.

Coronary artery bypass grafts and angioplasty

Decisions advising on the need for invasive coronary artery treatments are made in the same way as for non-diabetics. Coronary artery bypass grafts may offer an improved prognosis to those with diabetes after spontaneous Q wave infarction compared with angioplasty, and further investigations of this apparent benefit are in progress.

After MI

Half the patients after their first MI die during the following twelve months, and half of those die before they reach hospital. Continuing optimal management of all risk factors is strongly recommended.

Management of risk factors after MI

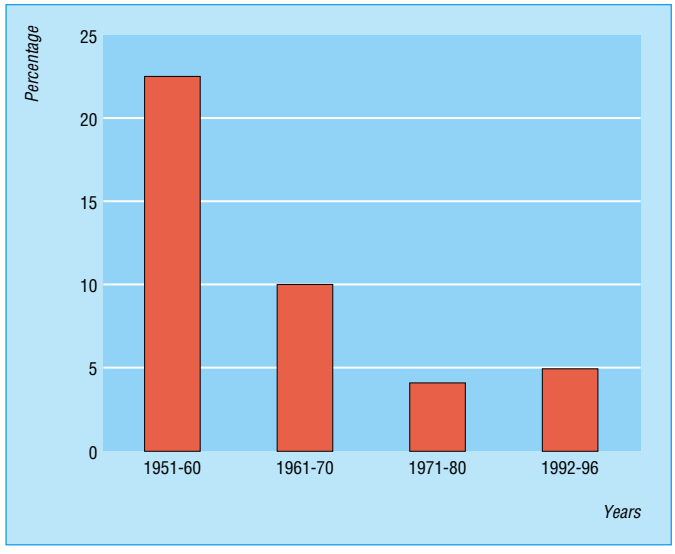
- Cessation of smoking
- Optimal blood pressure management
- Optimal diabetic control normally with insulin
- Optimal lipid control
- Use of aspirin, β blockers, ACE inhibitors as indicated

The histogram showing rates of serious cardiovascular events according to blood pressure in patients with hypertension and Type 2 diabetes is adapted from Hansson L, et al. *Lancet* 1998;351:1755-62. The figure showing Kaplan Meier estimates of the composite outcome of myocardial infarction, stroke, or death from cardiovascular causes in the HOPE study is adapted from Yusuf S, et al. *New Engl J Med* 2000;342:145-53. The figure showing simvastatin: vascular event by follow up duration is adapted from the Heart Protection Study website *www.ctsu.ox.ac.uk/~hps*. The table showing some causes of secondary hyperlipaedemia is adapted from International Task Force for Prevention of Coronary Heart Disease. *Nutr Metab Cardiovasc Dis* 1992;2:113-56.

18 Pregnancy

Fifty years ago, more than one quarter of diabetic pregnancies ended in fetal death. Now, due to major developments in obstetrics, diabetes, and paediatrics, most are successful, although perinatal mortality at 2 to 5% is still higher than normal (<1%). Major congenital malformations, some of which could be prevented by good preconception diabetes control, are still excessive.

Good results in diabetic pregnancy can only be achieved in centres where appropriate expertise exists. Joint clinics where care is shared between diabetes physicians and specialist nurses working alongside obstetricians with their team of midwives should be the norm. Attendance at separate clinics is a second-rate option: it is inconvenient and often confusing for patients and leads to inconsistent care and advice.



Diabetic pregnancies: perinatal mortality—King's College Hospital

Management

Pre-pregnancy counselling

Education of diabetic women of childbearing age is important: they should be advised that if they plan a pregnancy they should attend a pre-pregnancy counselling clinic (ideally linked to the antenatal clinic), start taking folic acid and aim for optimal diabetic control before conception in order to reduce the incidence of congenital malformations. The goal is a HbA_{1c} < 6·5% although this is sometimes difficult to achieve even after several months of intense effort, and pragmatic decisions are often needed.

Pregnant diabetic women should report to their doctor or clinic as early as possible in pregnancy and be referred without delay to a joint diabetic antenatal clinic. The midwife should encourage them to attend their local antenatal classes for social contacts.

Congenital malformations

Pregnancy in Type 1 diabetes is associated with an increased risk of congenital malformations. There is a clear relationship between poor diabetic control in the first trimester, and it is equally established that optimal control before conception reduces the rate of congenital malformations to near normal. Thus an HbA_{1c} more than 50% above the upper limit is associated with a two to fourfold increase in the abnormality rate. While in a non-diabetic population congenital malformations occur in 1 to 1·5% of pregnancies, this may be as high as 4 to 6% in those with diabetes, or even higher in those with exceptionally poor control during the first trimester.

The abnormalities are essentially the same as those occurring in the general population and include skeletal abnormalities such as spina bifida or hemivertebra, congenital heart disease, and neurological defects such as microcephaly or anencephaly. Sacral agenesis is greatly increased in diabetes but is still very rare.

Tight control of diabetes before and during pregnancy

Diabetic control should be optimal throughout pregnancy; if it is not, admission to hospital, even for short spells, should be arranged without delay. Home blood glucose monitoring is performed at least four times daily to determine preprandial readings, and about one and a half hours after the main meals perhaps twice weekly to determine postprandial levels,

Management of pregnancy and diabetes

- Pre-pregnancy counselling: understanding the need for tight diabetic control before conception, the need to take folic acid and of the commitment required for a good outcome
- Tight control of diabetes through pregnancy
- Management of diabetic complications during pregnancy
- Obstetric requirements and monitoring during pregnancy
- Management of diabetes in labour
- Management of the neonate

The incidence and relative risk of congenital malformations* in infants of diabetic mothers

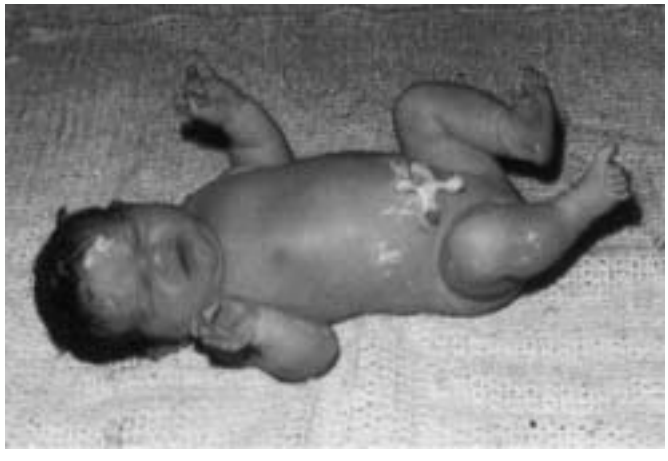
Malformation	Incidence/1000 births	Relative risk compared with infants of non-diabetic mothers
Cardiac	10	4
Anencephaly	3	5
Arthrogryphosis	0.3	28
Ureteral duplication	0.7	23
Cystic kidney	0.6	4
Renal agenesis	0.3	5
Anorectal atresia	0.3	4
Caudal regression	1.3	212
Pesudohermaphroditism	0.6	11

*Thought to be of greater incidence in infants of diabetic mothers than in infants of non-diabetic mothers

especially when optimal HbA_{1c} levels have not been achieved. Insulin injections are usually needed three or four times daily and patients should be taught how to adjust their own insulin doses. The dose of insulin needed often increases substantially during pregnancy, sometimes to twice or three times the usual amount and this needs to be anticipated. Good blood glucose control can be difficult to attain during the first trimester and attempts to do so are often complicated by severe hypoglycaemic episodes; thereafter most patients find that they can reach the standards of tight control without difficulty and without untoward hypoglycaemia. Suitable insulin regimens for tight control are described on page 21. Where serious difficulties (including severe hypoglycaemia) in achieving optimal control are encountered, continuous subcutaneous insulin infusion should be considered (see page 28-9). These patients and their families should be equipped with all the essential materials needed to cope with hypoglycaemia (see pages 34-5).

Type 2 diabetic pregnant women are increasingly seen especially among ethnic minorities. Treatment is by diet or diet and insulin only. Those who are on oral hypoglycaemic drugs should be changed to insulin, preferably before conception. There is however no evidence that tablet treatment is teratogenic, and where circumstances demand (for example, in countries where insulin is not readily available, or for patients who absolutely refuse insulin) they can be used. Indeed the use of glyburide has been examined and from a single large study in gestational diabetes appears safe, effective and, unlike some other sulphonylureas, does not cross the placenta, thus not provoking neonatal hypoglycaemia.

Preprandial blood glucose values should be maintained as near normal as possible and should preferably be kept below 4·0 to 5·5 mmol/l, though postprandial levels, which should be checked especially when HbA_{1c} is suboptimal, may be up to 7·0 mmol/l. The target for HbA_{1c} should ideally be within the normal range—that is, <6% if possible, although on occasions levels up to 7% represent the best that can be achieved



A typically large baby born to a diabetic mother

Management of diabetic complications during pregnancy

Retinopathy

Fundi should be examined routinely at the beginning, middle and end of pregnancy. If retinopathy is present early in pregnancy, more frequent examination is needed because occasionally progression is rapid during the course of pregnancy. If proliferative changes are present photocoagulation should be performed urgently.

Nephropathy

Patients with established proteinuria from nephropathy can expect problems during pregnancy, with an increased risk of fetal loss. When this is considered, together with the knowledge that the mother may need dialysis or transplantation within a few years, many of these patients should be advised to avoid pregnancy. However, with skilful management, most such patients whose renal function is still normal or near normal can expect a live infant. If however the serum creatinine exceeds 200 mmol/l the outlook for the fetus is bleak and patients should be strongly advised not to get pregnant.

Diabetic women with established nephropathy should normally be advised against pregnancy. If the serum creatinine exceeds 200 μmol/l, the outlook for the fetus is very poor indeed

Nephropathy and antihypertensive treatment

Patients with nephropathy who become pregnant can expect to become hypertensive and later often develop severe oedema. Fetal growth may be retarded, and because the problems can be severe, very early delivery of exceptionally small infants becomes necessary. With modern intensive neonatal care, most babies now survive, though a significant proportion develop mild or occasionally severe disabilities. Hypertension is treated with methyldopa, hydralazine, or occasionally labetalol or amlodipine. Frusemide can be given if a diuretic is needed, but

thiazides are contraindicated. Patients taking angiotensin converting enzyme inhibitors should be changed to other antihypertensive agents before pregnancy because of very serious later adverse effects on the fetus during pregnancy, though congenital malformations are not increased. Any patient with proteinuria from nephropathy may need protracted inpatient care, which should be undertaken in a specialist unit.

Obstetric requirements through pregnancy

An early ultrasound scan establishes the exact gestational age and detects major fetal abnormalities. A specific congenital abnormality ultrasound scan is required before 20 weeks. Thereafter a monthly scan documents the rate of fetal growth. There is considerable acceleration of growth after 28 weeks in the majority of diabetic pregnancies, even in the presence of good diabetic control. Retarded growth may be serious and requires specialist investigation. Assessment of fetal size and weight is important in determining the timing and mode of delivery.

Management in labour

Most patients should be admitted for a short period before planned delivery or for a longer period if there are either obstetric or diabetic problems. Fetal heart rate and its beat-to-beat variations are monitored regularly by an instantaneous heart rate meter and in cases of fetal distress other investigations are needed. Delivery is planned to take place as near to term as possible, unless there are medical or obstetric indications for earlier induction. The aim is for a vaginal delivery whenever possible. However, delivery is still on average at 37 weeks and about two-thirds of deliveries are still by caesarean section.

Glucose and insulin in labour

Glucose and insulin are given by intravenous infusion for all vaginal deliveries as follows:

Intravenous dextrose (10%): one litre every eight hours delivered at a steady rate.

Intravenous insulin: soluble insulin diluted in physiological saline (1 unit insulin/ml saline) and administered by infusion pump at about 1 unit/h (usual range 0·5-2 units/h). If very low infusion rates are used the insulin concentration can be halved.

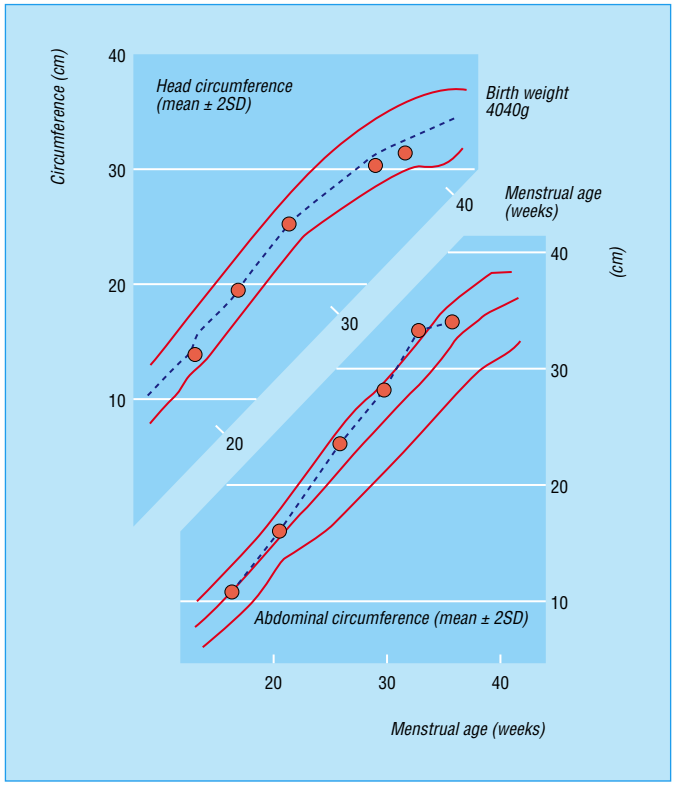
Blood glucose concentrations should be maintained in the range 4·0-7·0 mmol/l. Insulin infusion is continued until the patient can restart her normal meals. The pre-pregnancy insulin dose is then restarted, otherwise severe hypoglycaemia will occur; if the patient was not previously on insulin, the insulin dose is halved.

Premature labour

Because of the hazards of premature labour, attempts may be made to promote fetal lung maturity by giving dexamethasone. This causes severe hyperglycaemia unless an intravenous insulin infusion is started at the same time as the administration of dexamethasone. Large doses of insulin may be needed.

Caesarean section

Insulin infusion is always used, as described in the section on management during surgery (see page 40).



Ultrasound measurement of fetal growth showing excessive increase of abdominal girth, indicating delivery of a large baby

Blood glucose regimen in labour

Blood glucose	Infusion rate
<4·0 mmol/l*	0·5 U/h
4·0-7·0 mmol/l	1 U/h
>7·0 mmol/l	2 U/h

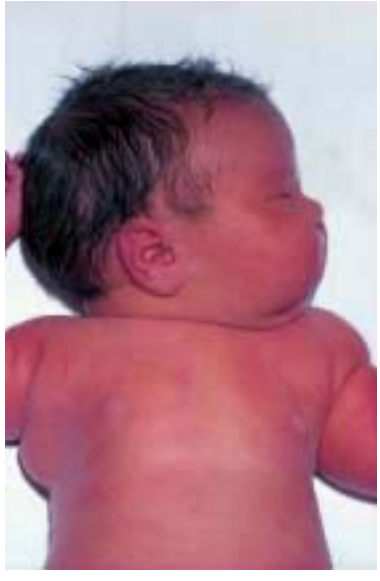
*If the blood glucose concentration decreases to <3·0 mmol/1, insulin infusion can be stopped for up to 30 minutes.

If dexamethasone is used in premature labour, insulin must be infused at the same time to avoid severe hyperglycaemia

The neonate

The babies of diabetic mothers are larger than normal and nearly one-third of those with Type 1 diabetes exceed the 97·5% centile. However they no longer need care routinely in special care baby units unless there are specific reasons. Respiratory distress syndrome is now rarely seen in these infants unless they are very premature. Blood glucose concentrations should be checked regularly, especially in jittery babies, because hypoglycaemia is still commoner than in the infants of non-diabetic mothers. Polycythaemia, hyperbilirubinaemia, and hypocalcaemia are also commoner among these infants.

Breast feeding
This is encouraged in diabetic mothers as in non-diabetic mothers. The mother's diet should be increased by about 50 g of carbohydrate daily and ample fluids taken. The insulin dose is not usually affected when these measures are followed. Breast feeding mothers should not use oral hypoglycaemic agents.



The neonate. "… they convey the distinct impression of having had such a surfeit of food and fluid pressed on them by an insistent hostess that they desire only peace so that they can recover from their excesses" (J. Farquhar)

Gestational diabetes (diabetes discovered during pregnancy)

The detection of gestational diabetes is undertaken by screening procedures in antenatal clinics. Babies born to women with gestational diabetes are frequently macrosomic and although both the effects on mortality and the benefits of intensive treatment are still uncertain, best practice requires optimal diabetic care for these patients.

Diagnosis
Routine blood glucose measurements are made in every pregnant woman between 26 and 28 weeks.

All women who have previously had gestational diabetes should have a glucose tolerance test between 26 and 30 weeks unless there are indications to perform it sooner.

Those from ethnic minorities carry a particularly high risk of gestational diabetes and at King's College Hospital those of African-Caribbean origin accounted for about 80% of them.

Diagnosis of gestational diabetes

- A random plasma glucose concentration ⩾11·1 mmol/l; fasting plasma glucose ⩾7·0 mmol/l; whole blood glucose ⩾6·1 mmol/l; are diagnostic of diabetes
- If a random blood glucose measurement is greater than 5·8 mmol/l perform a glucose tolerance test
- A 2 hour blood glucose ⩾7·8 mmol/l during a glucose tolerance test in pregnancy is diagnostic of gestational diabetes

Treatment
Gestational diabetes is initially treated by diet alone. If control deteriorates using the criteria described above, insulin should be used, and stopped immediately on delivery of the infant. Oral hypoglycaemics are not advised, though there is no evidence that they are harmful.

Postpartum
A glucose tolerance test should be repeated six weeks after delivery. More than half of the patients return to normal; these women however have an approximately 50% risk of developing Type 2 diabetes later in life, sometimes within one year. An annual check of fasting blood glucose concentration is desirable, and they should be advised to avoid excessive weight gain by a regular programme of healthy eating and exercise (see chapter 3). Those patients who remain diabetic after pregnancy should be treated in the usual way.

Contraception and diabetes

Good contraceptive advice for diabetic patients is vital to ensure that pregnancies are planned and conception takes place when diabetic control is optimal. All methods of contraception are available to diabetic women; the combined oral contraceptive pill is suitable in the absence of macrovascular disease or microvascular disease especially proteinuria, while progestogen-only methods provide a range of options including highly reliable long term methods, such as Depoprovera and implantable methods such as Implanon.

There is no evidence of any clinically significant effect on diabetic control from either combined oestrogen-progestogen pills or progestogen-only methods, nor is there any influence on the progression of diabetic complications. In the case of women with a history of gestational diabetes, the use of combined pills does not influence the subsequent development of Type 2 diabetes, but there is at present some doubt regarding the use of progestogen-only pills during lactation in women with recent gestational diabetes in whom there may be an increase in the subsequent incidence of Type 2 diabetes.

Intrauterine methods of contraception are suitable for women in stable relationships who have had at least one pregnancy, and may be ideal for diabetic women in whom hormonal (oestrogen containing) methods are contraindicated for either diabetic specific conditions or those unrelated to diabetes including cardiovascular or cerebrovascular disorders, venous thrombosis, pulmonary embolism or liver disease. Barrier methods of contraception have no metabolic consequences but are often insufficiently reliable, particularly if diabetic control is poor, when pregnancy should be rigorously avoided.

Sterilisation may be considered the ideal option when the family is complete, but it should be borne in mind that both Implanon and the progestogen delivering intrauterine system Mirena, provide more reliable contraception than laparoscopic sterilisation and are of course both reversible should circumstances change.

The table showing the incidence and relative risk of congenital malformations in infants of diabetic mothers is adapted from Combs CA, et al. *Clin Obstet Gynecol* 1991;5:315-31. The photograph of a typically large baby born to a diabetic mother is from Chamberlain G, Morgan M. *ABC of Antenatal Care*, 4th ed. London: BMJ Publishing Group, 2002.

The story of Mrs B-J continued: pregnancy

I had heard many tales about the trauma of diabetics who had babies, and I made up my mind I would never have any, even if I did get married. Once, when I was at the clinic, I was asked by the Sister if I would go up the maternity ward and have a chat with the two diabetic expectant mothers. They had been in hospital for nearly three months, which was normal in those days. The poor dears were bored stiff and glad to see a different face. I hope I cheered them up a bit. They convinced me that I would never want to follow their example and I never did. Dr Pyke told me a few years ago that things had changed now and expectant mothers no longer have to serve such a long sentence.

19 Organisation of diabetic care: primary-secondary care interface

Care of people with diabetes requires enthusiasm, commitment and organisation. There are various ways of undertaking it, but without interest and motivation, none will succeed.

It is best to bring diabetic patients together into properly organised clinics, whether in general practice or hospital, so that they can benefit from the wide range of services needed for their long-term care. A close liaison is needed between general practice and hospital specialists, and there are substantial new developments at the interface with Primary Care Groups and Primary Care Trusts. Patients can then have access to all members of the large team now involved in delivering the treatment and advice which they need. Diabetes shared care schemes are very advanced in many parts of the United Kingdom and set a model for other specialties, as they represent at the same time both efficient treatment and ideal links between hospitals and communities.

Any scheme aims to deliver the best care to patients. In order to do so there must be an efficient flow of information about patients, and the shared experience of a dedicated staff. Dissemination of expertise among all those concerned is essential. Schemes require nurturing, and an audit of outcome measures is needed for feedback to assess progress and identify courses of action. Local demographic trends must be understood, including an awareness of the needs of different ethnic groups. Education of the public is becoming increasingly important, and information is needed in schools as well. Research, innovation, and renewal are constantly needed. Increasingly, information regarding local diabetes services are provided on websites.

Requirements for diabetes care

A complex range of services for comprehensive diabetes care is needed as follows:

- to establish diagnosis and initiate treatment
- for patient education leading to independence
- to achieve optimal or appropriate diabetic control
- for screening and detection of diabetic complications
- to enable treatment of diabetic complications
- for care of those who are acutely or chronically ill
- for education of all medical and nursing staff involved in diabetes care.

The facilities needed to achieve these goals are provided by diabetes centres, which offer a common base for an integrated specialist and primary care diabetes service, and by the general practitioner.

Services provided by the general practitioner

- A diabetes register
- Dedicated time for care of people with diabetes
- Preferably one doctor with a special interest in diabetes
- A practice nurse who has received some specific training
- Access to all necessary laboratory services
- Facilities for complications screening and access to an eye screening programme
- Provision of access to the diabetes team to provide appropriate education, dietetic advice, and podiatry

St Vincent Declaration, 1989

A joint European initiative between the World Health Organization and the International Diabetes Federation resulted in the publication of the St Vincent Declaration in 1989, which calls for targets for improving the outlook for diabetic patients. The recommendations include:

- Reducing new blindness due to diabetes by a third or more
- Reducing numbers of people entering end stage diabetic renal failure by at least a third
- Reducing by a half the rate of limb amputations for diabetic gangrene
- Cutting morbidity and mortality from coronary heart disease in diabetic patients by vigorous programmes of risk reduction
- Achieving pregnancy outcome in diabetic women that approximates that of non-diabetic women

Services provided by diabetes centres

- Provide expertise, literature, and teaching aids based at the hospital department, and take the lead role in organisation of district diabetes services to co-ordinate hospital and general practitioner activities
- Provide mutually agreed guidelines on management
- Provide emergency access for patient and doctors (a direct, dedicated telephone line is essential)
- Establish a register of diabetic patients (and where possible also identify those at risk)
- The team should provide a clinical service for all new and established diabetic patients requiring specialist hospital attention
- Provide services jointly (ideally in joint clinics) with relevant specialists for the treatment of:
  - Retinopathy with an ophthalmologist
  - Pregnancy with obstetricians
  - Children and adolescents with a paediatrician
  - Foot problems (including peripheral vascular disease) with a vascular surgeon
  - Renal problems with a nephrologist
  - Neuropathy with a neurologist
  - Erectile dysfunction
  - Family and psychological problems
- Provide a screening service for detection of diabetic complications (see page 45)
- Oversee the care of all diabetic hospital inpatients by training ward-based “link” nurses
- Provide interpreters and patient advocates where necessary

The medical consultation

Every patient presenting for the first time should undergo a full clinical appraisal and physical examination. Subsequent consultations should not only include an assessment of the diabetes and its complications, but also specifically enquire regarding episodes of hypoglycaemia, supplementing this enquiry with reminders of how episodes should be managed and avoided. Inspection of insulin injection sites is often neglected. Assessment of other medical conditions and medication is also important because of potential interactions, and the need to offer each individual patient clinically appropriate advice.

The Physician's Visit by Jan Steen, 1663

Indications for referral to the hospital diabetes centre

Good communications between community and hospital are crucial. While referral patterns will depend on local expertise, the following guidelines are offered:
Group 1 These patients should normally attend hospital diabetic clinics:

- Type 1 diabetic patients
- all children and adolescents with diabetes
- those with problems from hypoglycaemia
- patients who need pre-pregnancy counselling and all those who are pregnant
- patients with significant complications
- those with active foot lesions or sepsis, or both.

Group 2 Referral to hospital desirable, depending on local practice expertise:

- decision to commence insulin
- patients in whom adequate control is not achieved (for criteria see page 10)
- newly diagnosed patients for assessment, education, and initiation of treatment.

Urgent expert assistance (usually at the hospital diabetes centre is needed):

- If an active foot lesion develops
- If there is a rapid decline of vision
- For reappraisal after a serious hypoglycaemic event
- If renal function deteriorates unexpectedly

The diabetes team

Because patient numbers are large, the disease lifelong, and its complications complex and diverse, a wide range of practitioners and specialists are involved in delivering a comprehensive service of high quality. The team needs a base, normally the diabetes centre in the local hospital, in order to maintain close communication with patients and professionals both in the hospital and in the community, as well as promoting innovation and research. Fragmented teams are likely to fail.

Diabetes specialist nurses and consultant nurses
The most important single innovation in diabetic care during the past three decades has been the increasing involvement of highly trained diabetes specialist nurses who can transform the standard of diabetic care, achieving liaison between hospital, general practitioner and patients at home, and offering a wide range of clinical and educational expertise. The Royal College of Nursing recommends that there should be one specialist nurse for a population of 50 000, and one for every 50 families with a diabetic child.

The training of nurses for diabetic care is of central importance and undertaken on specifically designed diploma and degree courses as well as at the diabetes centres themselves.

The diabetes team

Diabetes centre	General practice	Community
Diabetes physician	General practitioner	
	Practice nurse	
Diabetes specialist nurse	⟷	
⟷	Dietitians ⟷	
⟷	Podiatrists ⟷	
⟷	Retinal screening staff ⟷	
Paediatrician		
Ophthalmologist		
Obstetrician		
Orthopaedic surgeon		
Vascular surgeon		
Renal physician		
Neurologist		
Psychologist		

The many important roles of the diabetes specialist nurse are closely linked to those of consultants, and can be summarised as follows:

- treatment of individual patients, linked with advice and counselling, establishing the right techniques and motivation needed to achieve proper diabetic management and control
- care of patients in hospital wards
- education of patients and professionals (see below)
- involvement in community care
- research, audit, and setting standards.

Consultant nurses have been established recently, with the aim both of enhancing clinical care and stimulating research and service innovations as well.

Diabetic kitchen, King's College Hospital, 1935 (Diabetes UK)

The role of the consultant in the community

The specialist diabetic consultant has a major role to play not only in the management of diabetes itself but also in the delivery of services across the community. Experience in the community surrounding King's College Hospital suggests the potential benefits of the following model seen in the box.

Role of the consultant

- The hospital team regularly visits general practices to see selected patients attending consultation
- The hospital team comprises a consultant and a diabetes specialist nurse, accompanied by a specialist registrar and a medical student
- The practice team includes the host general practitioners, practice nurses, and others such as local district nurses, health visitors, podiatrists, and dieticians

Education

Education of patients

An integral aspect of diabetes care is to inform all patients of the nature of the disorder and its treatment, and to place the potential threat of complications in their true perspective. Educational facilities are offered by the whole of the diabetes team both to individuals and to groups.

Instructions to new patients are always given initially on an individual basis. Most centres also organise courses for groups, ranging from a single half-day to comprehensive weekly series. There should be separate education groups for Type 1 and Type 2 diabetic patients, and the courses should provide scope for discussion and questions as well as direct instruction. Ongoing education is also needed to refresh memories over the several decades following diagnosis.

Education of patients has become very sophisticated in the field of diabetes; it has reduced admissions and to some extent complication rates, notably amputations. It is a concept which could be applied much more extensively to other areas of medicine.

Aims of an education programme

- To explain the nature of the disease and its complications
- To explain the treatment, starting with the simplest ground rules and eventually provide comprehensive instructions on both treatment and monitoring, enabling patients to take control of their own condition
- To explain dietary and other lifestyle requirements
- To provide printed literature. "Starter" packs should contain:
  - A booklet about diabetes
  - Dietary instructions
  - Home monitoring booklet with full instructions
  - Information on driving
  - Essential telephone numbers
  - Leaflets on Diabetes UK

Education of health professionals

In order to maintain standards, all those involved in diabetes care require regular updating, and every locality must take responsibility for educational programmes. These should include practice nurses, specialist nurses, hospital nurses, as well as junior and senior medical staff in both hospitals and general practice. Organisation of educational programmes requires a considerable resource.

Records

Computers are essential for the maintenance of good records on diabetes, though no single system is clearly superior. Maintenance of a register of diabetic patients is an essential operation and becomes increasingly important for recalling patients for review and assessment. Records of varying complexity can be held on the computer and the presentation can be structured so as to present the necessary information described below.

DIAMOND COMPUTER SYSTEM

Name **Hospital Number** **D.O.B.:**

Date	Weight	Body mass index	Systolic pressure	Diastolic pressure	Random blood glucose	HbA_{1c}	Creatinine	Albumin/ Creatinine ratio	Abustix	Cholesterol	Retinopathy		Visual acuity		Foot pulses		Feet
											R	L	R	L	R	L	
12/05/97	95.8	32.38	145	98	12.9	8.1	135	40.0	+ +		Laser	Laser	6/18	6/9	+	0	Healthy
10/02/97	97.7	33.02	130	80	13.9	10.7	72		+ +	6.8	Laser	Laser	6/18	6/9	+	0	Foot ulcer
16/09/96	100.5	33.97	150	94	11.8				+		Pre-prolif	Pre-prolif	6/6	6/6	+	+	
09/09/96	100.2	33.87	140	90	12.1	6.5	78		+ +		B'gd	B'gd	6/6	6/6			
11/03/96	97.1	32.82	130	85	4.6	6.3		3.4	Negative		B'gd	B'gd			+	+	Healthy
11/12/95	96.2		165	98	9.7	6.2	70	2.9	Trace	5.7	0	0	6/6	6/5			
01/12/93	90.6		110	78		11.2		0.2	Negative		0	0	6/6	6/5	+	+	Healthy

Patient history record

Special records are needed for proper care of diabetic patients both in hospital and in general practice—however this is organised, records must be immediately available when required. In hospital there are huge advantages in maintaining a separate set of records for diabetic patients and this system is used by major departments. Many general practices also keep special records supplementary to the basic "Lloyd George" card. Such records must be designed for serial recording of factual data, including weight, blood glucose, urine tests, HbA_{1c}, visual acuity, complications (results of eye examination in particular), and treatment. There must always be space for recording the outcome of the medical consultation itself and for treatment recommendations. There should be a system to alert the staff to the presence of particular problems, for example, sight-threatening retinopathy, and to the date when the next examination is required (for example, blood pressure measurement or eye examination). Many also incorporate an education checklist which records when patients have attended sessions and ensure that essential advice has been given, for example on driving.

Early diabetic record in the hand of Dr R D Lawrence

Other facilities

Local diabetes service advisory groups (LDSAGs)
Coordination of services by LDSAGs is crucial to their success. Local committees can achieve this very effectively. They should comprise representatives of local purchasing authorities (for example health authorities or primary care trusts), providers (hospital consultants and general practitioners), diabetes specialist nurses, and consumers (diabetic patients). Effective discussions in this group can substantially enhance local services which might otherwise become seriously fragmented.

Diabetes facilitators
Support for general practices establishing diabetes services is crucial and a team for this purpose, comprising specialist nurses, facilitators, and a dietician, greatly enhances this process, assisting them with the development of optimal facilities and providing them with useful guidelines. The National Diabetes Facilitator's Group runs training courses.

Diabetes UK
The central resources of Diabetes UK provide direct advice for both patients and health professionals by printed literature, and access to scientific and epidemiological information. Diabetes UK also funds research, and provides scientific and educational meetings, children's camps, family weekends, and many other activities.

Local branches of Diabetes UK, organised by people with diabetes, serve as self help and fundraising groups, as well as helping to maintain high quality local services.

Juvenile Diabetes Research Foundation (JDRF)
This organisation was founded in the United States and is now established in the United Kingdom as well. Its chief aims are fund raising to support research, particularly in diabetes prevention and treatment, together with promotion of understanding of this condition by the public.

National Service Framework (NSF) for Diabetes: standards

The NSF for diabetes was published in 2002 and full details can be found on the website <www.doh.gov.uk/nsf/diabetes>

The standards have been divided into 12 sections relating to the following nine categories:

1 Prevention of Type 2 diabetes
2 Identification of people with Type 2 diabetes
3 Empowering people with diabetes
4 Clinical care of adults with diabetes
5 Clinical care of children and young people with diabetes
6 Management of diabetic emergencies
7 Care of people with diabetes during admission to hospital
8 Diabetes and pregnancy
9 Detection and management of long-term diabetic complications

Diabetes UK is at 10 Parkway, London NW1 7AA; telephone 020-7424-1000; <www.diabetes.org.uk>

The JDRF is at 19 Angel Gate, London ECIV 2PT; telephone 020-7713-2030 <www.jdrf.org.uk>

Conclusion

Obviously the requirements for diabetes care will vary from one geographical area to another, but anyone who undertakes the care of people with diabetes must heed the words of Dr Elliott P. Joslin: "To retain his patient for 20 years he must shun proprietary remedies as the devil does holy water, continually seek for new knowledge as eagerly as the diabetic grasps for life, but ever sift the wheat from the chaff remembering that faithful treatment in season and out of season is rewarded."

The Physician's Visit by Jan Steen, 1663 is reproduced with permission from the V&A Picture Library. The photograph of EP Josllin is from Joslin EP. *Diabetic Manual*, Lea and Febiger, 1941.

Elliott P Joslin, 1869-1962

The story of Mrs B-J continued: joining the Diabetic Association

I joined the Diabetic Association, as it was called then, when it was first formed (1934). Through them, I was sent to St Mary's Convalescent Home at Birchington to recuperate after the whooping cough attack. I was there for the Coronation of George V1 in 1936, and can well remember the lovely party we had that day. My mother had sent me a parcel of patriotic items such as red, white and blue crepe paper, ribbons, and flowers made of red, white and blue feathers. With these I made myself an outfit for the fancy dress parade and won first prize, a china plate coronation souvenir.

The war brought problems for me and at the end of it, in 1945, I was quite run down. I think this was due to lack of fresh fruit, especially oranges and bananas. Diabetics were allowed to have three times the normal ration of meat, butter, margarine and cheese. To get this, the sugar ration had to be surrendered. Later in the war, when milk was rationed to two pints a week for each adult, diabetics were allowed one pint a day. Apart from the extra milk, I never had extra rations, as my mother witnessed an unpleasant scene one day in Sainsburys, and it put her off ever getting the extras; when a lady with three diabetics in her family was handed a large joint, there was almost a riot and the poor lady was manhandled and the meat torn away from her. The manager called the police to restore order. My mother was so upset that she firmly refused to apply to the Food Office for the permits and said that I would have her rations if necessary.

20 Practical problems

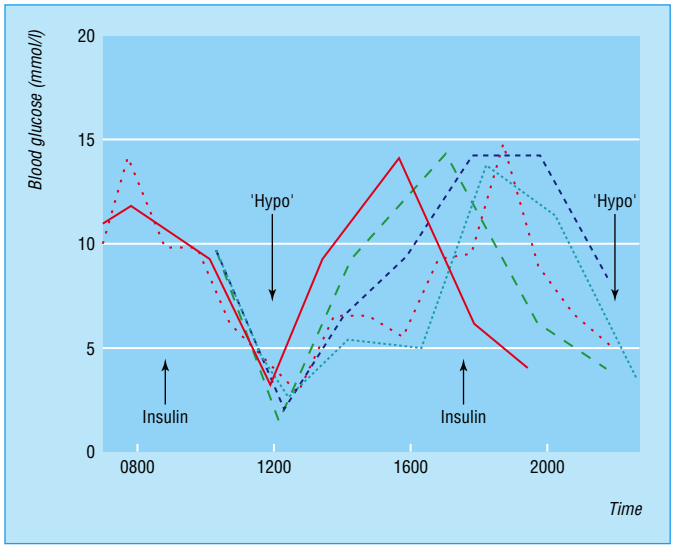
Employment and hobbies

Hypoglycaemia is the major hazard for any insulin treated diabetic patient

Most of the problems of people with diabetes in society result from the ever present possibility of hypoglycaemia in insulin treated patients. Although this hazard is small in many individuals, it is an unacceptable risk in some circumstances. The guiding principles in making the difficult assessments for employment or hobbies relate to whether the risk of confusion during hypoglycaemia affects only the individual or whether it also places the safety of others at risk; the magnitude of the risk of diminished awareness of hypoglycaemia; and the magnitude of the hazard should a hypoglycaemia related accident occur.

Individual firms and industries have generally established their own regulations about the suitability of those with diabetes for particular jobs. If the candidates for employment are rejected unreasonably, solely on account of diabetes they may appeal. People with diabetes are not normally accepted by the armed forces, the police, or the merchant navy, and those already in these occupations may be diverted away from active service to office work. Shiftwork, especially nightshift work, should be avoided if possible by those taking insulin, but some patients can make appropriate adjustments and many may successfully cope with such work.

Diabetic people treated by diet alone or with oral hypoglycaemic tablets who are otherwise fit should be permitted to undertake any occupation or hobby. Their risk of hypoglycaemia is negligible.



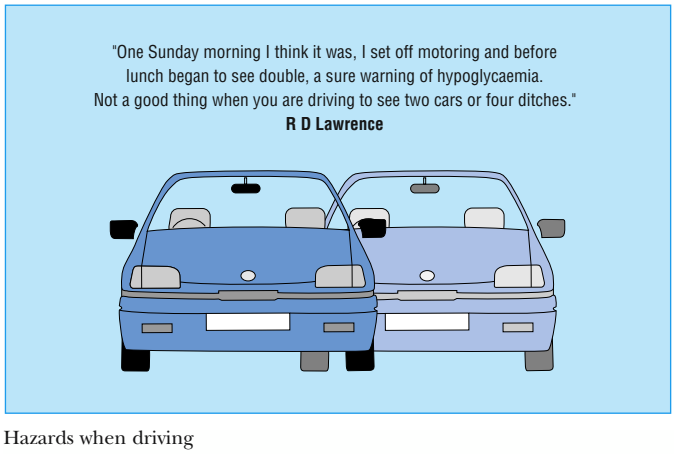
Reproducible profiles of blood glucose levels over five days in an individual patient, showing times when there is a risk of hypoglycaemia

Driving

All diabetic patients who are otherwise physically fit and not affected by blackouts, or other proscribed or relevant medical conditions, are normally allowed to hold ordinary Group 1 (category B) driving licences—that is, vehicles up to 3500 kg with up to nine seats and with a trailer up to 750 kg. The law demands that diabetic patients treated with tablets or insulin (but not those on diet alone) should inform the Driver and Vehicle Licensing Agency (DVLA) in Swansea. If applying for a licence for the first time, the appropriate declaration must be made on the application form. It is helpful to indicate whether insulin is being used. Driving licences are granted for three years and are reissued (at no extra fee) subject to a satisfactory medical report for those on insulin, and up to 70 years of age for those on diet or tablets, subject to any change of their health or treatment.

For any driving licence, visual acuity must be better than 6/12 in the better eye, and the visual field should exceed 120° horizontally and 20° above and below throughout 120°. Those who have had laser photocoagulation should report this to the DVLA so that appropriate visual field perimetry can be performed.

Healthy people with diabetes who are treated with diet or tablets are normally allowed to drive vehicles larger than those defined above (that is, Group 2), provided they pass a separate test to meet the higher Group 2 standards. Group 2 or vocational licences include those for large group vehicles and passenger carrying vehicles. Those taking insulin are not permitted to hold Group 2 licences, unless the license was granted before 1991 and subject to satisfactory medical reports, because of the serious potential consequences of hypoglycaemia, no matter how small the risk may be for any



Hazards when driving

individual diabetic patient. However, exceptional cases, based on individual assessment of the risk of experiencing diminished awareness of hypoglycaemia, can be identified (see Appendix 1) and enable some people to hold C1 licences (vehicles between 3500 and 7500 kg), although D1 vehicles (minibuses) do not fall into this category.

All insulin-treated diabetic patients who drive should always keep a supply of sugar in their cars. They should normally check blood glucose before driving, and should not drive if they are late for a meal, when the danger of hypoglycaemia is very great, especially around noon. If they experience warning symptoms of hypoglycaemia they should stop, switch off the ignition, and preferably leave the car, since they may otherwise be open to the charge of driving under the influence of drugs (insulin). Those unfortunate patients who are prone to hypoglycaemia without warning must not drive, and their doctors should make this advice very clear.

Insurance and pension

Driving insurance with a normal premium will usually be issued, subject to a satisfactory medical report. Life assurance premiums are often raised by amounts which depend on the result of a medical examination. It is worth looking for the "best buy". Sickness and holiday insurance premiums are often higher than normal. Diabetes UK offers helpful advice on insurance.

Travel

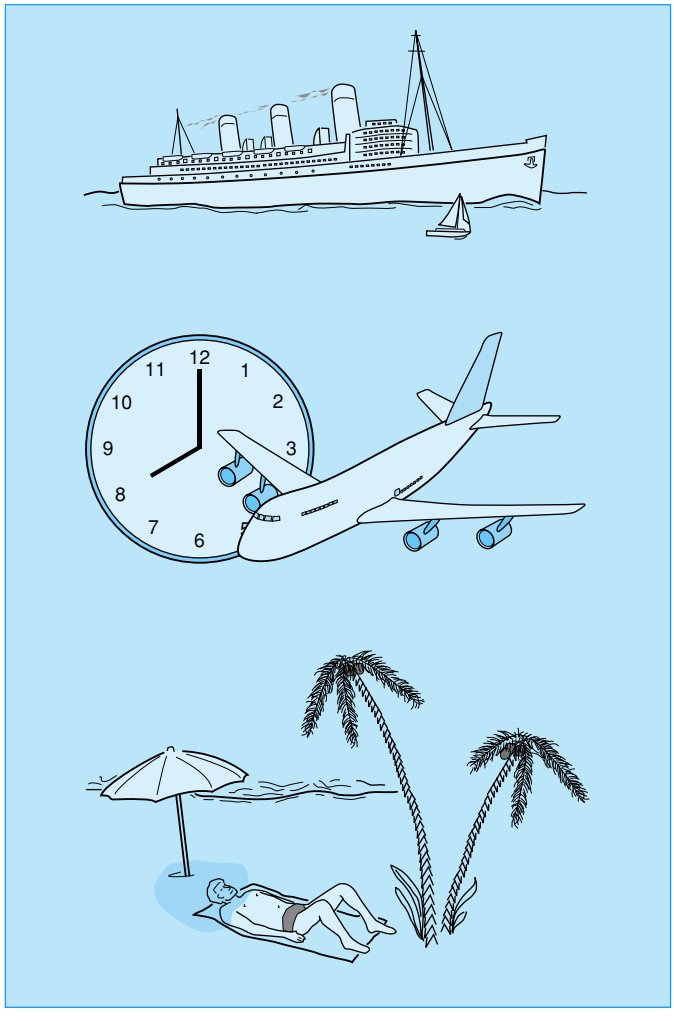
Diabetic control is easily upset by the rigours of travelling. People with diabetes should therefore undertake regular blood tests and adjust diet and insulin if necessary. Ideally those on insulin should travel with a companion if they are going to remote places. It is well worth carrying an identity disk which can be obtained from Diabetes UK, which also helps with appropriate foreign language leaflets. Some of the following circumstances present special problems.

Sea sickness and other stomach upsets causing vomiting
Diabetic patients may use the same anti-seasickness tablets as non-diabetics; these drugs do not change diabetic control. They do however tend to cause drowsiness, so it is best not to drive. If vomiting occurs insulin should be continued without fail and the situation dealt with as described on page 37.

Time changes on long-distance air travel
There are inevitably difficulties with diabetic control for a few days.

Flying west
The time between injections can, with little problem, be lengthened by two to three hours twice daily. Regular blood tests should be performed and if they are excessively high (around 15 mmol/l or more) extra soluble insulin (between 4 and 8 units) can be taken. If the time gap between injections is lengthened still further, a small supplementary injection of soluble insulin (between 4 and 8 units) is taken between the usual injections.

Flying east
The time between injections will need to be shortened by two to three hours each time, which could result in rather low blood glucose readings. Careful testing should be performed, and if required each dose can be reduced by a small amount



Hazards when travelling (sea sickness, time changes, burns on feet)

Other problems while abroad

- Vomiting either from motion sickness or stomach upsets (see page 37)
- Intercurrent illnesses affecting diabetic control, for example infections (see page 37)
- Hypoglycaemia (see page 34)
- Alterations of diabetic control due to major changes in diet or activity
- Burning feet on hot sand or stones, making foot protection with sandals or trainers extremely important

(4 to 8 units on average). Regular meals should be taken as normal. Many airlines will make special provision for those with diabetes if notified in advance; it is nevertheless strongly advisable to carry a food pack in case of delays or other emergencies.

Physical activity
More insulin may be needed if those with diabetes decide to be much less active when on holiday, or vice versa. Dietary indiscretions may also play havoc with control.

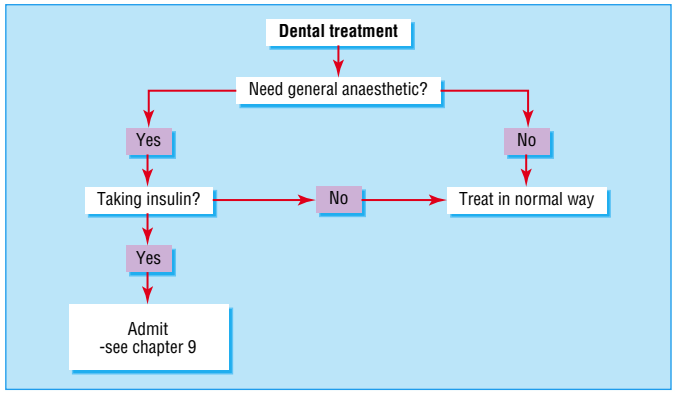
Breakage or loss of equipment
People with diabetes should carry ample supplies of syringes, insulin, needles, and testing equipment, and it is wise for a travelling companion to have a second set. The equipment in current use and that for emergency use is best kept in separate places. Soluble and isophane insulins are obtainable in most countries.

Storage of insulin
In a temperate climate insulin will keep for some months at room temperature (and furthermore injections sting less if the insulin is not chilled). Refrigeration is wise for prolonged visits to a tropical climate, and is also recommended for stocks kept at home for long periods, although exposed insulin actually deteriorates very little. Insulin should never be deep frozen and should not be left in the luggage hold of an aircraft, where it may freeze. Insulin is not damaged when passing through scanners.

Vaccination and inoculation
These are quite suitable for diabetic people and should be given for the same indications as for non-diabetics.

Dental treatment

Diabetic people may receive dental treatment in the normal way and without any special arrangements, except in the case of insulin treated patients needing a general anaesthetic, when a short admission to hospital is the wisest course. Dental infections and abscesses can of course upset diabetic control as is the case with any other source of infection.



Dental treatment

The story of Mrs B-J concluded

I had started insulin only ten years after its discovery, but I remember meeting an elderly man in the upstairs waiting room by the path lab telling us that he had become diabetic before insulin, and how he thanked God for it every day. I know how he felt, but sadly his prophecy, that diabetes would be treated only like a cold in a further ten years, was not fulfilled.

During her early years Mrs B-J was treated with a variety of twice daily insulin regimens changing with the fashions of time and the whims of both herself and her physicians. The insulin dose in 1960 was around 26 units daily and in 2001 it is still approximately 24 units daily. Her HbA_{1c} has ranged between 9% and 10% between 1995 and 2001-2.

At the time of writing early in 2002, Mrs B-J, now nearly 80 years old and with diabetes of nearly 70-years duration, still regularly attends the diabetic clinic at King's College Hospital and has done so without a break since 1932. Her medical records from 1932 are complete, and correspondence dates back to 1966.

Apart from the recent development of infection in one great toe, it is remarkable to observe that after all these years, she has no retinopathy (1999) and only mild lens opacities; there has never been any proteinuria and an absence of microalbuminuria was noted in 1999.

We must congratulate Mrs B-J for her courage and perseverance over so many years. We are grateful now for the privilege of being able to read the very personal account of her own diabetes which must give so much encouragement to others.

Appendix 1

Questionnaire to assess diminished awareness of hypoglycaemia

This questionnaire is recommended by the Driver and Vehicle Licensing Agency for applicants for certain driving licences as indicated on page 88 and is also of value in assessing patients involved in other hazardous activities or occupations. The applicant should be assessed by a consultant specialising in diabetes.

1 Please give details of medical supervision for diabetes.
- Date of interview
- Date of previous attendance
- Date of diagnosis of diabetes
- Date insulin treatment commenced

2 Are you satisfied that the applicant:
(a) knows what symptoms can occur as a consequence of hypoglycaemia?
(b) can recognise these symptoms if they occur?
(c) can take appropriate action?

3 Has the applicant, to your knowledge, experienced hypoglycaemia while driving within the last 12 months, which required assistance from another person?
If yes please give details/date(s):

4 Is there evidence of impaired awareness of hypoglycaemia in the past 12 months, during waking hours?
If yes please give details/date(s):

5 Is there a history of hypoglycaemia during waking hours in the last 12 months requiring assistance from a third party?
If yes please give details/date(s):

6 Does the applicant have a very clear understanding of diabetes and the necessary precautions for safe driving?

7 Does the applicant always carry an accessible supply of carbohydrate in the vehicle?

8 Does the applicant undertake blood glucose monitoring at least TWICE daily and at times relevant to driving on their current entitlement?

9 Have you examined the applicant's blood glucose records for the past three months?
If no please explain why this was not done:

10 Are you satisfied with the accuracy of the results?
If no, please explain why not:

11 Is there evidence of biochemical hypoglycaemia without symptoms (blood glucose below 3·0 mmol/l) on routine testing?
If yes please give details:

12 Is there any diabetic complication or other medical condition that could affect safe driving?

Further information

Diabetes associations for patients and health professionals

American Diabetes Association (Patient and Professional),
1660 Duke Street,
Alexandria,
Virginia VA 22314,
USA.
Tel: 001-703-549-1500
Fax: 001-703-549-6995

Australian Diabetes Society (Professional),
145 Macquarie Street,
Sydney, NSW 2000,
Australia.
Tel: 0061-9256-5462
Fax: 0061-9251-8174
<www.racp.edu.au> e-mail <sneylon@racp.edu.au>

Diabetes UK (Patient and Professional),
10 Parkway,
London NW1 7AA,
UK.
Tel: 0044-20-7424-1000
<www.diabetes.org.uk>

Canadian Diabetes Association (Patient and Professional),
PO Box 12013,
Station BRM B,
Toronto, Ontario, M7Y 2L3,
Canada.
<www.diabetes.ca>

Diabetes Australia (Patient),
1st Floor Churchill House,
218 Northbourne Avenue,
Braddon ACT 2612,
Australia.
<www.diabetesaustralia.com.au>

Diabetes New Zealand (Patient),
PO Box 54, 1 Conquest Street,
Oamaru,
New Zealand.
<www.diabetes.org.nz> e-mail <info@diabetes.org.nz>

European Association for Study of Diabetes (Professional),
Rheindorfer Weg 3, D-40591,
Dusseldorf,
Germany.
Tel: 0049-211-7584690
Fax: 0049-211-75846929
<www.easd.org>
e-mail <easd@uni-dusseldorf.de>

Diabetes Federation of Ireland (Patient),
76 Lower Gardiner Street,
Dublin 1,
Ireland.
Tel and fax: 00353-1836-3022

Juvenile Diabetes Research Foundation International,
120 Wall Street,
New York,
NY 10005-4001,
USA.
Tel: 001-212-785-9595
<www.jdrf.org/index.php>

Juvenile Diabetes Research Foundation,
19, Angel Gate,
London, ECIV 2PT,
UK.
Tel: 0044-20-7713-2030
Fax: 0044-20-7713-2031
<www.jdrf.org.uk>

National Diabetes International Clearing House,
Box NDIC,
1 Information Way,
Bethesda,
MD 20892-3560,
USA.

New Zealand Society for the Study of Diabetes (Professional),
East Riding,
Whiterocks Road,
6-D RD Oamaru,
New Zealand.
Tel and fax: 0064-343-48110

Society for Endocrinology, Metabolism and Diabetes of South Africa (Professional),
PO Box 783155,
Sandton 2146,
Johannesburg,
South Africa.
Tel: 0027-11/202-0500
Fax: 0027-11/807-7989
e-mail <rsh@novonordisk.com>

South African Diabetes Association (Patient),
PO Box 3943,
Cape Town 8000,
South Africa.
<www.sada.org.za>

Useful websites

Clinical Standards for Diabetes (Scotland)	<www.clinicalstandards.org>
Exeter Genetic Screening Service	<www.diabetesgenes.org>
Heart Protection Study	<www.hpsinfo.org>
National Service Framework for Diabetes	<www.doh.gov.uk/nsf/diabetes>
Scottish Intercollegiate Guidelines Network (SIGN)	<www.sign.ac.uk>
Warwick Centre for Diabetes Education and Research	<www.diabetescare.warwick.ac.uk>

Further reading

Alberti KGMM, Defronzo RA, Keen H, Zimmet P, eds. *International Textbook of Diabetes Mellitus, 2nd ed.* Chichester: John Wiley & Sons Ltd, 1997. [A two volume comprehensive review.]

Amiel S. Is there anything new about insulin therapy? In: Amiel S, ed. *Horizons in Medicine.* Royal College of Physicians of London, 2002.

Angel A, Dhalla N, Pierce G, Singal P. *Diabetes and cardiovascular disease.* Netherlands: Kluwer Academic Publishers, 2001.

Besser GM, Bodansky HJ, Cudworth AG. *Clinical Diabetes: An Illustrated Text.* Philadelphia: JB. Lippincott, 1988. [An extensively illustrated text.]

Bilous R. HOPE and other recent trials of antihypertensive therapy in Type 2 diabetes. In: Amiel S, ed. *Horizons in Medicine.* Royal College of Physicians of London, 2002.

Bliss M. *The Discovery of Insulin.* Edinburgh: Paul Harris Publishing, 1983. [A masterful account of the events surrounding the discovery of insulin.]

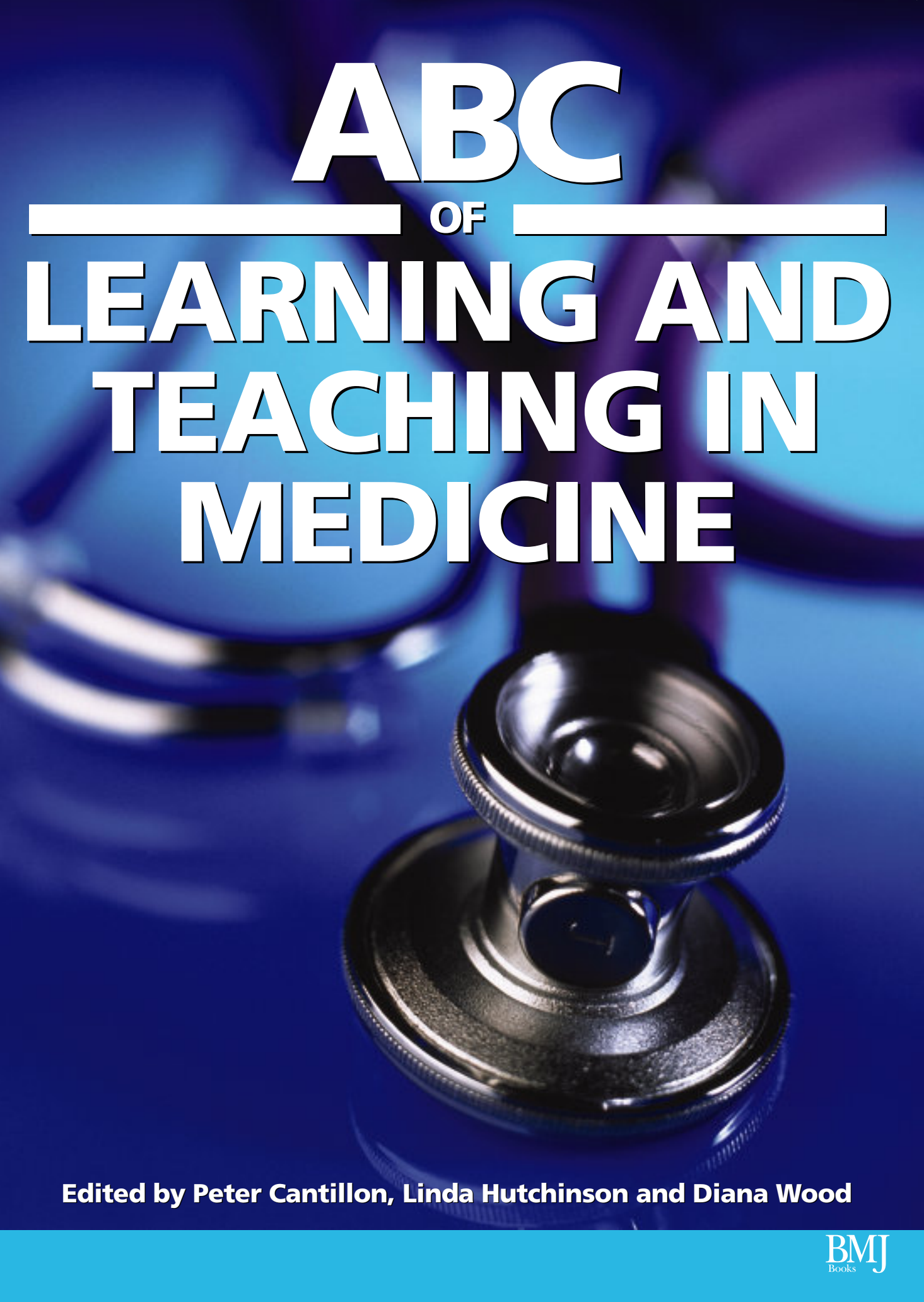
Bloom A, Ireland J. (with revisions by Watkins PJ). *A Colour Atlas of Diabetes, 2nd ed.* London: Wolfe Publishing Ltd, 1992.
Boulton AJM, Connor H, Cavanagh PR, eds. *The Foot in Diabetes, 3rd ed.* Chichester: John Wiley & Sons Ltd, 2000.
Bowker JH, Pfeifer MA, eds. *Levin and O'Neals the diabetic foot.* St Louis: Mosby, 2001.
Day JL. *Living with diabetes, 2nd ed.* Chichester: John Wiley & Sons Ltd, 2001.
Dornhorst A, Hadden DR, eds. *Diabetes and pregnancy.* Chichester: John Wiley & Sons, 1996.
Dyck PJ, Thomas PK, eds. *Diabetic Neuropathy, 2nd ed.* Philadelphia: WB Saunders Co, 1999.
Edmonds ME, Foster AVM. *Managing the diabetic foot.* Oxford: Blackwell Science, 2000.
Ekoe JM, Zimmet P, Williams R. *Epidemiology of diabetes mellitus.* Chichester: John Wiley & Sons Ltd, 2001.
Fox C, Pickering A. *Diabetes in the real world.* London: Class Publishing, 1995. [A practical handbook advising on the needs of diabetic patients in general practice].
Frier BM, Fisher M, eds. *Hypoglycaemia.* London: Edward Arnold, 1993. [The most authoritative reference book on the subject at present.]
Frier BM, Fisher M, eds. *Hypoglycaemia in clinical diabetes.* Chichester: John Wiley & Sons Ltd, 1999.
Gill G, Pickup J, Williams G, eds. *Difficult diabetes.* Oxford: Blackwell Science, 2001.
Gill G, Mbanya J-C, Alberti G, eds. *Diabetes in Africa.* Cambridge: FSG Communications Ltd, Reach, 1997.
Hasslacher C. *Diabetic nephropathy.* Chichester: John Wiley & Sons Ltd, 2001.
Hollingsworth DR. *Pregnancy, Diabetes and Birth: a Management Guide, 2nd ed.* London: Williams and Wilkins Ltd, 1991.
Jeffcoate W, MacFarlane R. *The Diabetic Foot: An Illustrated Guide to Management.* London: Chapman and Hall, 1995.
Kahn CR, Weir GC, eds. *Joslin's Diabetes Mellitus.* Philadelphia: Lea and Febiger, 1995.
Kelmark CJH. *Childhood and Adolescent Diabetes.* London: Chapman and Hall, 1994.
Kirby R, Holmes S, Carson C. *Erectile dysfunction, 2nd ed.* Oxford: Health Press, 1999.
Krentz AJ, Bailey CJ. *Type 2 diabetes in practice.* London: Royal Society of Medicine Press Ltd, 2001.
Leslie RDG, Robbins DC, eds. *Diabetes: Clinical Science in Practice.* Cambridge: Cambridge University Press, 1995.
MacKinnon M. *Providing Diabetes Care General Practice, 4th ed.* London: Class Publishing, 2002.
Malins JM. *Clinical Diabetes Mellitus.* London: Eyre and Spottiswoode, 1968. [A splendid clinical description of diabetes.]
Mathias CJ, Bannister R, eds. *Autonomic Failure, 4th ed.* Oxford: University Press, Oxford, 1999.
Mogensen CE, ed. *The Kidney and Hypertension in Diabetes Mellitus, 5th ed.* USA: Kluwer Academic Publishing, 2000.
Nattrass M. *Malins' Clinical Diabetes, 2nd ed.* London: Chapman and Hall, 1994.
Page S, Hall G. *Diabetes: emergency and hospital management.* London: BMJ Publishing Group, 1999.
Pickup JC, Williams G, eds. *Textbook of Diabetes, 2nd ed.* Oxford: Blackwell Scientific Publications, 1997. [A substantial account in two volumes; beautifully illustrated.]
Pickup JC, Williams G. *Chronic Complications of Diabetes.* Oxford: Blackwell Scientific Publications, 1994. [A one volume updated version taken from the authors' two volume textbook.]
Ritz R, Rychlik I, eds. *Nephropathy in Type 2 diabetes.* Oxford: Oxford University Press, 1999.
Shaw KM, ed. *Diabetic complications.* Chichester: John Wiley and Sons, 1996.
Sinclair A, Finucane P. *Diabetes in old age, 2nd ed.* Chichester: John Wiley & Sons Ltd, 2000.
Tooke J, ed. *Diabetic angiopathy.* London: Arnold, 1999.
Warren S, Le Compte PM, Legg MA. *The Pathology of Diabetes Mellitus.* London: Henry Kimpton, 1966. [A classic account.]
Watkins PJ. The diabetic traveller. In: Dawood R, ed. *Travelers' Health.* Random House, New York, 2002.
Watkins PJ, Amiel S, Howell SL, Turner E. *Diabetes and its management, 6th ed.* Oxford: Blackwell Science, 2003.
West KM. *Epidemiology of Diabetes and its Vascular Lesions.* New York: Elsevier, 1978. [A classic review of diabetes in a world setting.]
Williams R, Wareham N, Kinmonth AM, Herman WH. *Evidence base for diabetes care.* Chichester: John Wiley & Sons Ltd, 2002.
Young A, Harries M, eds. *Physical activity for patients: an exercise prescription.* London: Royal College of Physicians, 2001.
United Kingdom Prospective Diabetes Survey (UKPDS). Original publications in *BMJ* and *Lancet*, September 1998.

Journal references

Atkinson MA, Eisenbarth GS. Type 1 diabetes: new perspectives on disease pathogenesis and treatment. *Lancet* 2001;358:221-9.
Barnett AH, Eff C, Leslie RDG, Pyke DA. Diabetes in identical twins: a study of 200 pairs. *Diabetologia* 1981;20:87-93.
Barrett T. Inherited diabetic disorders. CME Bulletin Endocrinology and Diabetes 1999;2:47-50.
DCCT: Diabetes Control and Complications Research Group. The effect of diabetes on the development and progression of long-term complications in insulin dependent diabetes. *New Engl J Med* 1993;329:977-86.
DIGAMI Study Group. Prospective randomised study of intensive insulin treatment on long-term survival after acute myocardial infarction in patients with diabetes mellitus. *BMJ* 1997;314:1512-15.
Euclid Study Group. Randomised placebo-controlled trial of lisinopril in normotensive patients with insulin dependent diabetes and normoalbuminuria or microalbuminuria. *Lancet* 1997;349:1787-92.
Expert Committee on the Diagnosis and Classification of Diabetes mellitus. Report of the Expert Committee on the Diagnosis and Classification of Diabetes. *Diabetes Care* 1997;20:1183-97.
Heller SR. Sudden cardiac death in young people with diabetes. *CME Bulletin Endocrinology and Diabetes*, 2000;3:4-7.
HOPE study: Heart Outcomes Prevention Evaluation Study Investigators. Effects of ramipril on cardiovascular and microvascular outcomes in people with diabetes mellitus. *Lancet* 2000;355:253-9.
Hoppener JWM, Ahren B, Lips CJM. Islet amyloid and Type 2 diabetes. *New Engl J Med* 2000;343:411-19.
MRC/BHF Heart protection study of cholesterol lowering with simvastatin in 20536 high-risk individuals: a randomised controlled trial. *Lancet* 2002;360:7-8
Owens DR, Zinman B, Bolli GB. Insulins today and beyond. *Lancet* 2001;358:739-46.
Ritz E, Orth SR. Nephropathy in patients with Type 2 diabetes. *New Engl J Med* 1999;341:1127-33.

Index

Page numbers in *italic* type refer to material in boxes or illustrations


ABC
OF
LEARNING AND
TEACHING IN
MEDICINE
Edited by Peter Cantillon, Linda Hutchinson and Diana Wood
BMJ
Books

ABC OF
LEARNING AND TEACHING IN MEDICINE

ABC OF LEARNING AND TEACHING IN MEDICINE

Edited by

PETER CANTILLON
Senior lecturer in medical informatics and medical education, National University of Ireland, Galway, Republic of Ireland

LINDA HUTCHINSON
Director of education and workforce development and consultant paediatrician, University Hospital Lewisham

and

DIANA WOOD
Director of medical education, University of Cambridge School of Clinical Medicine, Addenbrookes Hospital, Cambridge

BMJ Books

ЗНАНИЕ

First published in 2003 by
BMJ Publishing Group Ltd, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 07279 16785

Typeset by BMJ Electronic Production
Printed and bound in Spain by GraphyCems, Navarra
Cover Image shows a stethoscope for listening to sounds within the body.
With permission from Colin Cuthbert/Science Photo Library

Contents

Contributors

Peter Cantillon
Senior lecturer in medical informatics and medical education, National University of Ireland, Galway, Republic of Ireland

Richard Farrow
Director of problem based learning at the Peninsula Medical School at the Universities of Exeter and Plymouth

Jill Gordon
Associate professor in the department of medical education at the University of Sydney, Australia

Linda Hutchinson
Director of education and workforce development and consultant paediatrician, University Hospital Lewisham, London

David Jaques
Independent consultant in learning and teaching in higher education

Carol Jollie
Project officer in the skills enhancement project for the Camden Primary Care Trust at St Pancras Hospital, London

David M Kaufman
Director of the Learning and Instructional Development Centre at Simon Fraser University, Burnaby, British Columbia, Canada

Judy McKimm
Head of curriculum development at Imperial College School of Medicine, London and an educational consultant

Jill Morrison
Professor of general practice and deputy associate dean for education at Glasgow University

John J Norcini
President and chief executive officer of the Foundation for Advancement of International Medical Education and Research, Philadelphia, Pennsylvania

David Prideaux
Professor and head of the Office of Medical Education in the School of Medicine at Flinders University, Adelaide, Australia

Lambert W T Schuwirth
Assistant professor in the department of educational development and research at the University of Maastricht in the Netherlands

Sydney Smee
Manager of the Medical Council of Canada's qualifying examination part II, in Ottawa, Canada

John Spencer
General practitioner and professor of medical education in primary health care at the University of Newcastle upon Tyne

Cees P M van der Vleuten
Professor and chair in the department of educational development and research at the University of Maastricht in the Netherlands

Diana Wood
Director of medical education, University of Cambridge School of Clinical Medicine, Addenbrookes Hospital, Cambridge

Preface

Although we would never allow a patient to be treated by an untrained doctor or nurse, we often tolerate professional training being delivered by untrained teachers. Traditionally students were expected to absorb most of their medical education by attending timetabled lectures and ward-rounds, moving rapidly from one subject to the next in a crowded curriculum. Our junior doctors learnt by watching their seniors in between endless menial tasks. In recent years the importance of active, self directed learning in higher education has been recognised. Outcome led structured programmes for trainees are being developed in the face of reduced working hours for both the learners and teachers. These all present new challenges for teachers in medicine of all levels of seniority.

Throughout the world there is great interest in developing a set of qualifications for medical teachers, both at the elementary "teaching the teacher" level and as part of progressive modular programmes leading to formal certification. In addition to acquiring new qualifications and standards, teachers also need access to literature resources that describe essential components in medical education and supply tips and ideas for teaching.

This ABC began as an expressed wish of the BMJ to publish an introductory and accessible text on medical education. It grew into a book covering the more generic topics of learning and teaching in medicine with the aim of illustrating how educational theory and research underpins the practicalities of teaching and learning. The editors invited an international group of authors on the basis of their acknowledged expertise in the particular topics assigned to them. Each chapter was edited and illustrated to ensure maximum accessibility for readers and subsequently peer reviewed by two educational experts. Their suggestions have been incorporated into the finished book.

The *ABC of Learning and Teaching in Medicine* would not have been possible without the tireless support of BMJ editorial staff, Julia Thompson, Eleanor Lines, Sally Carter, and Naomi Wilkinson. We would also like to thank Professor Paul O' Neill and Dr Ed Peile for their excellent and timely peer reviews for each of the chapters. Finally we would very much welcome comments and suggestions about this ABC from its most important reviewers, you the readers.

PC, DW, LH

1 Applying educational theory in practice

David M Kaufman

How many times have we as teachers been confronted with situations in which we really were not sure what to do? We "flew by the seat of our pants," usually doing with our learners what had been done with us. It would be useful to be able to turn to a set of guiding principles based on evidence, or at least on long term successful experience.

Fortunately, a body of theory exists that can inform practice. An unfortunate gap between academics and practitioners, however, has led to a perception of theory as belonging to an "ivory tower" and not relevant to practice. Yet the old adage that "there is nothing more practical than a good theory" still rings true today. This chapter describes several educational theories and guiding principles and then shows how these could be applied to three case studies relating to the "real world."

Adult learning theory

Malcolm Knowles introduced the term "andragogy" to North America, defining it as "the art and science of helping adults learn." Andragogy is based on five assumptions—about how adults learn and their attitude towards and motivation for learning.

Knowles later derived seven principles of andragogy. Most theorists agree that andragogy is not really a theory of adult learning, but they regard Knowles' principles as guidelines on how to teach learners who tend to be at least somewhat independent and self directed. His principles can be summarised as follows:

- Establish an effective learning climate, where learners feel safe and comfortable expressing themselves
- Involve learners in mutual planning of relevant methods and curricular content
- Involve learners in diagnosing their own needs—this will help to trigger internal motivation
- Encourage learners to formulate their own learning objectives—this gives them more control of their learning
- Encourage learners to identify resources and devise strategies for using the resources to achieve their objectives
- Support learners in carrying out their learning plans
- Involve learners in evaluating their own learning—this can develop their skills of critical reflection.

Self directed learning

Self directed learning can be viewed as a method of organising teaching and learning in which the learning tasks are largely within the learners' control (as with the adult learning principles described above).

It can also be viewed as a goal towards which learners strive so that they become empowered to accept personal responsibility for their own learning, personal autonomy, and individual choice. Success in the first view would lead to attaining the second.

Philip Candy identified in the literature about 100 traits associated with self direction, which he synthesised as the ability to be methodical and disciplined; logical and analytical; collaborative and interdependent; curious, open, creative, and motivated; persistent and responsible; confident and competent at learning; and reflective and self aware.

Andragogy—five assumptions about adult learning

- Adults are independent and self directing
- They have accumulated a great deal of experience, which is a rich resource for learning
- They value learning that integrates with the demands of their everyday life
- They are more interested in immediate, problem centred approaches than in subject centred ones
- They are more motivated to learn by internal drives than by external ones

Learners need to feel safe and comfortable expressing themselves

Self directed learning

- Organising teaching and learning so that learning is within the learners' control
- A goal towards which learners strive so that they become able to accept responsibility for their own learning

How do we develop these traits in our learners? Most importantly, learners must have the opportunity to develop and practise skills that directly improve self directed learning. These skills include asking questions, critically appraising new information, identifying their own knowledge and skill gaps, and reflecting critically on their learning process and outcomes.

Learners should identify their own knowledge gaps and critically appraise new information

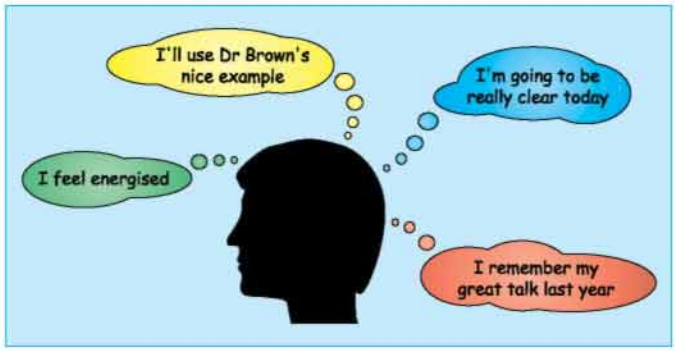
Self efficacy

According to Albert Bandura, people's judgments of their own ability to deal with different situations is central to their actions. These actions include what they choose to do, how much effort they invest in activities, how long they persist in the face of adversity, and whether they approach the tasks anxiously or assuredly.

These judgments, called "self efficacy," may or may not be accurate, but they arise from four main information sources. In decreasing order of their strength, these sources are: performance attainments, observations of other people, verbal persuasion, and physiological state. Successes raise our self efficacy, while failures lower it. Failures are particularly likely to lower our self efficacy if they occur early in the learning process and are not due to lack of effort or difficult situations.

Observing other people similar to us performing successfully can strengthen our beliefs that we can perform similar tasks, especially when the tasks are unfamiliar. Verbal persuasion from a credible source also can help.

Finally, we (both teachers and learners) need to re-interpret our anxiety or nervousness in difficult situations as excitement or anticipation, rather than as an ominous sign of vulnerability.



Teachers and learners need to view any anxiety or nervousness in difficult situations as excitement or anticipation

Self efficacy—roles for the teacher

- Modelling or demonstration
- Setting a clear goal or image of the desired outcome
- Providing basic knowledge and skills needed as the foundation for the task
- Providing guided practice with corrective feedback
- Giving students the opportunity to reflect on their learning

Constructivism

Constructivism has important implications for teaching and learning. Firstly, the teacher is viewed not as a transmitter of knowledge but as a guide who facilitates learning. Secondly, as learning is based on prior knowledge, teachers should provide learning experiences that expose inconsistencies between students' current understandings and their new experiences. Thirdly, teachers should engage students in their learning in an active way, using relevant problems and group interaction. Fourthly, if new knowledge is to be actively acquired, sufficient time must be provided for in-depth examination of new experiences.

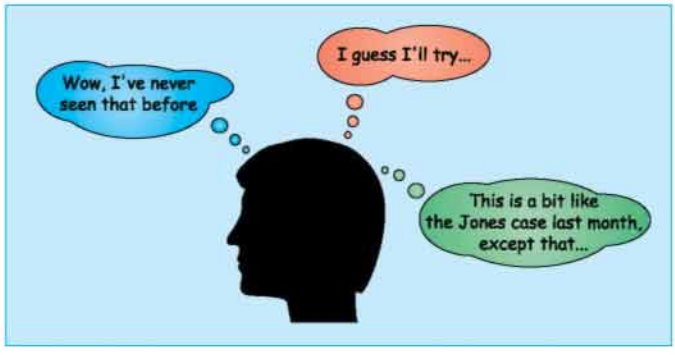
The primary idea of constructivism is that learners "construct" their own knowledge on the basis of what they already know. This theory posits that learning is active, rather than passive, with learners making judgments about when and how to modify their knowledge

Reflective practice

The theory of reflective practice is attributed primarily to Donald Schön, whose work is based on the study of a range of professions. He argues that formal theory acquired through professional preparation is often not useful to the solution of the real life "messy, indeterminate" problems of practice.

Schön labels professionals' automatic ways of practising as professional "zones of mastery"—that is, areas of competence. Unexpected events or surprises trigger two kinds of reflection.

The first, "reflection in action," occurs immediately. It is the ability to learn and develop continually by creatively applying current and past experiences and reasoning to unfamiliar events while they are occurring. The second, "reflection on action," occurs later. It is a process of thinking back on what happened in a past situation, what may have contributed to the unexpected event, whether the actions taken were appropriate, and how this situation may affect future practice.



"Reflection in action"

Through the process of reflecting both "in practice" and "on practice," practitioners continually reshape their approaches and develop "wisdom" or "artistry" in their practice. Activities such as debriefing with peers or learners, seeking feedback from learners on a regular basis, and keeping a journal can provide vehicles for reflective practice.

Converting theory into practice

Each of the educational theories presented here can guide our teaching practices. Some theories will be more helpful than others in particular contexts. However, several principles also emerge from these theories, and these can provide helpful guidance for medical educators.

Three cases studies
The boxes (right) describe three "real world" case studies representing situations encountered in medical education settings. The educational theories described above, and the principles which emerge from them, can guide us in solving the problems posed in these three cases.

Case 1 solution
You could present an interactive lecture on the autonomic nervous system. You could distribute a notetaking guide. This would contain key points, space for written notes, and two key multiple choice or "short answer" questions requiring higher level thinking (principle 1, see box above). You could stop twice during the lecture and ask the students to discuss their response to each question with their neighbours (principles 1, 3, and 5). A show of hands would determine the class responses to the question (checking for understanding) and you could then give the correct answer (principle 5). Finally, you could assign a learning issue for the students to research in their own time (principle 4).

Case 2 solution
You could assign the students to small groups of four to six, and ask each group to submit two case studies describing clinical ethics issues in their local hospitals (principles 1 and 2). The ethics theory and approach needed to analyse these cases could be prepared by experts and presented on a website in advance of the sessions (principles 4, 5). The first of the six blocks of two hours could be used to discuss the material on the website and clarify any misunderstandings (principle 5). You could then show the students how to work though a case, with participation by the class (principle 7). The other five blocks could then be used for each small group to work through some of the cases prepared earlier, followed by a debriefing session with the whole class (principles 5 and 6).

Case 3 solution
You could first invite the registrar to observe you with patients, and do a quick debrief at the end of the day (principles 2, 6, and 7). With help from you, she could then develop her own learning goals, based on the certification requirements and perceived areas of weakness (principles 1, 3, and 4). These goals would provide the framework for assessing the registrar's performance with patients (principles 5, 6). You could observe and provide feedback (principle 5). Finally, the registrar could begin to see patients alone and keep a journal (written or electronic) in which she records the results of "reflection on practice" (principle 6). She could also record in her journal the personal learning issues arising from her patients, could conduct self directed learning on these, and could document

Seven principles to guide teaching practice

1 The learner should be an active contributor to the educational process
2 Learning should closely relate to understanding and solving real life problems
3 Learners' current knowledge and experience are critical in new learning situations and need to be taken into account
4 Learners should be given the opportunity and support to use self direction in their learning
5 Learners should be given opportunities and support for practice, accompanied by self assessment and constructive feedback from teachers and peers
6 Learners should be given opportunities to reflect on their practice; this involves analysing and assessing their own performance and developing new perspectives and options
7 Use of role models by medical educators has a major impact on learners. As people often teach the way they were taught, medical educators should model these educational principles with their students and junior doctors. This will help the next generation of teachers and learners to become more effective and should lead to better care for patients

Case 1: Teaching basic science

You have been asked to give a lecture on the autonomic nervous system to a first year medical class of 120 students. This has traditionally been a difficult subject for the class, particularly as it has not been explicitly covered by faculty in the problem based anatomy course. You wonder how you can make this topic understandable to the class in a 50-minute lecture.

Case 2: Ethics education

You are a member of a course committee in the department of internal medicine, which is charged with the task of integrating the topic of ethics into the third year medicine rotation. Your committee has been given six blocks of two hours over a 12 week period. You wonder how to make the material engaging, understandable, and useful to the students.

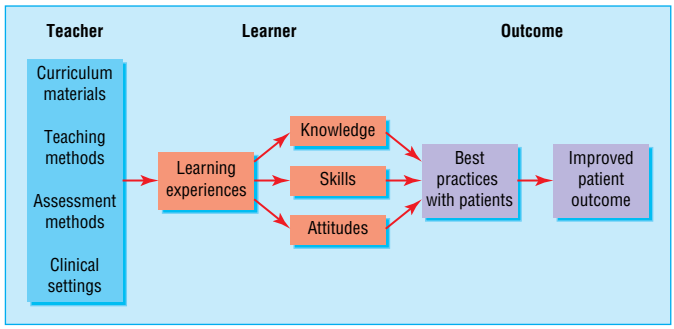
Case 3: General practice training

You are the trainer for a first year registrar in her first year of a general practice training programme. Your practice is so busy that you have very little time to spend with her. You wonder how you can contribute to providing a valuable learning experience for your trainee.

her findings in the journal (principles 1, 4, and 6). You could provide feedback on the journal (principle 5). If practical, the cohort of registrars could communicate via the internet to discuss their insights and experiences (principle 6).

Conclusions

This article has attempted to show how the gap between educational theory and practice can be bridged. By using teaching and learning methods based on educational theories and derived principles, medical educators will become more effective teachers. This will enhance the development of knowledge, skills, and positive attitudes in their learners, and improve the next generation of teachers. Ultimately, this should result in better trained doctors who provide an even higher level of patient care and improved patient outcomes.



From theory to practice

Further reading

- Bandura A. *Social foundations of thought and action: a social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
- Candy PC. *Self-direction for lifelong learning: a comprehensive guide to theory and practice*. San Francisco: Jossey-Bass, 1991.
- Kaufman DM, Mann KV, Jennett P. *Teaching and learning in medical education: how theory can inform practice*. London: Association for the Study of Medical Education, 2000. (Monograph.)
- Knowles MS and Associates. *Andragogy in action: applying modern principles of adult learning*. San Francisco: Jossey-Bass, 1984.
- Schön DA. *Educating the reflective practitioner: toward a new design for teaching and learning in the professions*. San Francisco: Jossey-Bass, 1987.

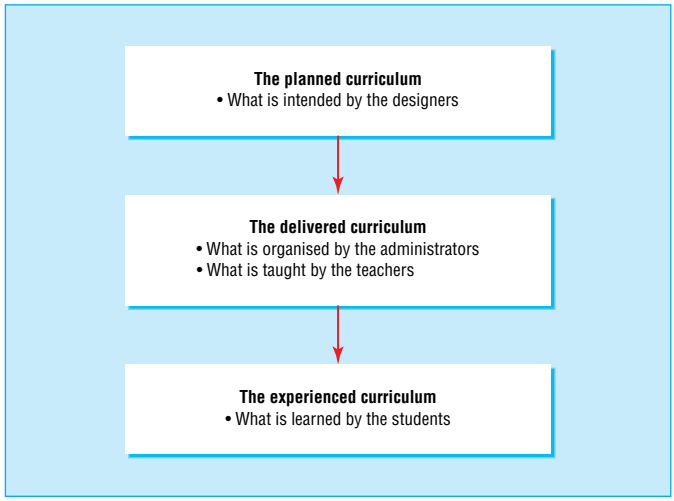
2 Curriculum design

David Prideaux

The curriculum represents the expression of educational ideas in practice. The word curriculum has its roots in the Latin word for track or race course. From there it came to mean course of study or syllabus. Today the definition is much wider and includes all the planned learning experiences of a school or educational institution.

The curriculum must be in a form that can be communicated to those associated with the learning institution, should be open to critique, and should be able to be readily transformed into practice. The curriculum exists at three levels: what is planned for the students, what is delivered to the students, and what the students experience.

A curriculum is the result of human agency. It is underpinned by a set of values and beliefs about what students should know and how they come to know it. The curriculum of any institution is often contested and problematic. Some people may support a set of underlying values that are no longer relevant. This is the so called sabretoothed curriculum, which is based on the fable of the cave dwellers who continued to teach about hunting the sabretoothed tiger long after it became extinct. In contemporary medical education it is argued that the curriculum should achieve a "symbiosis" with the health services and communities in which the students will serve. The values that underlie the curriculum should enhance health service provision. The curriculum must be responsive to changing values and expectations in education if it is to remain useful.



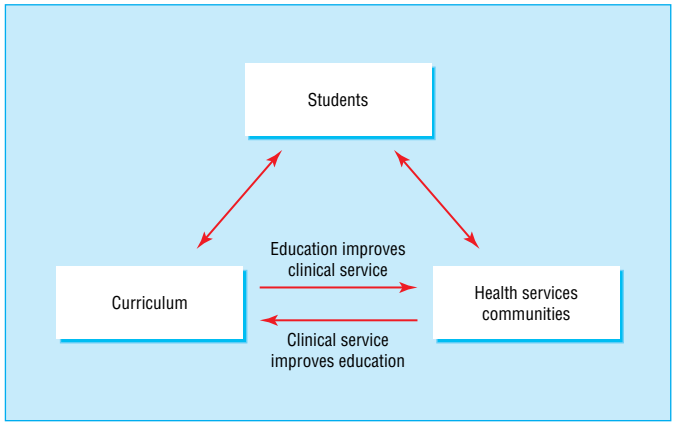
Three levels of a curriculum

Elements of a curriculum

If curriculum is defined more broadly than syllabus or course of study then it needs to contain more than mere statements of content to be studied. A curriculum has at least four important elements: content; teaching and learning strategies; assessment processes; and evaluation processes.

The process of defining and organising these elements into a logical pattern is known as curriculum design. Curriculum writers have tried to place some order or rationality on the process of designing a curriculum by advocating models.

There are two main types: prescriptive models, which indicate what curriculum designers should do; and descriptive models, which purport to describe what curriculum designers actually do. A consideration of these models assists in understanding two additional key elements in curriculum design: statements of intent and context.



"Symbiosis" necessary for a curriculum. From Bligh J et al (see "Further reading" box)

Curriculum models

Prescriptive models
- What curriculum designers should do
- How to create a curriculum

Descriptive models
- What curriculum designers actually do
- What a curriculum covers

Prescriptive models

Prescriptive models are concerned with the ends rather than the means of a curriculum. One of the more well known examples is the "objectives model," which arose from the initial work of Ralph Tyler in 1949. According to this model, four important questions are used in curriculum design.

The first question, about the "purposes" to be obtained, is the most important one. The statements of purpose have become known as "objectives," which should be written in terms of changed behaviour among learners that can be easily measured. This was interpreted very narrowly by some people and led to the specification of verbs that are acceptable and those that are unacceptable when writing the so called

Objectives model—four important questions*
- What educational purposes should the institution seek to attain?
- What educational experiences are likely to attain the purposes?
- How can these educational experiences be organised effectively?
- How can we determine whether these purposes are being attained?

*Based on Tyler R. *Basic principles of curriculum and instruction.* Chicago: Chicago University Press, 1949

"behavioural objectives." Once defined, the objectives are then used to determine the other elements of the curriculum (content; teaching and learning strategies; assessment; and evaluation).

This model has attracted some criticism—for example, that it is difficult and time consuming to construct behavioural objectives. A more serious criticism is that the model restricts the curriculum to a narrow range of student skills and knowledge that can be readily expressed in behavioural terms. Higher order thinking, problem solving, and processes for acquiring values may be excluded because they cannot be simply stated in behavioural terms. As a result of such criticism the objectives model has waned in popularity. The importance of being clear about the purpose of the curriculum is well accepted.

More recently, another prescriptive model of curriculum design has emerged. "Outcomes based education" is similar in many respects to the objectives model and again starts from a simple premise—the curriculum should be defined by the outcomes to be obtained by students. Curriculum design proceeds by working "backwards" from outcomes to the other elements (content; teaching and learning experiences; assessment; and evaluation).

The use of outcomes is becoming more popular in medical education, and this has the important effect of focusing curriculum designers on what the students will do rather than what the staff do. Care should be taken, however, to focus only on "significant and enduring" outcomes. An exclusive concern with specific competencies or precisely defined knowledge and skills to be acquired may result in the exclusion of higher order content that is important in preparing medical professionals.

Although debate may continue about the precise form of these statements of intent (as they are known), they constitute an important element of curriculum design. It is now well accepted that curriculum designers will include statements of intent in the form of both broad curriculum aims and more specific objectives in their plans. Alternatively, intent may be expressed in terms of broad and specific curriculum outcomes. The essential function of these statements is to require curriculum designers to consider clearly the purposes of what they do in terms of the effects and impact on students.

Descriptive models

An enduring example of a descriptive model is the situational model advocated by Malcolm Skilbeck, which emphasises the importance of situation or context in curriculum design. In this model, curriculum designers thoroughly and systematically analyse the situation in which they work for its effect on what they do in the curriculum. The impact of both external and internal factors is assessed and the implications for the curriculum are determined.

Although all steps in the situational model (including situational analysis) need to be completed, they do not need to be followed in any particular order. Curriculum design could begin with a thorough analysis of the situation of the curriculum or the aims, objectives, or outcomes to be achieved, but it could also start from, or be motivated by, a review of content, a revision of assessment, or a thorough consideration of evaluation data. What is possible in curriculum design depends heavily on the context in which the process takes place.

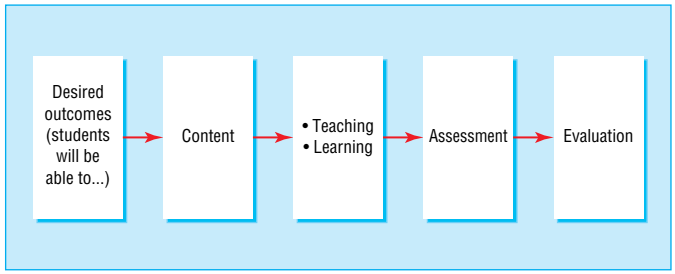
All the elements in curriculum design are linked. They are not separate steps. Content should follow from clear statements of intent and must be derived from considering external and internal context. But equally, content must be delivered by

Behavioural objectives*

Acceptable verbs	Unacceptable verbs
• To write	• To know
• To recite	• To understand
• To identify	• To really understand
• To differentiate	• To appreciate
• To solve	• To fully appreciate
• To construct	• To grasp the significance of
• To list	• To enjoy
• To compare	• To believe
• To contrast	• To have faith in

*From Davies I. *Objectives in curriculum design*. London: McGraw Hill, 1976

Clearly stated objectives provide a good starting point, but behavioural objectives are no longer accepted as the "gold standard" in curriculum design



Outcomes based curriculum (defining a curriculum "backwards"—that is, from the starting point of desired outcomes)

Example of statements of intent

Aim
- To produce graduates with knowledge and skills for treating common medical conditions

Objectives
- To identify the mechanisms underlying common diseases of the circulatory system
- To develop skills in history taking for diseases of the circulatory system

Broad outcome
- Graduates will attain knowledge and skills for treating common medical conditions
- Students will identify the mechanisms underlying common diseases of the circulatory system
- Students will acquire skills in history taking for diseases of the circulatory system

Situational analysis*

External factors	Internal factors
• Societal expectations and changes	• Students
• Expectations of employers	• Teachers
• Community assumptions and values	• Institutional ethos and structure
• Nature of subject disciplines	• Existing resources
• Nature of support systems	• Problems and shortcomings in existing curriculum
• Expected flow of resources	

*From Reynolds J, Skilbeck M. *Culture and the classroom*. London: Open Books, 1976

appropriate teaching and learning methods and assessed by relevant tools. No one element—for example, assessment—should be decided without considering the other elements.

Curriculum maps

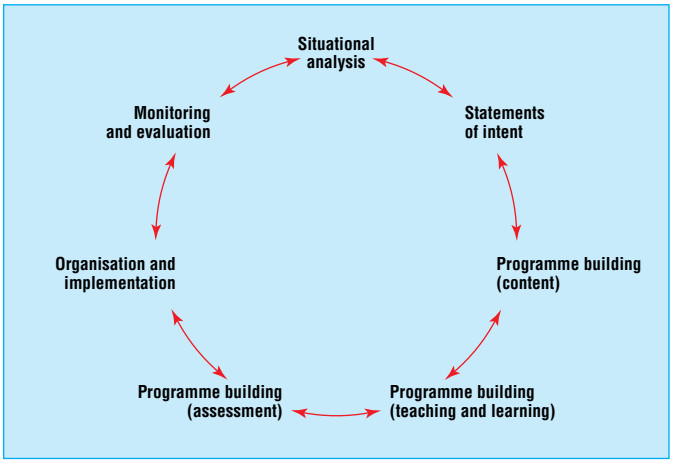
Curriculum maps provide a means of showing the links between the elements of the curriculum. They also display the essential features of the curriculum in a clear and succinct manner. They provide a structure for the systematic organisation of the curriculum, which can be represented diagrammatically and can provide the basis for organising the curriculum into computer databases.

The starting point for the maps may differ depending on the audience. A map for students will place them at the centre and will have a different focus from a map prepared for teachers, administrators, or accrediting authorities. They all have a common purpose, however, in showing the scope, complexity, and cohesion of the curriculum.

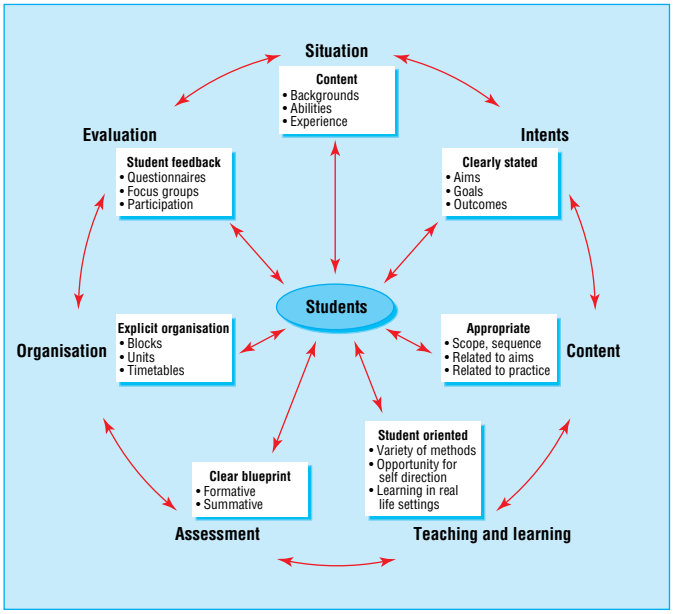
Curriculum maps with computer based graphics with "click-on" links are an excellent format. The maps provide one way of tracing the links between the curriculum as planned, as delivered, and as experienced. But like all maps, a balance must be achieved between detail and overall clarity of representation.

Further reading

- Bligh J, Prideaux D, Parsell G. PRISMS: new educational strategies for medical education. *Med Educ* 2001;35:520-1.
- Harden R, Crosby J, Davis M. Outcome based education: part 1—an introduction to outcomes-based education. *Med Teach* 1991;21(1):7-14.
- Harden R. Curriculum mapping: a tool for transparent and authentic teaching and learning. *Med Teach* 2000;23(2):123-7.
- Prideaux D. The emperor's new clothes: from objectives to outcomes. *Med Educ* 2000;34:168-9.
- Print M. *Curriculum development and design*. Sydney: Allen and Unwin, 1993.



The situational model, which emphasises the importance of situation or context in curriculum design



Example of a curriculum map from the students' perspective. Each of the boxes representing the elements of design can be broken down into further units and each new unit can be related to the others to illustrate the interlinking of all the components of the curriculum

3 Problem based learning

Diana Wood

Problem based learning is used in many medical schools in the United Kingdom and worldwide. This article describes this method of learning and teaching in small groups and explains why it has had an important impact on medical education.

What is problem based learning?

In problem based learning (PBL) students use "triggers" from the problem case or scenario to define their own learning objectives. Subsequently they do independent, self directed study before returning to the group to discuss and refine their acquired knowledge. Thus, PBL is not about problem solving per se, but rather it uses appropriate problems to increase knowledge and understanding. The process is clearly defined, and the several variations that exist all follow a similar series of steps.

Group learning facilitates not only the acquisition of knowledge but also several other desirable attributes, such as communication skills, teamwork, problem solving, independent responsibility for learning, sharing information, and respect for others. PBL can therefore be thought of as a small group teaching method that combines the acquisition of knowledge with the development of generic skills and attitudes. Presentation of clinical material as the stimulus for learning enables students to understand the relevance of underlying scientific knowledge and principles in clinical practice.

However, when PBL is introduced into a curriculum, several other issues for curriculum design and implementation need to be tackled. PBL is generally introduced in the context of a defined core curriculum and integration of basic and clinical sciences. It has implications for staffing and learning resources and demands a different approach to timetabling, workload, and assessment. PBL is often used to deliver core material in non-clinical parts of the curriculum. Paper based PBL scenarios form the basis of the core curriculum and ensure that all students are exposed to the same problems. Recently, modified PBL techniques have been introduced into clinical education, with "real" patients being used as the stimulus for learning. Despite the essential ad hoc nature of learning clinical medicine, a "key cases" approach can enable PBL to be used to deliver the core clinical curriculum.

What happens in a PBL tutorial?

PBL tutorials are conducted in several ways. In this article, the examples are modelled on the Maastricht "seven jump" process, but its format of seven steps may be shortened.

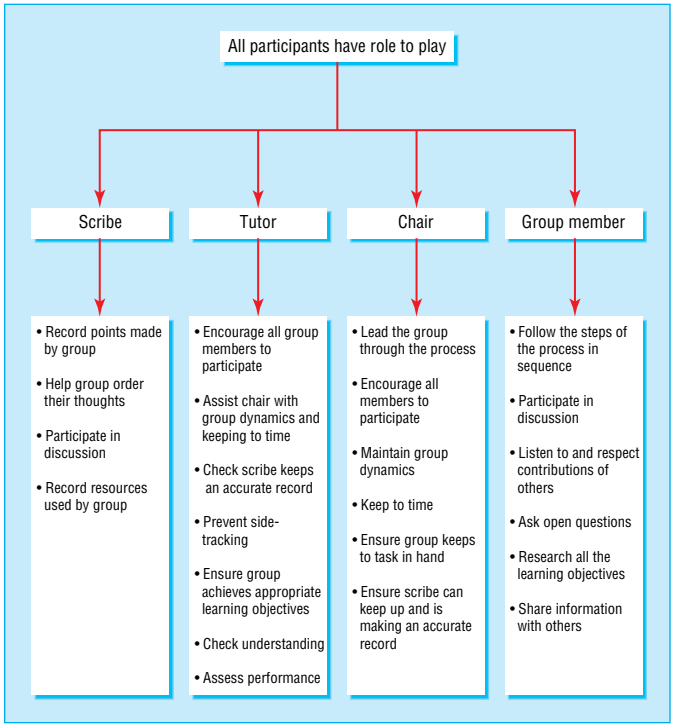
A typical PBL tutorial consists of a group of students (usually eight to 10) and a tutor, who facilitates the session. The length of time (number of sessions) that a group stays together with each other and with individual tutors varies between institutions. A group needs to be together long enough to allow good group dynamics to develop but may need to be changed occasionally if personality clashes or other dysfunctional behaviour emerges.

Students elect a chair for each PBL scenario and a "scribe" to record the discussion. The roles are rotated for each scenario. Suitable flip charts or a whiteboard should be used for recording the proceedings. At the start of the session,

The group learning process: acquiring desirable learning skills

Generic skills and attitudes

- Teamwork
- Chairing a group
- Listening
- Recording
- Cooperation
- Respect for colleagues' views
- Critical evaluation of literature
- Self directed learning and use of resources
- Presentation skills



Roles of participants in a PBL tutorial

depending on the trigger material, either the student chair reads out the scenario or all students study the material. If the trigger is a real patient in a ward, clinic, or surgery then a student may be asked to take a clinical history or identify an abnormal physical sign before the group moves to a tutorial room. For each module, students may be given a handbook containing the problem scenarios, and suggested learning resources or learning materials may be handed out at appropriate times as the tutorials progress.

The role of the tutor is to facilitate the proceedings (helping the chair to maintain group dynamics and moving the group through the task) and to ensure that the group achieves appropriate learning objectives in line with those set by the curriculum design team. The tutor may need to take a more active role in step 7 of the process to ensure that all the students have done the appropriate work and to help the chair to suggest a suitable format for group members to use to present the results of their private study. The tutor should encourage students to check their understanding of the material. He or she can do this by encouraging the students to ask open questions and ask each other to explain topics in their own words or by the use of drawings and diagrams.

PBL in curriculum design

PBL may be used either as the mainstay of an entire curriculum or for the delivery of individual courses. In practice, PBL is usually part of an integrated curriculum using a systems based approach, with non-clinical material delivered in the context of clinical practice. A module or short course can be designed to include mixed teaching methods (including PBL) to achieve the learning outcomes in knowledge, skills, and attitudes. A small number of lectures may be desirable to introduce topics or provide an overview of difficult subject material in conjunction with the PBL scenarios. Sufficient time should be allowed each week for students to do the self directed learning required for PBL.

Writing PBL scenarios

PBL is successful only if the scenarios are of high quality. In most undergraduate PBL curriculums the faculty identifies learning objectives in advance. The scenario should lead students to a particular area of study to achieve those learning objectives (see box on page 11).

How to create effective PBL scenarios*

- Learning objectives likely to be defined by the students after studying the scenario should be consistent with the faculty learning objectives
- Problems should be appropriate to the stage of the curriculum and the level of the students' understanding
- Scenarios should have sufficient intrinsic interest for the students or relevance to future practice
- Basic science should be presented in the context of a clinical scenario to encourage integration of knowledge
- Scenarios should contain cues to stimulate discussion and encourage students to seek explanations for the issues presented
- The problem should be sufficiently open, so that discussion is not curtailed too early in the process
- Scenarios should promote participation by the students in seeking information from various learning resources

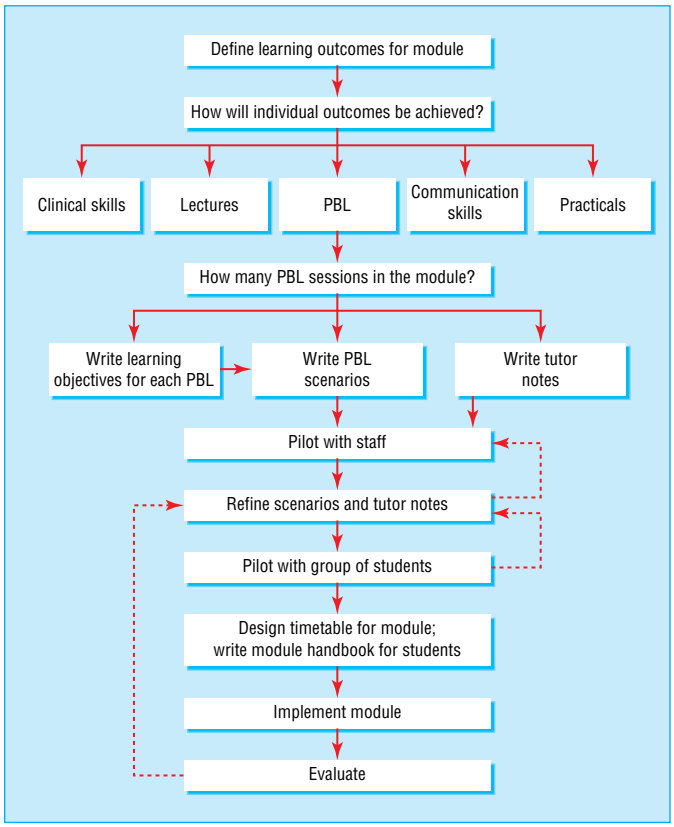
*Adapted from Dolmans et al. *Med Teacher* 1997;19:185-9

Examples of trigger material for PBL scenarios

- Paper based clinical scenarios
- Experimental or clinical laboratory data
- Photographs
- Video clips
- Newspaper articles
- All or part of an article from a scientific journal
- A real or simulated patient
- A family tree showing an inherited disorder

PBL tutorial process

Step 1—Identify and clarify unfamiliar terms presented in the scenario; scribe lists those that remain unexplained after discussion
Step 2—Define the problem or problems to be discussed; students may have different views on the issues, but all should be considered; scribe records a list of agreed problems
Step 3—"Brainstorming" session to discuss the problem(s), suggesting possible explanations on basis of prior knowledge; students draw on each other's knowledge and identify areas of incomplete knowledge; scribe records all discussion
Step 4—Review steps 2 and 3 and arrange explanations into tentative solutions; scribe organises the explanations and restructures if necessary
Step 5—Formulate learning objectives; group reaches consensus on the learning objectives; tutor ensures learning objectives are focused, achievable, comprehensive, and appropriate
Step 6—Private study (all students gather information related to each learning objective)
Step 7—Group shares results of private study (students identify their learning resources and share their results); tutor checks learning and may assess the group



Designing and implementing a curriculum module using PBL supported by other teaching methods

Staff development

Introducing PBL into a course makes new demands on tutors, requiring them to function as facilitators for small group learning rather than acting as providers of information. Staff development is essential and should focus on enabling the PBL tutors to acquire skills in facilitation and in management of group dynamics (including dysfunctional groups).

Tutors should be also given information about the institution's educational strategy and curriculum programme so that they can help students to understand the learning objectives of individual modules in the context of the curriculum as a whole. Methods of assessment and evaluation should be described, and time should be available to discuss anxieties.

Staff may feel uncertain about facilitating a PBL tutorial for a subject in which they do not themselves specialise. Subject specialists may, however, be poor PBL facilitators as they are more likely to interrupt the process and revert to lecturing. None the less, students value expertise, and the best tutors are subject specialists who understand the curriculum and have excellent facilitation skills. However, enthusiastic non-specialist tutors who are trained in facilitation, know the curriculum, and have adequate tutor notes, are good PBL tutors.

Assessment of PBL

Student learning is influenced greatly by the assessment methods used. If assessment methods rely solely on factual recall then PBL is unlikely to succeed in the curriculum. All assessment schedules should follow the basic principles of testing the student in relation to the curriculum outcomes and should use an appropriate range of assessment methods.

Assessment of students' activities in their PBL groups is advisable. Tutors should give feedback or use formative or summative assessment procedures as dictated by the faculty assessment schedule. It is also helpful to consider assessment of the group as a whole. The group should be encouraged to reflect on its PBL performance including its adherence to the process, communication skills, respect for others, and individual contributions. Peer pressure in the group reduces the likelihood of students failing to keep up with workload, and the award of a group mark—added to each individual's assessment schedule—encourages students to achieve the generic goals associated with PBL.

Conclusion

PBL is an effective way of delivering medical education in a coherent, integrated programme and offers several advantages over traditional teaching methods. It is based on principles of adult learning theory, including motivating the students, encouraging them to set their own learning goals, and giving them a role in decisions that affect their own learning.

Predictably, however, PBL does not offer a universal panacea for teaching and learning in medicine, and it has several well recognised disadvantages. Traditional knowledge based assessments of curriculum outcomes have shown little or no difference in students graduating from PBL or traditional curriculums. Importantly, though, students from PBL curriculums seem to have better knowledge retention. PBL also generates a more stimulating and challenging educational environment, and the beneficial effects from the generic attributes acquired through PBL should not be underestimated.

A dysfunctional group: a dominant character may make it difficult for other students to be heard

Advantages and disadvantages of PBL

Advantages of PBL	Disadvantages of PBL
Student centred PBL—It fosters active learning, improved understanding, and retention and development of lifelong learning skills	*Tutors who can't "teach"*—Tutors enjoy passing on their own knowledge and understanding so may find PBL facilitation difficult and frustrating
Generic competencies—PBL allows students to develop generic skills and attitudes desirable in their future practice	*Human resources*—More staff have to take part in the tutoring process
Integration—PBL facilitates an integrated core curriculum	*Other resources*—Large numbers of students need access to the same library and computer resources simultaneously
Motivation—PBL is fun for students and tutors, and the process requires all students to be engaged in the learning process	*Role models*—Students may be deprived access to a particular inspirational teacher who in a traditional curriculum would deliver lectures to a large group
"Deep" learning—PBL fosters deep learning (students interact with learning materials, relate concepts to everyday activities, and improve their understanding)	*Information overload*—Students may be unsure how much self directed study to do and what information is relevant and useful
Constructivist approach—Students activate prior knowledge and build on existing conceptual knowledge frameworks	

Further reading

- Davis MH, Harden RM. AMEE medical education guide number 15: problem-based learning: a practical guide. *Med Teacher* 1999;21:130-40.
- Norman GR, Schmidt HG. Effectiveness of problem-based learning curricula: theory, practice and paper darts. *Med Educ* 2000;34:721-8.
- Albanese M. Problem based learning: why curricula are likely to show little effect on knowledge and clinical skills. *Med Educ* 2000;34:729-38.

Christ and St John with Angels by Peter Paul Rubens is from the collection of the Earl of Pembroke/BAL.The Mad Hatter's Tea Party is by John Tenniel.

PBL scenarios: the importance of linking to faculty learning objectives

PBL scenario 1

A 35 year old part time nurse, presented to her general practitioner, Dr Smith, with a six month history of weight loss (12.7 kg). When questioned, she said she was eating well but had diarrhoea. She also felt exhausted and had developed insomnia. On further questioning she admitted to feeling increasingly hot and shaky and to having muscle weakness in her legs, particularly when climbing stairs. She was normally well and had not seen the doctor since her last pregnancy eight years ago.

A blood test showed the following results:
Free thyroxine 49.7 pmol/l (normal range 11 to 24.5)
Total thyroxine 225 nmol/l (normal range 60 to 150)
Thyroid stimulating hormone $<$ 0.01 mU/l (0.4 to 4.0)

Dr Smith referred her to an endocrinologist at the local hospital where initial investigations confirmed a diagnosis of Graves' disease. She was treated with carbimazole and propranolol for the first month of treatment followed by carbimazole alone. After discussing the therapeutic options, she opted to have iodine-131 treatment.

Faculty learning objectives

- Describe the clinical features of thyrotoxicosis and diagnostic signs of Graves' disease
- Interpret basic thyroid function tests in the light of the pituitary thyroid axis and feedback mechanisms
- List the types of treatment for thyrotoxicosis including their indications, mode of action, and potential side effects

Notes

This scenario is part of a core endocrinology and metabolism module for third year undergraduate medical students. The faculty learning objectives relate to the scenario; the problem is relevant to the level of study and integrates basic science with clinical medicine. The combination of basic science, clinical medicine, and therapeutics should lead to extensive discussion and broadly based self directed learning

PBL scenario 2

Mr JB, a 58 year old car mechanic with a history of chronic obstructive pulmonary disease, was at work when he complained of pain in his chest. The pain steadily got worse and he described an aching in his jaw and left arm. One hour after the pain started he collapsed and his colleagues called an ambulance. When he arrived at the local accident and emergency department, Mr JB was pale, sweaty, and in severe pain.

Examination showed:
Blood pressure 80/60 mm Hg
Heart rate 64 beats/min
Electrocardiography showed anterolateral myocardial infarction

He was treated with diamorphine, metoclopramide, and aspirin. As the accident and emergency staff were preparing to give him streptokinase, he had a cardiac arrest. Electrocardiography showed asystolic cardiac arrest. Despite all efforts, resuscitation failed.

Faculty learning objectives

- List the risk factors for myocardial infarction
- Describe a rehabilitation programme for patients who have had a myocardial infarction

Notes

This scenario is part of a core module in the cardiorespiratory system for first year undergraduate medical students. The scenario is complex for students with limited clinical experience. The faculty learning objectives relate to public health and epidemiological aspects of ischaemic heart disease. For increased impact, the faculty illustrated the case with a dramatic scenario. Students would be unlikely to arrive at the same objectives, probably concentrating on clinical aspects of acute myocardial infarction and its management

4 Evaluation

Jill Morrison

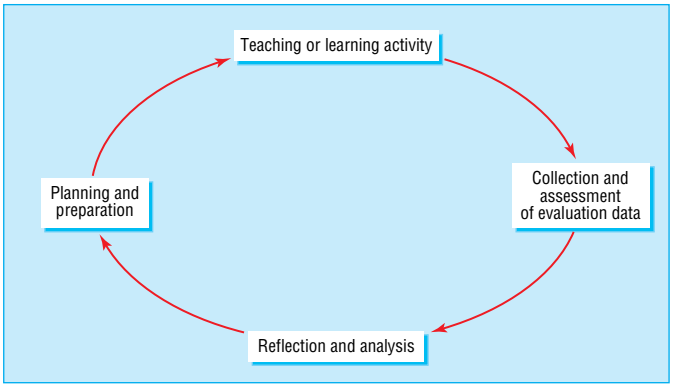
Evaluation is an essential part of the educational process. The focus of evaluation is on local quality improvement and is analogous to clinical audit. Medical schools require evaluation as part of their quality assurance procedures, but the value of evaluation is much greater than the provision of simple audit information. It provides evidence of how well students' learning objectives are being achieved and whether teaching standards are being maintained. Importantly, it also enables the curriculum to evolve. A medical curriculum should constantly develop in response to the needs of students, institutions, and society. Evaluation can check that the curriculum is evolving in the desired way. It should be viewed positively as contributing to the academic development of an institution and its members.

Purpose of evaluation

- To ensure teaching is meeting students' learning needs
- To identify areas where teaching can be improved
- To inform the allocation of faculty resources
- To provide feedback and encouragement for teachers
- To support applications for promotion by teachers
- To identify and articulate what is valued by medical schools
- To facilitate development of the curriculum

Evaluation versus research

Evaluation and educational research are similar activities but with important differences. Research is usually aimed at producing generalisable results that can be published in peer reviewed literature, and it requires ethical and other safeguards. Evaluation is generally carried out for local use and does not usually require ethics committee approval. Evaluation has to be carefully considered by curriculum committees, however, to ensure that it is being carried out ethically. Finally, evaluation is a continuous process, whereas research may not become continuous if the answer to the question is found.



Evaluation cycle. From Wilkes et al (see "Further reading" box)

What should be evaluated

Evaluation may cover the process and/or outcome of any aspect of education, including the delivery and content of teaching. Questions about delivery may relate to organisation—for example, administrative arrangements, physical environment, and teaching methods. Information may also be sought about the aptitude of the teacher(s) involved. The content may be evaluated for its level (it should not be too easy or too difficult), its relevance to curriculum objectives, and integration with previous learning.

Outcome measures may show the impact of the curriculum on the knowledge, skills, attitudes, and behaviour of students. Kirkpatrick described four levels on which to focus evaluation; these have recently been adapted for use in health education evaluation by Barr and colleagues. Some indication of these attributes may be obtained by specific methods of inquiry—for example, by analysing data from student assessments.

Kirkpatrick's four levels on which to focus evaluation*

- Level 1—Learner's reactions
- Level 2a—Modification of attitudes and perceptions
- Level 2b—Acquisition of knowledge and skills
- Level 3—Change in behaviour
- Level 4a—Change in organisational practice
- Level 4b—Benefits to patients or clients

*Adapted by Barr et al (see "Further reading" box)

The full impact of the curriculum may not be known until some time after the student has graduated

Evaluation in curriculum planning

Evaluation should be designed at the start of developing a curriculum, not added as an afterthought. When an educational need has been identified, the first stage is to define the learning outcomes for the curriculum. The goals of the evaluation should be clearly articulated and linked to the outcomes.

Clarifying the goals of evaluation will help to specify the evidence needed to determine success or failure of the training. .A protocol should then be prepared so that individual responsibilities are clearly outlined.

Questions to ask when planning an evaluation

- What are the goals of the evaluation?
- From whom and in what form will data be collected?
- Who will collect and analyse data?
- What type of analysis, interpretation, and decision rules will be used and by whom?
- Who will see the results of the evaluation?

Designing evaluation

An ideal evaluation method would be reliable, valid, acceptable, and inexpensive. Unfortunately, ideal methods for evaluating teaching in medical schools are scarce.

Establishing the reliability and validity of instruments and methods of evaluation can take many years and be costly. Testing and retesting of instruments to establish their psychometric properties without any additional benefit for students or teachers is unlikely to be popular with them. There is a need for robust "off the shelf" instruments that can be used to evaluate curriculums reliably. The process of evaluation itself may produce a positive educational impact if it emphasises those elements that are considered valuable and important by medical schools.

Characteristics of an ideal evaluation

- Reliability
- Validity
- Acceptability—to evaluator and to person being evaluated
- Inexpensiveness

To reduce possible bias in evaluation, collect views from more than one group of people—for example, students, teachers, other clinicians, and patients

Participation by students

Several issues should be considered before designing an evaluation that collects information from students.

Competence—Students can be a reliable and valid source of information. They are uniquely aware of what they can consume, and they observe teaching daily. They are also an inexpensive resource. Daily contact, however, does not mean that students are skilled in evaluation. Evaluation by students should be limited to areas in which they are competent to judge.

Ownership—Students who are not committed to an evaluation may provide poor information. They need to feel ownership for an evaluation by participating in its development. The importance of obtaining the information and the type of information needed must be explicit. Usually the results of an evaluation will affect only subsequent cohorts of students, so current students must be convinced of the value of providing data.

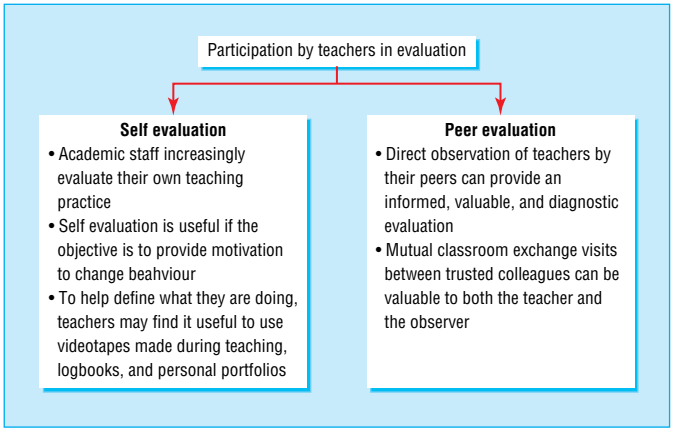
Sampling—Students need to feel that their time is respected. If they are asked to fill out endless forms they will resent the waste of their time. If they become bored by tedious repetition, the reliability of the data will deteriorate. One solution is to use different sampling strategies for evaluating different elements of a curriculum. If reliable information can be obtained from 100 students, why collect data from 300?

Anonymity is commonly advocated as a guard against bias when information is collected from students. However, those who support asking students to sign evaluation forms say that this helps to create a climate of responsible peer review. If students are identifiable from the information they provide, this must not affect their progress. Data should be collected centrally and students' names removed so that they cannot be identified by teachers whom they have criticised.

Feedback—Students need to know that their opinions are valued, so they should be told of the results of the evaluation and given details of the resulting action.

Areas of competence of students to evaluate teaching and curriculum

- Design: whether the curriculum enables students to reach their learning objectives; whether it fits well with other parts of the curriculum
- Delivery: attributes of teacher and methods used
- Administrative arrangements

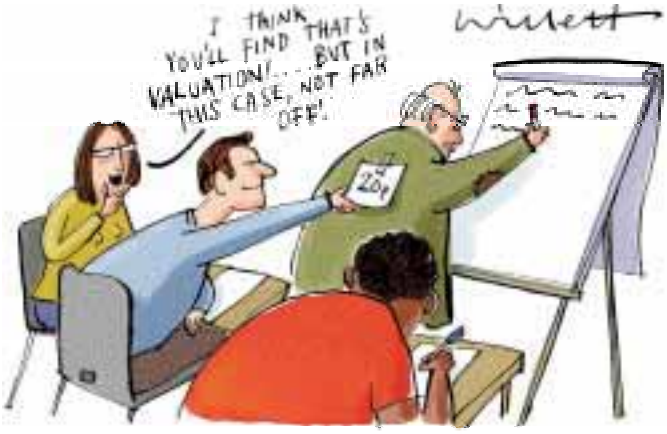


Issues relating to students' participation in evaluation may also apply to teachers, but self evaluation and peer evaluation are also relevant

Methods of evaluation

Evaluation may involve subjective and objective measures and qualitative and quantitative approaches. The resources devoted to evaluation should reflect its importance, but excessive data collection should be avoided. A good system should be easy to administer and use information that is readily available.

Interviews—Individual interviews with students are useful if the information is sensitive—for example, when a teacher has received poor ratings from students, and the reasons are not clear. A group interview can provide detailed views from students or teachers. A teaching session can end with reflection by the group.



Surveys—Questionnaires are useful for obtaining information from large numbers of students or teachers about the teaching process. Electronic methods for administering questionnaires may improve response rates. The quality of the data, however, is only as good as the questions asked, and the data may not provide the reasons for a poorly rated session.

Information from student assessment—Data from assessment are useful for finding out if students have achieved the learning outcomes of a curriculum. A downward trend in examination results over several cohorts of students may indicate a deficiency in the curriculum. Caution is needed when interpreting this source of information, as students' examination performance depends as much on their application, ability, and motivation as on the teaching.

Completing the evaluation cycle

The main purpose of evaluation is to inform curriculum development. No curriculum is perfect in design and delivery. If the results of an evaluation show that no further development is needed, doubt is cast on the methods of evaluation or the interpretation of the results.

This does not mean that curriculums should be in a constant state of change, but that the results of evaluation to correct deficiencies are acted on, that methods continue to improve, and that content is updated. Then the process starts all over again.

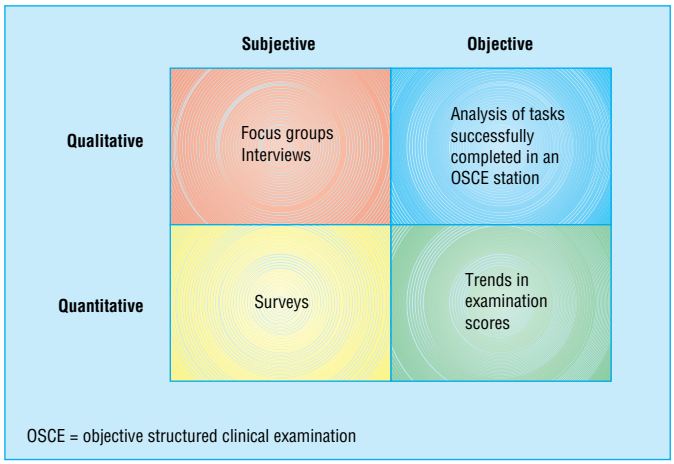
Worked example

Background
Clinical teaching staff think that students are becoming weaker at examining cranial nerves. The examination scores for that part of the objective structured clinical examination (OSCE) carried out at the end of year three show a decline over several years. Three focus groups are held with students in year four, and several clinical teachers are interviewed. The results suggest that the decline is due to fewer appropriate patients presenting at outpatient sessions where cranial nerve examination is taught and to a lack of opportunities for practising examination skills.

Intervention
A teaching session is designed for delivery in the clinical skills centre. After that, students should be able to do a systematic examination of cranial nerves. They should also recognise normal signs and know which common abnormalities to look for. Sessions are timetabled for practising skills learnt during the teaching session.

Evaluation
A questionnaire is developed for completion by a third of students. It seeks their views on the teaching process, including the teaching skills of the tutor, physical aspects of the teaching environment, appropriateness of the teaching material, and opportunities for practising examination skills. Outcome measures include comparison of examination scores for students in the previous cohort with those participating in the teaching session, plus a questionnaire for all clinical supervisors for neurology in the following year to get their views about students' examination skills. A tenth of students with a range of scores in the relevant part of the OSCE are interviewed to find out the reasons for their varied scores. The evaluation results are disseminated widely to staff and students.

Questionnaire surveys are the most common evaluation tool



Examples of methods of evaluation

Key points

Evaluation should:
- Enable strategic development of a curriculum
- Be a positive process that contributes to the academic development of a medical school

The goals of an evaluation should:
- Be clearly articulated
- Be linked to the outcomes of the teaching

When carrying out an evaluation:
- More than one source and type of information should be sought
- The results should be fed back to participants and details of the resulting action given

Learners need:
- To be involved in developing an evaluation
- To feel their time is respected
- To know their opinions are valued and acted on

Evaluators must:
- Act on the results of the evaluation to correct deficiencies, improve methods, and update content
- Repeat the process

Further reading

- Robson C. *Small scale evaluation.* London: Sage, 2000.
- Cohen L, Manion L. *Research methods in education.* 4th ed. London: Routledge, 1994.
- Snell L, Tallett S, Haist S, Hays R, Norcini J, Prince K, et al. A review of the evaluation of clinical teaching: new perspectives and challenges. *Med Educ* 2000;34:862-70.
- Barr H, Freeth D, Hammick M, Koppel, Reeves S. *Evaluations of interprofessional education: a United Kingdom review of health and social care.* London: CAIPE/BERA, 2000.
- Wilkes M, Bligh J. Evaluating educational interventions. *BMJ* 1999;318:1269-72.

5 Teaching large groups

Peter Cantillon

Lecturing or large group teaching is one of the oldest forms of teaching. Whatever their reputation, lectures are an efficient means of transferring knowledge and concepts to large groups. They can be used to stimulate interest, explain concepts, provide core knowledge, and direct student learning.

However, they should not be regarded as an effective way of teaching skills, changing attitudes, or encouraging higher order thinking. Large group formats tend to encourage passive learning. Students receive information but have little opportunity to process or critically appraise the new knowledge offered.

How can lectures be used to maximise learning and provide opportunities for student interaction? This article will supply some of the answers and should help you to deliver better, more interactive lectures.

Getting your bearings

It is important to find out as much as possible about the context of the lecture—that is, where it fits into the course of which it is part.

An understanding of the context will allow you to prepare a lecture that is both appropriate and designed to move students on from where they are.

Helping students to learn in lectures

An important question for any lecturer to consider when planning a teaching session is, "how can I help my students to learn during my lecture?" There are several different techniques you can use to aid student learning in a large group setting.

Helping your students to learn

- Use concrete examples to illustrate abstract principles
- Give handouts of the lecture slides, with space to write notes
- Give handouts with partially completed diagrams and lists for the students to complete during or after the lecture
- Allow for pauses in the delivery to give students time to write notes
- Check for understanding by asking questions or by running a mini quiz

Planning your lecture

It is important to distinguish between the knowledge and concepts that are essential (need to know) and those which, though interesting, are not part of the core message (nice to know).

The aims of the lecture should be clearly defined ("what do I hope to achieve with this lecture?"). These will help to define the teaching methods and the structure. If, for example, the purpose of the lecture is to introduce new knowledge and concepts, then a classic lecture structure might be most appropriate.

On the other hand, if the purpose is to make the students aware of different approaches to a particular clinical problem, a problem oriented design in which alternative approaches are presented and discussed might be a more appropriate format.

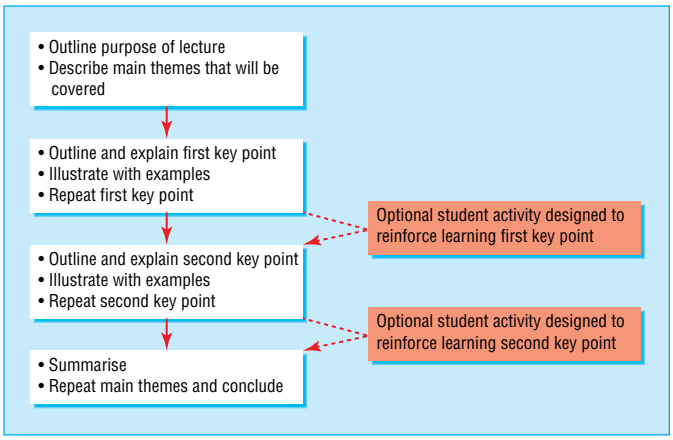
A lecturer holds forth …

What you need to know before planning a lecture

- How your lecture fits into the students' course or curriculum
- The students' knowledge of your subject—try to get a copy of the lecture and tutorial list for the course
- How the course (and your lecture) will be assessed
- The teaching methods that the students are accustomed to

The successful teacher is no longer on a height, pumping knowledge at high pressure into passive receptacles … he is a senior student anxious to help his juniors.

William Osler (1849-1919)



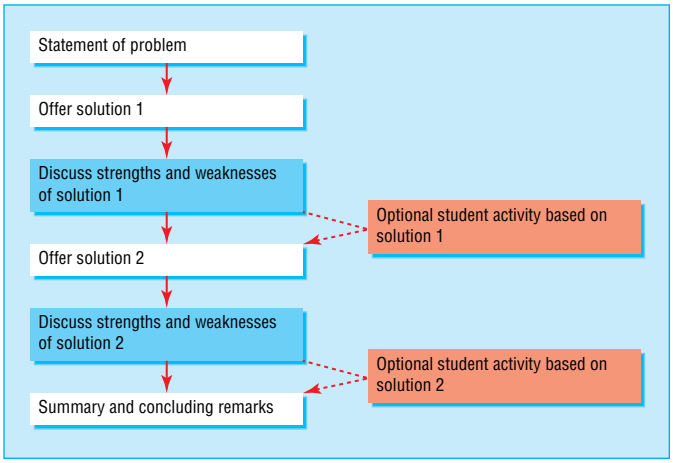
Example of a lecture plan with a classic structure

Choosing teaching media

When you have selected the content of the lecture and placed it into a working structure, the next consideration is how to deliver the message. Which teaching media should be used (for example, slides, overheads, handouts, quizzes)? The most appropriate media will differ depending on the venue, class size, and topic.

Choosing the medium for delivering the lecture

- Which teaching media are available at the teaching venue?
- Which teaching media are you familiar with? (It is not always appropriate to experiment with new media)
- Which medium will best illustrate the concepts and themes that you want to teach the students?
- Which medium would encourage students to learn through interaction during your lecture?



Example of a lecture plan with a problem oriented structure

Getting started

In the first moments of a lecture it is important that the students are given some sense of place and direction. Thus a brief summary of the previous lecture and an indication of the major themes and learning objectives for the current session provide both you and the students with a relatively easy start. If you are working with a new group it may be useful to indicate the ground rules for the session—for example, "switch off mobile phones," or "ask questions at any time."

Handouts

- Handouts can encourage better learning if they allow students more time to listen and think
- Handouts should provide a scaffold on which students can build their understanding of a topic
- Handouts should provide a summary of the major themes while avoiding an exhaustive explanation of each
- Handouts can be used to direct further learning, by including exercises and questions with suggested reading lists

Encouraging students to interact

Students learn well by "doing." Yet there is an understandable tendency for students to regard lectures as an opportunity to sit back, be entertained, and "soak up" the learning. However, you can use various methods to encourage students to take a more active part in the learning process.

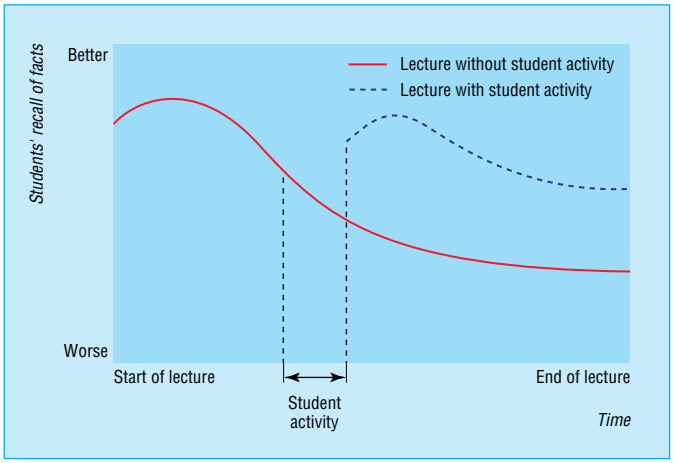
Students' attention (and recall) is best at the beginning and end of a lecture. Recall can be improved by changing the format of your lecture part way through. It is also important when planning a lecture to think about activities and exercises that will break up the presentation.

Ask questions

It is useful to ask questions of the group at various stages in the lecture, to check comprehension and promote discussion. Many lecturers are intimidated by the silence following a question and fall into the trap of answering it themselves. Wait for the answers to come. It takes time for students to move from listening to thinking mode. A simple tip is to count slowly to 10 in your head—a question is almost certain to arrive.

Get students to ask you questions

An alternative to getting students to answer questions is to ask them to direct questions at you. A good way of overcoming students' normal fear of embarrassment is to ask them to prepare questions in groups of two or three. Questions can then be invited from groups at random. When asked a question, you should repeat it out loud to ensure that the whole group is aware of what was asked. Seeking answers to the question from other students, before adding your own views, can increase the level of interaction further.



Graph showing effect of students' interaction on their ability to recall what they have heard in a lecture. Adapted from Bligh, 2000 (see "Recommended reading" box)

"Tell me, and I forget. Show me, and I remember. Involve me, and I understand"

Chinese proverb

Brainstorming
Brainstorming is a technique for activating the students' knowledge or current understanding of an issue or theme. The lecturer invites answers to a question or problem from the audience and writes them, without comment, on a board or overhead. After a short period, usually about two or three minutes, the lecturer reviews the list of "answers" with the class. The answers can be used to provide material for the next part of the lecture or to give students an idea of where they are before they move on. By writing answers in a way that can be seen by everyone in the audience, you allow the students to learn from each other.

Buzz groups
Buzz groups also encourage interaction. They consist of groups of two to five students working for a few minutes on a question, problem, or exercise set by the lecturer. Buzz group activity is a useful means of getting students to process and use new information to solve problems. At the end of the buzz group session, the teacher can either continue with the lecture or check the results of the exercise by asking one or two groups to present their views. Remember that in an amphitheatre lecture hall, students can sit on their own desks to interact with the students behind them.

Mini-assessments
Mini-assessments and exercises are used in lectures to help students to recognise gaps in their learning and to encourage them to use new material in practice. Brief assessments can also allow the lecturer to measure how well the messages are being understood. Students could be asked, for example, to complete a brief, multiple choice questionnaire or a "one-minute" paper. The timing of quizzes and exercises will depend on what is required. An assessment of prior learning would be best at the start of a lecture, whereas an estimate of learning from the current session might be best carried out towards the end of the lecture.

How to end your lecture

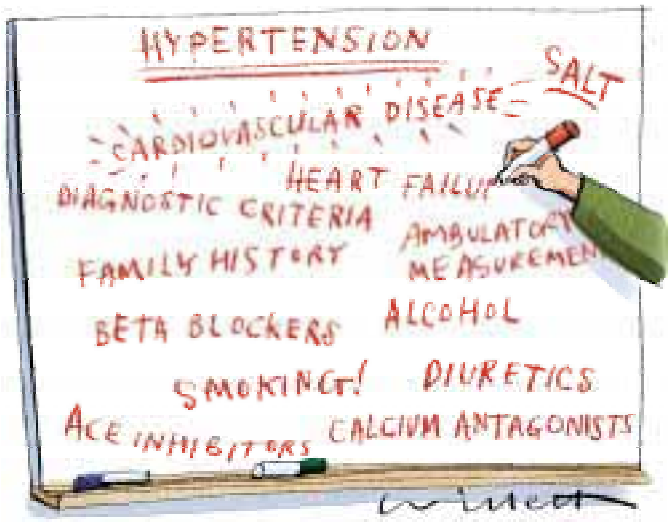
At the end of a lecture it is important to summarise the key points and direct students toward further learning. You may present the key points on a slide or overhead. Alternatively, you may go through the main headings on a handout. Students are encouraged to learn more about a subject if they are set tasks or exercises that will require them to look further than the lecture notes for answers and ideas. The end of a lecture is also a common time for questions. Students may find the use of a one-minute paper a useful tool to help them to identify concepts and impressions that need clarification.

Evaluating your lecture

Practice does make perfect, but the process of developing as a lecturer is greatly helped if some effort is made to evaluate performance. Evaluation involves answering questions such as "how did I do?" or "what did the students learn?"

A lecture can be evaluated in different ways. If the students are to be used as a source of feedback, the following methods are useful:

- Ask a sample of the students if you can read their lecture notes—this exercise gives some insight into what students have learned and understood
- Ask for verbal feedback from individual students
- Ask the students to complete a one-minute paper



"One-minute" paper worksheet

Name: _____

Date: _____

Lecture title: _____

Directions: Take a moment to think about the lecture you have just attended, and then answer the following questions.

1. What was the most important thing you learned in today's lecture?

_____

2. What question remains uppermost in your mind at the end of today's lecture?

_____

3. What was the "muddiest point" in today's lecture?

_____

Example of a one-minute paper

Please rate the lecture on the following items

	Strongly agree	Slightly agree	Slightly disagree	Strongly disagree
Clear				
Interesting				
Easy to take notes from				
Well organised				
Relevant to the course				

Example of an evaluation form focusing on the lecture. Adapted from Brown et al, 2001 (see "Recommended reading" box)

- Ask the students to complete an evaluation questionnaire.

If you want to evaluate your teaching style and delivery, peers can be a useful source of feedback:

- Ask a colleague to observe part or all of a lecture and provide feedback afterwards. It is important to inform the observer what aspects of the lecturing process you want evaluated—for example, clarity, logical flow, effectiveness of the media used
- Videotape the lecture for private viewing, and arrange a joint viewing with a colleague later.

Lectures are still a common teaching method in both undergraduate and postgraduate medical education. Their continued popularity is due to the fact that they represent an effective and efficient means of teaching new concepts and knowledge. This article has emphasised the importance of good lecture planning and of the inclusion of student interaction to ensure effective learning.

Recommended reading

- Newble DI, Cannon R. *A handbook for medical teachers.* 4th ed. Dordrecht, Netherlands: Kluwer Academic, 2001.
- Gibbs G, Habeshaw T. *Preparing to teach.* Bristol: Technical and Educational Services, 1989.
- Bligh DA. *What's the use of lectures?* San Francisco: Jossey-Bass, 2000.
- Brown G, Manogue M. AMEE medical education guide No 22: refreshing lecturing: a guide for lecturers. *Medical Teacher* 2001;23:231-44.

Please rate the lecturer on the following items

	Strongly agree	Slightly agree	Slightly disagree	Strongly disagree
Was enthusiastic				
Was clearly audible				
Seemed confident				
Gave clear explanations				
Encouraged participation				

Example of an evaluation form focusing on the lecturer rather than the lecture. Adapted from Brown et al, 2001 (see "Recommended reading" box)

6 Teaching small groups

David Jaques

Group discussion plays a valuable role in the all-round education of students, whether in problem based learning and team projects or in the more traditional academic scenario of the tutorial or seminar. When it works well, discussion can allow students to negotiate meanings, express themselves in the language of the subject, and establish closer contact with academic staff than more formal methods permit. Discussion can also develop the more instrumental skills of listening, presenting ideas, persuading, and working as part of a team. But perhaps most importantly, discussion in small groups can or should give students the chance to monitor their own learning and thus gain a degree of self direction and independence in their studies.

All these worthy aims require active participation and the ready expression of ideas. Yet it frequently doesn't work out this way. Indeed many tutors too readily fall back on their reserve positions of authority, expert, and prime talker. Many of the problems associated with leading small groups effectively are likely to be exacerbated with larger groups. So how can we avoid the common traps?

If you are leading a group discussion you will need to consider both the configuration of the group and your own behaviour. Groups often communicate poorly because the physical conditions make it difficult to communicate. For example, in a group of 10 students seated round a rectangular table, at least four students on either side of the table have no eye contact with each other, thus reducing participation. If you ask and answer questions all the time, even less interaction is likely.

If a group sits in a circle without a table, communication is likely to be easier. When the discussion has started, it is your responsibility as discussion leader to listen to and respond to the whole group. Listening becomes a problem when the students regard you as an expert or you engage with one or two of the more vocal students rather than the whole group.

More structure, less intervention

Being a democratic discussion leader involves making the right sort of nudges and interventions. The role can be made a lot less demanding by using more structure and less intervention in the group process. The rest of this article shows how clear and purposeful group structures can help to bypass many of the problems outlined above, by delegating responsibility for group interaction (and therefore for learning) to the students.

Group structures and processes

You can minimise your internal involvement in the group process by organising or structuring groups into smaller units, especially when the group process is likely to be problematical. This is particularly so when you wish to mobilise a sense of coherence and full participation among a largish group of students. A sequence of tasks might then be set. For example:

- Students work individually for five minutes drawing up a list
- They share their ideas in pairs for 10 minutes
- In groups of four to six, students write up categories on a large sheet of paper
- This is followed by 25 minutes of open discussion among the groups.

"By separating teaching from learning, we have teachers who do not listen and students who do not talk"

Based on Palmer P (*The Courage to Teach*. Jossey Bass, 1998)

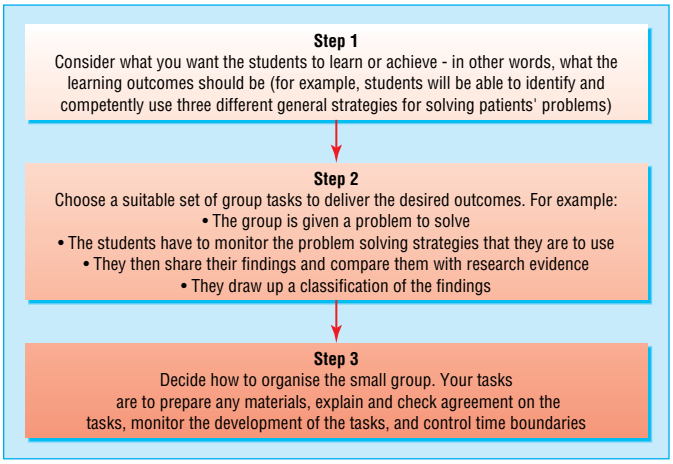
Problems associated with leading effective small groups

- The teacher gives a lecture rather than conducting a dialogue
- The teacher talks too much
- Students cannot be encouraged to talk except with difficulty; they will not talk to each other, but will only respond to questions from the tutor
- Students do not prepare for the sessions
- One student dominates or blocks the discussion
- The students want to be given the solutions to problems rather than discuss them

Your own behaviour can have an enormous effect on how the group functions

Techniques for effective facilitation in group discussion

- Ensure that group members have an agreed set of ground rules—for example, not talking at the same time as another group member
- Ensure that the students are clear about the tasks to be carried out
- When you present a question don't answer it yourself or try to reformulate it—count to 10 silently before speaking again
- When you have something you *could* say (which could be most of the time), count to 10 again
- Look round the group both when you are speaking and when a student is speaking. That way the students will quickly recognise that they are addressing the group rather than just you. It will allow you to pick up cues from those who want to speak but are either a bit slow or inhibited



Planning the structure of a small discussion group

Your role in this kind of situation may be to move round checking that everyone understands and accepts the task and is doing it in an appropriate way and to encourage completion as the end point approaches. You could leave the room for a while and let the groups work without supervision.

The following group structures require some assertive leadership to set up but allow you to take a back seat as the process itself takes over the direction of events.

To encourage group interaction consider breaking a larger group into smaller groups of five or six students; organise membership on a heterogeneous or random basis to prevent cliques forming

Group round

Each person has a brief time—say, 20 seconds or one minute—to say something in turn round the group. The direction round the group can be decided by the first contributor, or members can speak in a random order. More interest and energy is usually generated, however, if the first person chooses who should go second, the second who should go third, and so on.

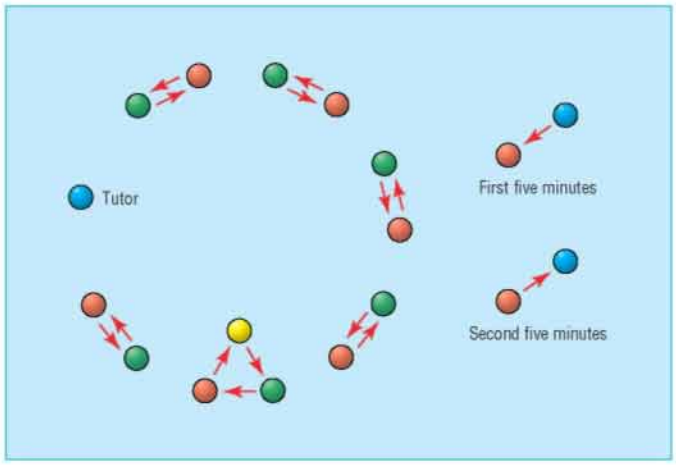
Group rounds are particularly useful at the beginning of any meeting so that everyone is involved from the start and, depending on what the group is asked to speak about, as a way of checking on learning issues

Buzz groups

With larger groups a break is often needed:

- To provide a stimulating change in the locus of attention
- For you to gain some idea of what the students know
- For the students to check their own understanding.

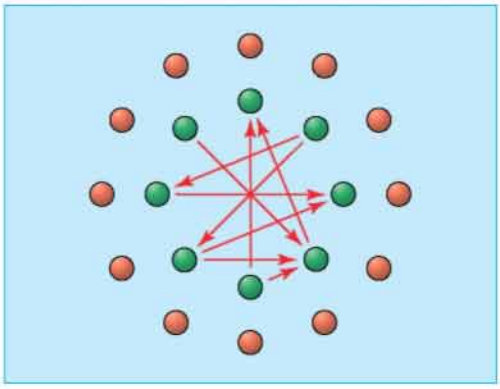
During a discussion students could be asked to turn to their neighbour to discuss for a few minutes any difficulties in understanding, to answer a prepared question, or to speculate on what they think will happen next in the proceedings. This will bring a sense of participation and some lively feedback. Buzz groups enable students to express difficulties they would have been unwilling to reveal to the whole class. (A variation is to allocate three or five minutes each way to the pairs—each phase is for one-way communication.)



Buzz groups, with pairs for one-way, five-minute communication

Snowball groups

Snowball groups (or pyramids) are an extension of buzz groups. Pairs join up to form fours, then fours to eights. These groups of eight report back to the whole group. This developing pattern of group interaction can ensure comprehensive participation, especially when it starts with individuals writing down their ideas before sharing them. To avoid students becoming bored with repeated discussion of the same points, it is a good idea to use increasingly sophisticated tasks as the groups gets larger.



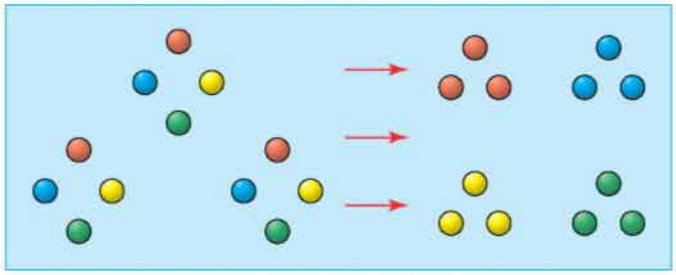
Fishbowl structure—inside group discusses, outside group listens in

Fishbowls

The usual fishbowl configuration has an inner group discussing an issue or topic while the outer group listens, looking for themes, patterns, or soundness of argument or uses a group behaviour checklist to give feedback to the group on its functioning. The roles may then be reversed.

Crossover groups

Students are divided into subgroups that are subsequently split up to form new groups in such a way as to maximise the crossing over of information. A colour or number coding in the first groupings enables a simple relocation—from, for example, three groups of four students to four groups of three, with each group in the second configuration having one from each of the first.

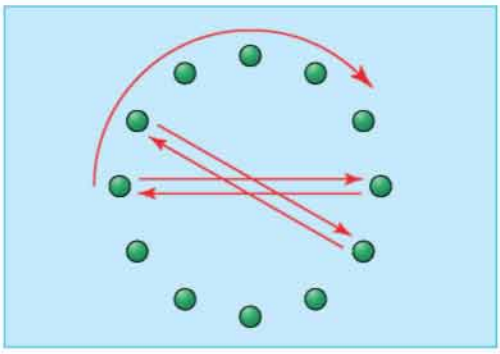


Crossover groups—redistribution of 12 students (each allocated one of four colours) for second period of session

Circular questioning

In circular questioning each member of the group asks a question in turn. In its simplest version, one group member formulates a question relevant to the theme or problem and puts it to the person opposite, who has a specified time (say, one

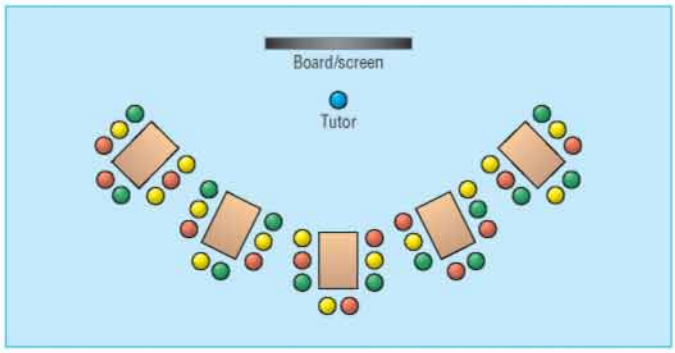
or two minutes) to answer it. Follow up questions can be asked if time permits. The questioning and answering continues clockwise round the group until everyone has contributed, at which time a review of questions and answers can take place. This could also include answers that others would like to have given. Alternatively, you or the students could prepare the questions on cards. You could also mix the best of the students' questions with some of your own.



Circular questioning

Horseshoe groups

This method allows you to alternate between the lecture and discussion formats, a common practice in workshops. Groups are arranged around tables, with each group in a horseshoe formation with the open end facing the front. You can thus talk formally from the board for a time before switching to presenting a group task. Subsequent reporting from each group can induce boredom. To avoid this danger, the tutor can circulate written reports for comment; get groups to interview each other publicly or get one member of each group to circulate; ask groups to produce and display posters; ask the reporters from each group to form an inner group in a fishbowl formation; or use the crossover method to move students around.



Horseshoe groups

The group structures described require an explicit task and topic, and they are possible only if the furniture is movable. Tutors could also consider experimenting with furniture to see if other group structures work. The physical configuration is a strong determinant of social (and hence learning) processes, as is the sequence of activities

Conclusion

This article has emphasised the choices available to you in working with groups. Some of these involve more skilled and sensitive handling of group process from within the group; others require imaginative management in the setting of tasks and the organising of purposeful activities for subgroups. Well organised and purposeful group discussion can create a firm foundation for qualities such as openness, networking, and proactive communication—important ingredients in the process of personal and organisational change. The value of effective group management in professional development and lifelong learning cannot be underestimated.

Recommended reading

- Brookfield S, Preskill S. *Discussion as a way of teaching—tools and techniques for university teachers*. Buckingham: Open University Press, 1999.
- Forster F, Hounsell D, Thompson S. *Tutoring and demonstrating—a handbook*. Sheffield: Universities' and Colleges' Staff Development Agency, 1995.
- Habeshaw T, Habeshaw S, Gibbs G. *53 interesting things to do in your seminars and tutorials*. Bristol: Technical and Educational Services, 1992.
- Jaques D. *Learning in groups*. 3rd ed. London: Kogan Page, 2000.
- Tiberius R. *Small group teaching: a trouble-shooting guide*. London: Kogan Page, 1999.

7 One to one teaching and feedback

Jill Gordon

"My method (is) to lead my students by hand to the practice of medicine, taking them every day to see patients in the public hospital, that they may hear the patients' symptoms and see their physical findings. Then I question the students as to what they have noted in the patients and about their thoughts and perceptions regarding the cause of the illness and the principles of treatment"

Dr Franciscus de la Boe Sylvius, 17th century professor of medicine at the University of Leyden, Netherlands

Dr Franciscus de la Boe Sylvius

Although it is not clear whether Dr Sylvius (above) was describing his teaching method in relation to a group of students or to a succession of individual students, he understood the essential features of clinical education. He understood, for example, the need for active learning in an authentic clinical setting.

Dr Sylvius also understood another important feature of one to one teaching—close behavioural observation (of each other, teacher and learner). No other setting provides the same opportunity for this. Dr Sylvius led his students "by [the] hand." He cared about his role as a teacher. In the closely observed one to one relationship your unguarded statements, your reactions under pressure, and your opinions about other people and the world at large are all magnified.

Just as you cannot hide from the learner, so the learner's knowledge, skills, and attitudes will become apparent to you. Provided that you have created a trusting relationship, you can discuss his or her personal and professional attitudes and values in a way that is seldom possible in a larger group. This is perhaps one of the key benefits of one to one teaching.

Another feature of one to one teaching is the opportunity to adjust what you teach to the learner's needs—"customise" your teaching. In 1978 Ausubel and colleagues suggested that the secret of education is to find out what the learner already knows and teach accordingly. In a lecture, tutorial, or seminar you cannot hope to diagnose and respond to every individual's learning needs, but a one to one relationship provides an opportunity to match the learning experience to the learner.

One to one teaching is perhaps one of the most powerful ways of "influencing students." You can create opportunities for active learning in authentic clinical settings while modelling desirable personal and professional attributes.

Stott and Davis in 1979 promoted the idea that one to one primary care consultations offer exceptional but often unrealised potential. The principles used in primary care consultations can be applied to one to one teaching, and the secret is forethought and planning.

Wards, operating theatres, general practice, and community clinics provide a context for active learning

As a teacher, you are an important role model whether you wish it or not

What's different about one to one teaching?

	Lecture	Seminar	PBL group	Clinical tutorial	One to one clinical attachment
Efficiency*	High	Medium	Low	Low	Very low
Active learning	Low (usually)	Variable	High	Medium to high	Very high
Mutual feedback	Low	Medium	High	Medium to high	Very high
Modelling behaviour in real life setting	Low	Low	Medium	High	Very high

PBL = problem based learning.
*Based on student numbers.

Exceptional potential of one to one teaching

- It tackles current learning needs
- It promotes autonomy and self directed learning
- It links prior knowledge with new clinical experiences
- It enables opportunistic teaching

Plan ahead—ask yourself some important questions

- What is the main purpose of the one to one attachment?
- Do you know why it is part of the learning programme?
- What are the learner's needs?
- How will you gauge how effectively you have met the learner's needs?
- How would you like this learner to describe the experience to a peer?

Provide an orientation

Most of us recall clinical teachers whose social skills amounted to a brief glance and a grunt. Times have changed, or should have. Find out and remember the learner's name—a simple but important courtesy. Outline the special opportunities and benefits that the attachment can provide. Ask the learner to prepare a learning plan and then compare the learner's plan to your own expectations. Once the plan has been agreed, don't shelve it—refer to it during the attachment and modify as necessary.

Find out and remember the learner's name—a simple but important courtesy

Agree on the ground rules

Ground rules are both practical (punctuality, dress, access to patient records) and philosophical (respect for patients and colleagues, confidentiality, consent, openness to different points of view). If these have been spelled out on day 1, you won't be caught out later. Make sure that the learner knows how much time you will be able to spend in observing, teaching, and giving feedback and what you expect in return.

Ask helpful questions

Open ended questions are generally better than closed questions at the beginning of the exchange. A small number of closed questions later in the conversation help you to "diagnose" just how much the learner knows and understands. Avoid questions that require nothing but recall. Try to formulate questions that assume an appropriate amount of knowledge, but build in higher order thinking and/or higher order skills. You might ask the learner, for example, to explain to you (as if you were the patient) the mechanisms behind a condition such as asthma or hypertension. This simulates clinical interface with a patient—testing recall, understanding, and communications skills all at once.

Skilful teaching is not unlike skilful history taking

Give feedback

Learners value feedback highly, and valid feedback is based on observation. Deal with observable behaviours and be practical, timely, and concrete. The one to one relationship enables you to give feedback with sensitivity and in private. Begin by asking the learner to tell you what he or she feels confident of having done well and what he or she would like to improve. Follow up with your own observations of what was done well (be specific), and then outline one or two points that could help the student to improve.

"If musicians learned to play their instruments as physicians learn to interview patients, the procedure would consist of presenting in lectures or maybe in a demonstration or two the theory and mechanisms of the music-producing ability of the instrument and telling him to produce a melody. The instructor of course, would not be present to observe or listen to the student's efforts, but would be satisfied with the student's subsequent verbal report of what came out of the instrument."

George Engel, after visiting 70 medical schools in North America

Encourage reflection

Just as many learning opportunities are wasted if they are not accompanied by feedback from an observer, so too are they wasted if the learner cannot reflect honestly on his or her performance. One to one teaching is ideally suited to encouraging reflective practice, because you can model the way a reflective practitioner behaves. Two key skills are *(a)* "unpacking" your clinical reasoning and decision making processes and *(b)* describing and discussing the ethical values and beliefs that guide you in patient care.

Monitor progress

- Identify deficiencies
- Ask the learner, half way through the attachment, to do a self assessment of how things are going. If both you and the learner can identify deficiencies within a safe learning environment, you can work together to tackle them well before the attachment ends
- If you have serious concerns, you have an obligation to make them known to the learner and to the medical school or training authority
- It is not appropriate to diagnose serious problems and hand the learner on to the next stage of training in the hope that the problems will somehow be correct themselves

Use other one to one teachers

Senior medical students, junior doctors, registrars, nurses, and allied health professionals are all potential teachers. When

junior colleagues interact with a learner, you can encourage them with positive feedback on their teaching.

Every patient interview and every physical examination places the learner in a privileged relationship with a patient. We all have patients whom we especially admire—particularly people who have coped bravely with a chronic illness or a major disability, a disaster such as war, or other misfortunes. Such patients activate an emotional response in the learner, imprinting an enriched memory of the patient and the patient's illness.

Promote active learning

- Time is limited in most clinical settings, and it can be tempting to revert to a passive observational teaching model
- Think about strategies to promote active learning
- Brief students to observe specific features of a consultation or procedure
- Ask patients for permission for the learner to carry out all or part of the physical examination or a procedure while you observe
- If space is available, allow students to interview patients in a separate room or cubicle before presenting them to you
- If possible videotape consultations for a debriefing session at a more convenient time
- Arrange for the learner to see the same patient over time, or in another context, such as a home visit

Reap the rewards

The role of the teacher is frequently undervalued, and yet teaching is potentially rewarding and enjoyable. It is also one of the defining features of a profession. Without teaching to ensure the transmission of knowledge, medicine becomes just another "job."

One to one teaching inevitably exposes you to evaluation by the learner. If the learner trusts you, he or she will be able to tell you what has worked well, and what could be improved. Respond to feedback by reflecting rather than by explaining, excusing, or offering counter arguments to defend your particular style. Openness to feedback is another professional attribute that is best modelled one to one.

Learners for whom you have been a role model and mentor are likely to repay you many times over. Sometimes you have the opportunity to observe their personal and professional development long after the one to one attachment has finished. Dr Sylvius knew how to do it, and Cicero understood its rejuvenating qualities.

"Cultivate the society of the young, remain interested and never stop learning" (Cicero)

Points to remember

Do

- Welcome
- Set shared achievable goals
- Put yourself in the learner's shoes
- Ask interesting questions
- Monitor progress and give feedback
- Encourage

Don't

- Appear unprepared
- Be vague about your expectations
- Confine the learner to passive roles
- Be "nit-picking"
- Leave feedback to the final assessment
- Humiliate

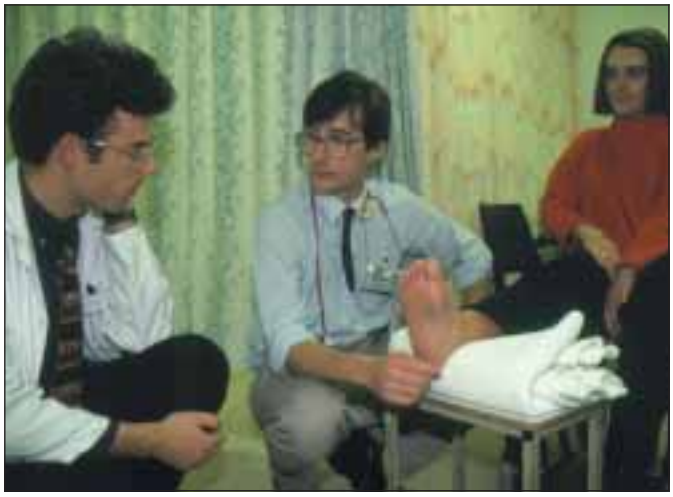
Further reading

- Parsell G, Bligh J. Recent perspectives on clinical teaching. *Med Educ* 2001;35:409-14.
- Paulman PM, Susman JL, Abboud CA, eds. *Precepting medical students in the office*. London: Johns Hopkins University Press, 2000.
- Westberg J, Jason H. *Collaborative clinical education: the foundation of effective health care*. New York: Springer, 1993.
- Whitehouse C, Roland M, Campion P, eds. *Teaching medicine in the community: a guide for undergraduate education*. New York: Oxford University Press, 1997.
- Stott N, Davis R. The exceptional potential in each primary care consultation. *J R Coll Gen Pract* 1979;29:201-5.
- Ausubel D, Novak J, Hanesian H. *Educational psychology: a cognitive view*. New York: Rinehart and Winston, 1978.

8 Learning and teaching in the clinical environment

John Spencer

Clinical teaching—that is, teaching and learning focused on, and usually directly involving, patients and their problems—lies at the heart of medical education. At undergraduate level, medical schools strive to give students as much clinical exposure as possible; they are also increasingly giving students contact with patients earlier in the course. For postgraduates, "on the job" clinical teaching is the core of their professional development. How can a clinical teacher optimise the teaching and learning opportunities that arise in daily practice?



Clinical teaching in general practice

Strengths, problems, and challenges

Learning in the clinical environment has many strengths. It is focused on real problems in the context of professional practice. Learners are motivated by its relevance and through active participation. Professional thinking, behaviour, and attitudes are "modelled" by teachers. It is the only setting in which the skills of history taking, physical examination, clinical reasoning, decision making, empathy, and professionalism can be taught and learnt as an integrated whole. Despite these potential strengths, clinical teaching has been much criticised for its variability, lack of intellectual challenge, and haphazard nature. In other words, clinical teaching is an educationally sound approach, all too frequently undermined by problems of implementation.

Challenges of clinical teaching

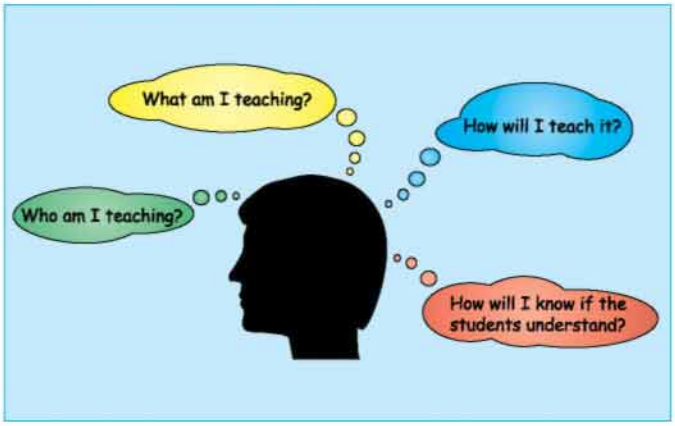
- Time pressures
- Competing demands—clinical (especially when needs of patients and students conflict); administrative; research
- Often opportunistic—makes planning more difficult
- Increasing numbers of students
- Fewer patients (shorter hospital stays; patients too ill or frail; more patients refusing consent)
- Often under-resourced
- Clinical environment not "teaching friendly" (for example, hospital ward)
- Rewards and recognition for teachers poor

Common problems with clinical teaching

- Lack of clear objectives and expectations
- Focus on factual recall rather than on development of problem solving skills and attitudes
- Teaching pitched at the wrong level (usually too high)
- Passive observation rather than active participation of learners
- Inadequate supervision and provision of feedback
- Little opportunity for reflection and discussion
- "Teaching by humiliation"
- Informed consent not sought from patients
- Lack of respect for privacy and dignity of patients
- Lack of congruence or continuity with the rest of the curriculum

The importance of planning

Many principles of good teaching, however, can (and should) be incorporated into clinical teaching. One of the most important is the need for planning. Far from compromising spontaneity, planning provides structure and context for both teacher and students, as well as a framework for reflection and evaluation. Preparation is recognised by students as evidence of a good clinical teacher.



Questions to ask yourself when planning a clinical teaching session

How doctors teach

Almost all doctors are involved in clinical teaching at some point in their careers, and most undertake the job conscientiously and enthusiastically.

However, few receive any formal training in teaching skills, and in the past there has been an assumption that if a person simply knows a lot about their subject, they will be able to teach it. In reality, of course, although subject expertise is important, it

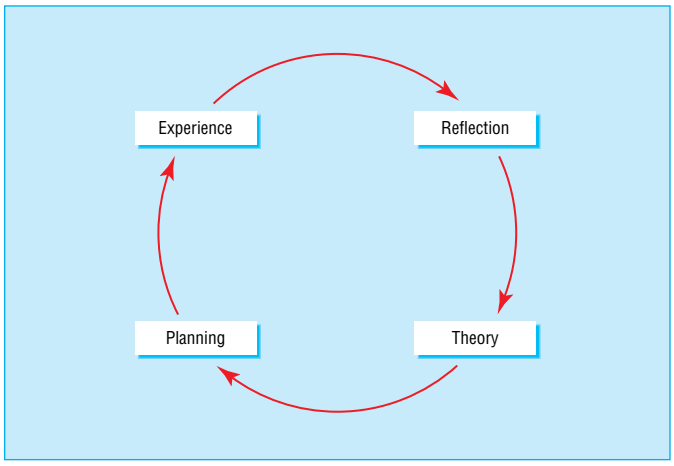
is not sufficient. Effective clinical teachers use several distinct, if overlapping, forms of knowledge.

How students learn

Understanding the learning process will help clinical teachers to be more effective. Several theories are relevant (see first article in the series, 25 January). All start with the premise that learning is an active process (and, by inference, that the teacher's role is to act as facilitator). Cognitive theories argue that learning involves processing information through interplay between existing knowledge and new knowledge. An important influencing factor is what the learner knows already. The quality of the resulting new knowledge depends not only on "activating" this prior knowledge but also on the degree of elaboration that takes place. The more elaborate the resulting knowledge, the more easily it will be retrieved, particularly when learning takes place in the context in which the knowledge will be used.

Experiential learning

Experiential learning theory holds that learning is often most effective when based on experience. Several models have been described, the common feature being a cyclical process linking concrete experience with abstract conceptualisation through reflection and planning. Reflection is standing back and thinking about experience (What did it mean? How does it relate to previous experience? How did I feel?). Planning involves anticipating the application of new theories and skills (What will I do next time?). The experiential learning cycle, which can be entered at any stage, provides a useful framework for planning teaching sessions.

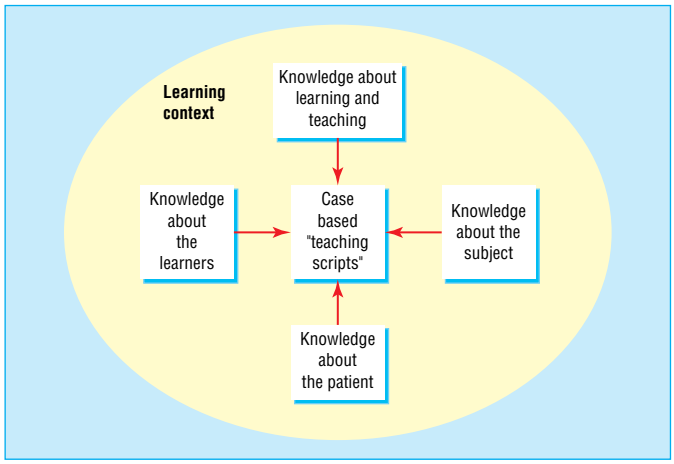


Experiential learning cycle: the role of the teacher is to help students to move round, and complete, the cycle

Questions

Questions may fulfil many purposes, such as to clarify understanding, to promote curiosity, and to emphasise key points. They can be classified as "closed," "open," and "clarifying" (or "probing") questions.

Closed questions invoke relatively low order thinking, often simple recall. Indeed, a closed question may elicit no response at all (for example, because the learner is worried about being wrong), and the teacher may end up answering their own question.



Various domains of knowledge contribute to the idiosyncratic teaching strategies ("teaching scripts") that tutors use in clinical settings

How to use cognitive learning theory in clinical teaching

Help students to identify what they already know

- "Activate" prior knowledge through brainstorming and briefing

Help students elaborate their knowledge

- Provide a bridge between existing and new information—for example, use of clinical examples, comparisons, analogies
- Debrief the students afterwards
- Promote discussion and reflection
- Provide relevant but variable contexts for the learning

Example of clinical teaching session based on experiential learning cycle

Setting—Six third year medical students doing introductory clinical skills course based in general practice

Topic—History taking and physical examination of patients with musculoskeletal problems (with specific focus on rheumatoid arthritis); three patients with good stories and signs recruited from the community

The session

Planning—Brainstorm for relevant symptoms and signs: this activates prior knowledge and orientates and provides framework and structure for the task

Experience—Students interview patients in pairs and do focused physical examination under supervision: this provides opportunities to implement and practise skills

Reflection—Case presentations and discussion: feedback and discussion provides opportunities for elaboration of knowledge

Theory—Didactic input from teacher (basic clinical information about rheumatoid arthritis): this links practice with theory

Planning— "What have I learned?" and "How will I approach such a patient next time?" Such questions prepare students for the next encounter and enable evaluation of the session

Effective teaching depends crucially on the teacher's communication skills. Two important areas of communication for effective teaching are questioning and giving explanations. Both are underpinned by attentive listening (including sensitivity to learners' verbal and non-verbal cues). It is important to allow learners to articulate areas in which they are having difficulties or which they wish to know more about

In theory, open questions are more likely to promote deeper thinking, but if they are too broad they may be equally ineffective. The purpose of clarifying and probing questions is self evident.

Questions can be sequenced to draw out contributions or be built on to promote thinking at higher cognitive levels and to develop new understanding

Explanation

Teaching usually involves a lot of explanation, ranging from the (all too common) short lecture to "thinking aloud." The latter is a powerful way of "modelling" professional thinking, giving the novice insight into experts' clinical reasoning and decision making (not easily articulated in a didactic way). There are close analogies between teacher-student and doctor-patient communication, and the principles for giving clear explanations apply. If in doubt, pitch things at a low level and work upwards. As the late Sydney Jacobson, a journalist, said, "Never underestimate the person's intelligence, but don't overestimate their knowledge." Not only does a good teacher avoid answering questions, but he or she also questions answers.

Exploiting teaching opportunities

Most clinical teaching takes place in the context of busy practice, with time at a premium. Many studies have shown that a disproportionate amount of time in teaching sessions may be spent on regurgitation of facts, with relatively little on checking, probing, and developing understanding. Models for using time more effectively and efficiently and integrating teaching into day to day routines have been described. One such, the "one-minute preceptor," comprises a series of steps, each of which involves an easily performed task, which when combined form an integrated teaching strategy.

Teaching on the wards

Despite a long and worthy tradition, the hospital ward is not an ideal teaching venue. None the less, with preparation and forethought, learning opportunities can be maximised with minimal disruption to staff, patients, and their relatives.

Approaches include teaching on ward rounds (either dedicated teaching rounds or during "business" rounds); students seeing patients on their own (or in pairs—students can learn a lot from each other) then reporting back, with or without a follow up visit to the bedside for further discussion; and shadowing, when learning will inevitably be more opportunistic.

Key issues are careful selection of patients; ensuring ward staff know what's happening; briefing patients as well as students; using a side room (rather than the bedside) for discussions about patients; and ensuring that all relevant information (such as records and *x* ray films) is available.

Teaching in the clinic

Although teaching during consultations is organisationally appealing and minimally disruptive, it is limited in what it can achieve if students remain passive observers.

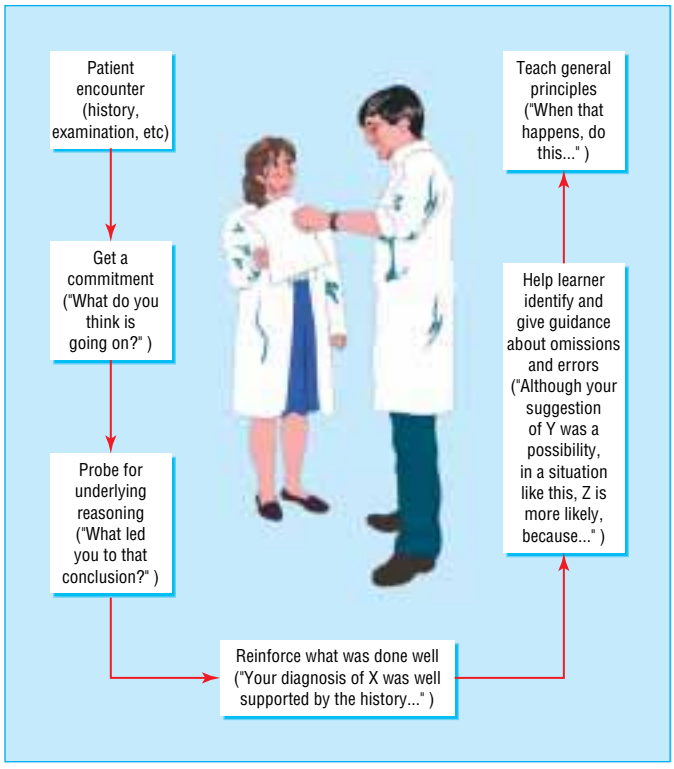
With relatively little impact on the running of a clinic, students can participate more actively. For example, they can be

How to use questions

- Restrict use of closed questions to establishing facts or baseline knowledge (What? When? How many?)
- Use open or clarifying/probing questions in all other circumstances (What are the options? What if?)
- Allow adequate time for students to give a response—don't speak too soon
- Follow a poor answer with another question
- Resist the temptation to answer learners' questions—use counter questions instead
- Statements make good questions—for example, "students sometimes find this difficult to understand" instead of "Do you understand?" (which may be intimidating)
- Be non-confrontational—you don't need to be threatening to be challenging

How to give effective explanations

- Check understanding before you start, as you proceed, and at the end—non-verbal cues may tell you all you need to know about someone's grasp of the topic
- Give information in "bite size" chunks
- Put things in a broader context when appropriate
- Summarise periodically ("so far, we've covered . . .") and at the end; asking learners to summarise is a powerful way of checking their understanding
- Reiterate the take home messages; again, asking students will give you feedback on what has been learnt (but be prepared for some surprises)



"One-minute preceptor" model

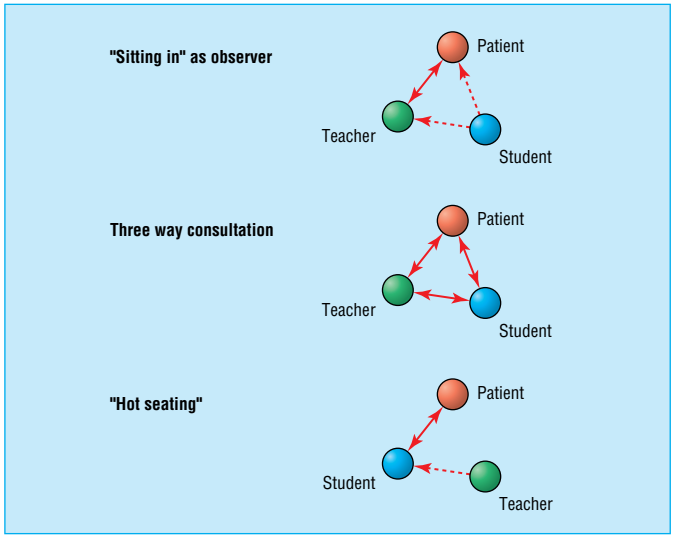
Teaching during consultations has been much criticised for not actively involving learners

asked to make specific observations, write down thoughts about differential diagnosis or further tests, or note any questions—for discussion between patients. A more active approach is "hot seating." Here, the student leads the consultation, or part of it. His or her findings can be checked with the patient, and discussion and feedback can take place during or after the encounter. Students, although daunted, find this rewarding. A third model is when a student sees a patient alone in a separate room, and is then joined by the tutor. The student then presents their findings, and discussion follows. A limitation is that the teacher does not see the student in action. It also inevitably slows the clinic down, although not as much as might be expected. In an ideal world it would always be sensible to block out time in a clinic to accommodate teaching.

The patient's role

Sir William Osler's dictum that "it is a safe rule to have no teaching without a patient for a text, and the best teaching is that taught by the patient himself" is well known. The importance of learning from the patient has been repeatedly emphasised. For example, generations of students have been exhorted to "listen to the patient—he is telling you the diagnosis." Traditionally, however, a patient's role has been essentially passive, the patient acting as interesting teaching material, often no more than a medium through which the teacher teaches. As well as being potentially disrespectful, this is a wasted opportunity. Not only can patients tell their stories and show physical signs, but they can also give deeper and broader insights into their problems. Finally, they can give feedback to both learners and teacher. Through their interactions with patients, clinical teachers—knowingly or unknowingly—have a powerful influence on learners as role models.

Drs Gabrielle Greveson and Gail Young gave helpful feedback on early drafts.



Seating arrangements for teaching in clinic or surgery

Working effectively and ethically with patients

- Think carefully about which parts of the teaching session require direct patient contact—is it necessary to have a discussion at the bedside?
- Always obtain consent from patients before the students arrive
- Ensure that students respect the confidentiality of all information relating to the patient, verbal or written
- Brief the patient before the session—purpose of the teaching session, level of students' experience, how the patient is expected to participate
- If appropriate, involve the patient in the teaching as much as possible
- Ask the patient for feedback—about communication and clinical skills, attitudes, and bedside manner
- Debrief the patient after the session—they may have questions, or sensitive issues may have been raised

Suggested reading

- Cox K. Planning bedside teaching. (Parts 1 to 8.) *Med J Australia* 1993;158:280-2, 355-7, 417-8, 493-5, 571-2, 607-8, 789-90, and 159:64-5.
- Parsell G, Bligh J. Recent perspectives on clinical teaching. *Med Educ* 2001;35:409-14.
- Hargreaves DH, Southworth GW, Stanley P, Ward SJ. *On-the-job learning for physicians.* London: Royal Society of Medicine, 1997.

9 Written assessment

Lambert W T Schuwirth, Cees P M van der Vleuten

Choosing the most appropriate type of written examination for a certain purpose is often difficult. This article discusses some general issues of written assessment then gives an overview of the most commonly used types, together with their major advantages and disadvantages

Some misconceptions about written assessment may still exist, despite being disproved repeatedly by many scientific studies. Probably the most important misconception is the belief that the format of the question determines what the question actually tests. Multiple choice questions, for example, are often believed to be unsuitable for testing the ability to solve medical problems. The reasoning behind this assumption is that all a student has to do in a multiple choice question is recognise the correct answer, whereas in an open ended question he or she has to generate the answer spontaneously. Research has repeatedly shown, however, that the question's format is of limited importance and that it is the content of the question that determines almost totally what the question tests.

This does not imply that question formats are always interchangeable—some knowledge cannot be tested with multiple choice questions, and some knowledge is best not tested with open ended questions.

Five criteria can be used to evaluate the advantages and disadvantages of question types: reliability, validity, educational impact, cost effectiveness, and acceptability. Reliability pertains to the accuracy with which a score on a test is determined. Validity refers to whether the question actually tests what it is purported to test.

Educational impact is important because students tend to focus strongly on what they believe will be in the examinations. Therefore they will prepare strategically depending on the question types used. Whether different preparation leads to different types of knowledge is not fully clear, however. When teachers are forced to use a particular question type, they will tend to ask about the themes that can be easily assessed with that question type, and they will neglect the topics for which the question type is less well suited. Therefore, it is wise to vary the question types in different examinations.

Cost effectiveness and acceptability are important as the costs of different examinations have to be taken into account, and even the best designed examination will not survive if it is not accepted by teachers and students.

Reliability

- A score that a student obtains on a test should indicate the score that this student would obtain in any other given (equally difficult) test in the same field ("parallel test")
- A test represents at best a sample—selected from a range of possible questions. So if a student passes a particular test one has to be sure that he or she would not have failed a parallel test, and vice versa
- Two factors influence reliability negatively:

  Sample error—The number of items may be too small to provide a reproducible result

  Sample too narrow—If the questions focus only on a certain element, the scores cannot generalise to the whole discipline

Validity

- The validity of a test is the extent to which it measures what it purports to measure
- Most competencies cannot be observed directly (body length, for example, can be observed directly; intelligence has to be derived from observations). Therefore, in examinations it is important to collect evidence to ensure validity:

  One simple piece of evidence could be, for example, that experts score higher than students on the test

  Alternative approaches include *(a)* an analysis of the distribution of course topics within test elements (a so called blueprint) and *(b)* an assessment of the soundness of individual test items.
- Good validation of tests should use several different pieces of evidence

"True or false" questions

The main advantage of "true or false" questions is their conciseness. A question can be answered quickly by the student, so the test can cover a broad domain. Such questions, however, have two major disadvantages. Firstly, they are quite difficult to construct flawlessly—the statements have to be defensibly true or absolutely false. Teachers must be taught thoroughly how to construct these question types. Secondly, when a student answers a "false" question correctly, we can conclude only that the student knew the statement was false, not that he or she knew the correct fact.

True or false questions are most suitable when the purpose of the question is to test whether students are able to evaluate the correctness of an assumption; in other cases they are best avoided

"Single, best option" multiple choice questions

Multiple choice questions are well known, and there is extensive experience worldwide in constructing them. Their main advantage is the high reliability per hour oftesting—mainly

Multiple choice questions can be used in any form of testing, except when spontaneous generation of the answer is essential, such as in creativity, hypothesising, and writing skills

because they are quick to answer—so a broad domain can be covered. They are often easier to construct than true or false questions and are more versatile. If constructed well, multiple choice questions can test more than simple facts. Unfortunately though, they are often used to test only facts, as teachers often think this is all they are fit for.

Teachers need to be taught well how to write good multiple choice questions

Multiple true or false questions

These questions enable the teacher to ask a question to which there is more than one correct answer. Although they take somewhat longer to answer than the previous two types, their reliability per hour of testing time is not much lower.

Construction, however, is not easy. It is important to have sufficient distracters (incorrect options) and to find a good balance between the number of correct options and distracters. In addition, it is essential to construct the question so that correct options are defensibly correct and distracters are defensibly incorrect. A further disadvantage is the rather complicated scoring procedure for these questions.

Which of the following drugs belong to the ACE inhibitor group?

(a) atenolol
(b) pindolol
(c) amiloride
(d) furosemide (frusemide)
(e) enalapril
(f) clopamide
(g) epoprostenol
(h) metoprolol
(i) propranolol
(j) triamterene
(k) captopril
(l) verapamil
(m) digoxin

Example of a multiple, true or false question

"Short answer" open ended questions

Open ended questions are more flexible—in that they can test issues that require, for example, creativity, spontaneity—but they have lower reliability. Because answering open ended questions is much more time consuming than answering multiple choice questions, they are less suitable for broad sampling. They are also expensive to produce and to score. When writing open ended questions it is important to describe clearly how detailed the answer should be—without giving away the answer. A good open ended question should include a detailed answer key for the person marking the paper. Short answer, open ended questions are not suitable for assessing factual knowledge; use multiple choice questions instead.

Short answer, open ended questions should be aimed at the aspects of competence that cannot be tested in any other way.

Open ended questions are perhaps the most widely accepted question type. Their format is commonly believed to be intrinsically superior to a multiple choice format. Much evidence shows, however, that this assumed superiority is limited

Essays

Essays are ideal for assessing how well students can summarise, hypothesise, find relations, and apply known procedures to new situations. They can also provide an insight into different aspects of writing ability and the ability to process information. Unfortunately, answering them is time consuming, so their reliability is limited.

When constructing essay questions, it is essential to define the criteria on which the answers will be judged. A common pitfall is to "over-structure" these criteria in the pursuit of objectivity, and this often leads to trivialising the questions. Some structure and criteria are necessary, but too detailed a structure provides little gain in reliability and a considerable loss of validity. Essays involve high costs, so they should be used sparsely and only in cases where short answer, open ended questions or multiple choice questions are not appropriate.

"Key feature" questions

In such a question, a description of a realistic case is followed by a small number of questions that require only essential decisions; these questions may be either multiple choice or open ended, depending on the content of the question. Key feature questions seem to measure problem solving ability

"Key feature" questions aim to measure problem solving ability validly without losing too much reliability

validly and have good reliability. In addition, most people involved consider them to be a suitable approach, which makes them more acceptable.

However, the key feature approach is rather new and therefore less well known than the other approaches. Also, construction of the questions is time consuming; inexperienced teachers may need up to three hours to produce a single key feature case with questions. Experienced writers, though, may produce up to four an hour. Nevertheless, these questions are expensive to produce, and large numbers of cases are normally needed to prevent students from memorising cases. Key feature questions are best used for testing the application of knowledge and problem solving in "high stakes" examinations.

Extended matching questions

The key elements of extended matching questions are a list of options, a "lead-in" question, and some case descriptions or vignettes. Students should understand that an option may be correct for more than one vignette, and some options may not apply to any of the vignettes. The idea is to minimise the recognition effect that occurs in standard multiple choice questions because of the many possible combinations between vignettes and options. Also, by using cases instead of facts, the items can be used to test application of knowledge or problem solving ability. They are easier to construct than key feature questions, as many cases can be derived from one set of options. Their reliability has been shown to be good. Scoring of the answers is easy and could be done with a computer.

The format of extended matching questions is still relatively unknown, so teachers need training and practice before they can write these questions. There is a risk of an under-representation of certain themes simply because they do not fit the format. Extended matching questions are best used when large numbers of similar sorts of decisions (for example, relating to diagnosis or ordering of laboratory tests) need testing for different situations.

Conclusion

Choosing the best question type for a particular examination is not simple. A careful balancing of costs and benefits is required. A well designed assessment programme will use different types of question appropriate for the content being tested.

Example of a key feature question

Case

You are a general practitioner. Yesterday you made a house call on Mr Downing. From your history taking and physical examination you diagnosed nephrolithiasis. You gave an intramuscular injection of 100 mg diclofenac, and you left him some diclofenac suppositories. You advised him to take one when in pain but not more than two a day. Today he rings you at 9 am. He still has pain attacks, which respond well to the diclofenac, but since 5 am he has also had a continuous pain in his right side and a fever (38.9°C).

Which of the following is the best next step?
(a) Ask him to wait another day to see how the disease progresses
(b) Prescribe broad spectrum antibiotics
(c) Refer him to hospital for an intravenous pyelogram
(d) Refer him urgently to a urologist

Example of an extended matching question

(a) Campylobacter jejuni, (b) Candida albicans, (c) Giardia lamblia, (d) Rotavirus, (e) Salmonella typhi, (f) Yersinia enterocolitica, (g) Pseudomonas aeruginosa, (h) Escherichia coli, (i) Helicobacter pylori, (j) Clostridium perfringens, (k) Mycobacterium tuberculosis, (l) Shigella flexneri, (m) Vibrio cholerae, (n) Clostridium difficile, (o) Proteus mirabilis, (p) Tropheryma whippelii

For each of the following cases, select (from the list above) the micro-organism most likely to be responsible:

- A 48 year old man with a chronic complaint of dyspepsia suddenly develops severe abdominal pain. On physical examination there is general tenderness to palpation with rigidity and rebound tenderness. Abdominal radiography shows free air under the diaphragm
- A 45 year old woman is treated with antibiotics for recurring respiratory tract infections. She develops a severe abdominal pain with haemorrhagic diarrhoea. Endoscopically a pseudomembranous colitis is seen

Using only one type of question throughout the whole curriculum is not a valid approach

Further reading

- Case SM, Swanson DB. Extended-matching items: a practical alternative to free response questions. *Teach Learn Med* 1993;5:107-15.
- Frederiksen N. The real test bias: influences of testing on teaching and learning. *Am Psychol* 1984;39:193-202.
- Bordage G. An alternative approach to PMPs: the "key-features" concept. In: Hart IR, Harden R, eds. *Further developments in assessing clinical competence; proceedings of the second Ottawa conference.* Montreal: Can-Heal Publications, 1987:59-75.
- Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987;12:220-46.
- Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement* 1982;6(1):1-11.
- Schuwirth LWT. *An approach to the assessment of medical problem solving: computerised case-based testing.* Maastricht: Datawyse Publications, 1998. (Thesis from Department of Educational Development and Research, Maastricht University.)

10 Skill based assessment

Sydney Smee

Skill based assessments are designed to measure the knowledge, skills, and judgment required for competency in a given domain. Assessment of clinical skills has formed a key part of medical education for hundreds of years. However, the basic requirements for reliability and validity have not always been achieved in traditional "long case" and "short case" assessments. Skill based assessments have to contend with case specificity, which is the variance in performance that occurs over different cases or problems. In other words, case specificity means that performance with one patient related problem does not reliably predict performance with subsequent problems.

For a reliable measure of clinical skills, performance has to be sampled across a range of patient problems. This is the basic principle underlying the development of objective structured clinical examinations (OSCEs). Several other structured clinical examinations have been developed in recent years, including modified OSCEs—such as the Royal College of Physicians' Practical Assessment of Clinical Examination Skills (PACES) and the objective structured long case (OSLER). This article focuses mainly on OSCEs to illustrate the principles of skill based assessment.

Written tests can assess knowledge acquisition and reasoning ability, but they cannot so easily measure skills

OSCEs

The objective structured clinical examination (OSCE) was introduced over 30 years ago as a reliable approach to assessing basic clinical skills. It is a flexible test format based on a circuit of patient based "stations."

At each station, trainees interact with a patient or a "standardised patient" to demonstrate specified skills. Standardised patients are lay people trained to present patient problems realistically. The validity of interactions with real patients, however, may be higher than that with standardised patients, but standardised patients are particularly valuable when communication skills are being tested.

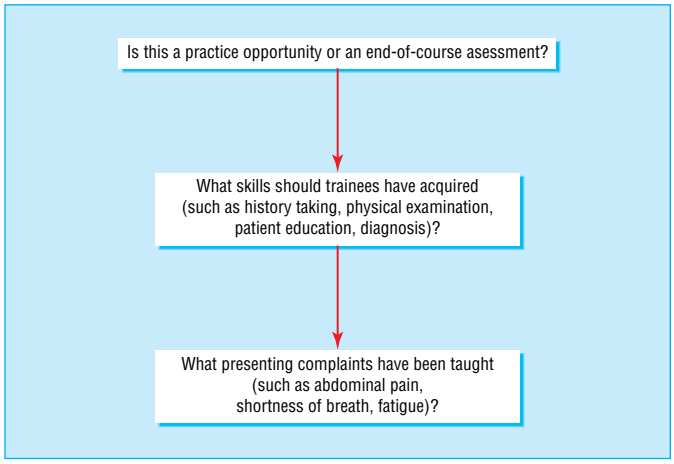
OSCE stations may be short (for eample, five minutes) or long (15-30 minutes). There may be as few as eight stations or more than 20. Scoring is done with a task specific checklist or a combination of checklist and rating scale. The scoring of the students or trainees may be done by observers (for example, faculty members) or patients and standardised patients.

Patient-doctor interaction for assessing clinical performance

Design

The design of an OSCE is usually the result of a compromise between the assessment objectives and logistical constraints; however, the content should always be linked to the curriculum, as this link is essential for validity.

Using many short stations should generate scores that are sufficiently reliable for making pass-fail decisions within a reasonable testing time. (Whether any OSCE is sufficiently reliable for grading decisions is debatable.) Fewer but longer stations maximise learning relative to the selected patient problems, especially when students or trainees receive feedback on their performance. The number of students, time factors, and the availability of appropriate space must also be considered.



Questions to answer when designing an OSCE

Planning

Planning is critical. Patients and standardised patients can be recruited only after stations are written. Checklists must be reviewed before being printed, and their format must be compatible with the marking method, ideally computerised. OSCEs generate a lot of data—for 120 students in a 20 station OSCE there will be 2400 mark sheets!)

Stations are the backbone of an OSCE, and yet the single most common problem is that station materials are incomplete and subject to last minute changes. The result is increased cost and wasted time.

If OSCE scores are being used for making pass-fail decisions, then it is also necessary to set strandards. Several methods for setting standards have been used, with the Angoff method described below being the most commonly used.

Plans should allow sufficient time to process and analyse the scores carefully.

Tasks to do ahead

- Create blueprint
- Set timeline (how long do we need?)
- Get authors for a case-writing workshop
- Review and finalise cases
- Arrange workshop on setting standards
- Recruit standardised patients; recruit faculty members as examiners
- Train standardised patients
- Print marking sheets, make signs
- List all supplies for set-up of OSCE stations
- Remind everyone of date
- Make sure students have all the information
- Plans for the examination day: diagram of station layout; directions for examiners, standardised patients, and staff; possible registration table for students; timing and signals (for example, stopwatch and whistles); procedures for ending the examination
- Anything else?

Costs

OSCE costs vary greatly because the number of stations determines the number of standardised patients, examiners, and staff required. Whether or not faculty members volunteer to write cases, set standards, and examine is also a significant factor.

The fixed costs of running an OSCE remain much the same regardless of the number of examination candidates. Administering an OSCE twice in one day only slightly increases the fixed costs, although the examiners' time is an important cost

Developing the stations

OSCE stations have three components.

Stem

A standardised format for the "stem" (task) is helpful—for example, providing the patient's name, age, presenting complaint, and the setting (such as clinic, emergency, or ward) for all stations. The stem must clearly state the task—for example, "in the next eight minutes, conduct a relevant physical examination."

Checklist

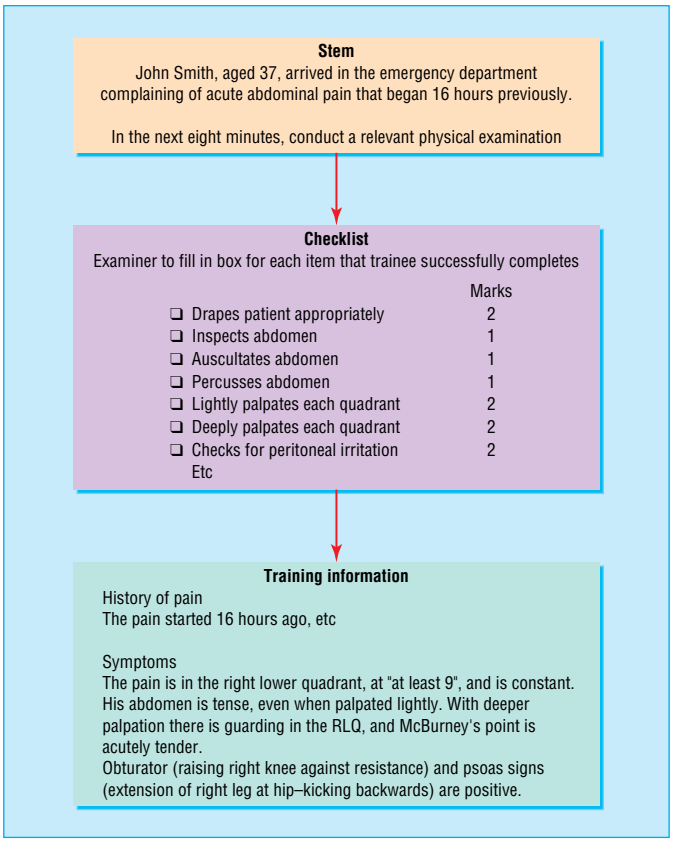
The checklist items are the actions that should be taken in response to the information in the stem. These items should be reviewed and edited to ensure that *(a)* they are appropriate for the level of training being assessed, *(b)* they are task based, and *(c)* they are observable (so the observer can score them).

The length of the checklist depends on the clinical task, the time allowed, and who is scoring. A checklist for a five minute station that is testing history taking may have up to 25 items if a faculty observer is doing the scoring. If a patient or standardised patient is doing the scoring, then fewer items should be used. Use of detailed items will guide scorers: for example, "examines the abdomen" is a general item that might better be separated into a series of items such as "inspects the abdomen," "auscultates the abdomen," "lightly palpates all four quadrants," and so on.

A score must be assigned to every item. Items may be scored 1 or 0, or relative weights may be assigned, with more critical items being worth more. Weights may not change the overall pass-fail rate of an OSCE, but they may improve the validity of a checklist and can affect which trainees pass or fail.

Training information

For standardised patients, directions should use patient based language, specify the patient's perception of the problem (for example, serious, not serious), provide only relevant information, and specify pertinent negatives. Responses to all checklist items should be included. The patient's behaviour and affect should be described in terms of body language, verbal



Components of OSCE station

tone, and pace. Symptoms to be simulated need to be described.

Limitations

Skill based assessments are based on tasks that approximate performance in the area of interest. The assumption is that the closer the tasks are to "real world" tasks, the more valid the assessment.

Three aspects of an OSCE limit how closely the stations approximate clinical practice. Firstly, time limited stations often require trainees to perform isolated aspects of the clinical encounter. This deconstructs the doctor-patient encounter and may be unacceptable for formative assessments. The trade-off is that limiting the time allows for more stations, which can provide performance snapshots that allow for reliable, summative decision making.

Secondly, OSCEs rely on task specific checklists, which assume that the doctor-patient interaction can be described as a list of actions. As a result, checklists tend to emphasise thoroughness, and this may become a less relevant criterion as the clinical experience of candidates increases. Thirdly, there are limits to what can be simulated, and this constrains the nature of the patient problems that can be sampled. Again, this becomes more of an issue as candidates' level of training and clinical experience increases.

Limitations of OSCEs

- Stations often require trainees to perform isolated aspects of the clinical encounter, which "deconstructs" the doctor-patient encounter
- OSCEs rely on task specific checklists, which tend to emphasise thoroughness. But with increasing experience, thoroughness becomes less relevant
- The limitations on what can be simulated constrain the type of patient problems that can be used

None of these limitations is prohibitive, but they should be considered when selecting an OSCE as an assessment tool and when making inferences from OSCE scores

Other approaches to skill based assessment

Traditional approaches

The oral examination (also known as the "viva") and the "long case" have long been used for assessing clinical competence. The oral examination is traditionally an unstructured face to face session with the examiners. This allows them to explore the trainee's understanding of topics deemed relevant to clinical practice. The long case is patient based, but the interaction with the patient is usually not observed. Instead, trainees summarise the patient problem for the examiners and respond to examiners' questions about findings, diagnosis or management, and other topics deemed relevant by examiners. The strength of the long case is the validity that comes from the complexities of a complete encounter with a real patient. However, the difficulty and relevance of these assessments varies greatly as the content is limited to one or two patient problems (selected from the available patients), and decisions are made according to unknown criteria, as examiners make holistic judgments. For this reason traditional unstructured orals and long cases have largely been discontinued in North America.

Alternative formats

Alternative formats tackle the problems associated with traditional orals and long cases by *(a)* having examiners observe the candidate's complete interaction with the patient, *(b)* training examiners to a structured assessment process, and/or *(c)* increasing the number of patient problems. For a short case assessment, for example, one or two examiners may direct a trainee through a series of five or six encounters with real patients. They observe, ask questions, and make a judgment based on the candidate's performance with all the patients. Similarly, a structured oral examination is still a face to face session with examiners, but guidelines for the topics to be covered are provided. Alternatively, a series of patient scenarios and agreed questions may be used so that the content and difficulty of the assessment is standardised across the trainees. Each of these adaptations is aimed at improving reliability, but

An alternative way to assess skills is to observe candidates' interaction with patients

the most important improvement comes from greatly increasing the number of patient problems, which may well cause an impractical increased testing time.

Reliability and validity

The reliability of a test describes the degree to which the test consistently measures what it is supposed to measure. The more reliable a test, the more likely it is that a similar result will be obtained if the test is readministered. Reliability is sensitive to the length of the test, the station or item discrimination, and the heterogeneity of the cohort of candidates. Standardised patients' portrayals, patients' behaviour, examiners' behaviour, and administrative variables also affect reliability.

The validity of a test is a measure of the degree to which the test actually measures what it is supposed to measure. Validity is a property of test scores and justifies their interpretation for a specific purpose. The most basic evidence of validity comes from documenting the links between the content of the assessment and the curriculum's objectives and from the qualifications of those who develop the assessment.

Factors leading to lower reliability

- Too few stations or too little testing time
- Checklists or items that don't discriminate (that is, are too easy or too hard)
- Unreliable patients or inconsistent portrayals by standardised patients
- Examiners who score idiosyncratically
- Administrative problems (such as disorganised staff or noisy rooms)

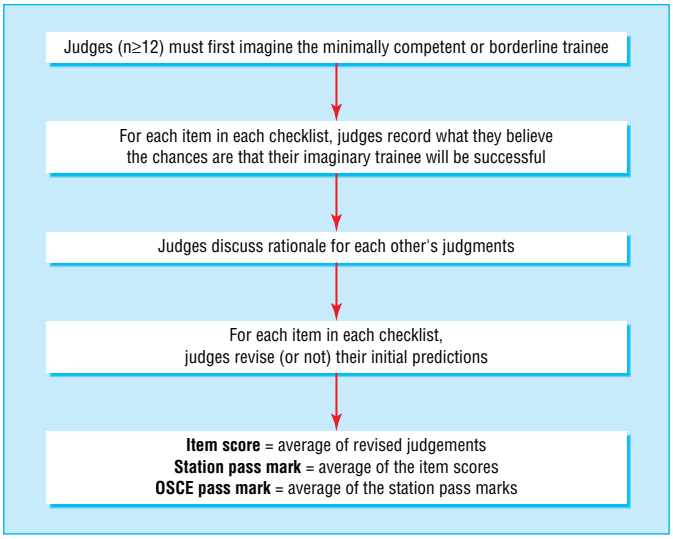
Questions to ensure validity

- Are the patient problems relevant and important to the curriculum?
- Will the stations assess skills that have been taught?
- Have content experts (generalists and specialists) reviewed the stations?

Setting standards

Checklists generate scores; judges set standards. The validity of a standard depends on the judges' qualifications and the reasonableness of the procedure they use to set it. When pass-fail decisions are being made, a skill based assessment should be "criterion referenced" (that is, trainees should be assessed relative to performance standards rather than to each other or to a reference group). An Angoff approach is commonly used to set the standard for an OSCE.

Skill based assessments do not replace knowledge based tests, but they do assess aspects of competence that knowledge based tests cannot assess. Although the use of OSCEs for skill based assessment is increasingly widespread, modifying more traditional formats may be appropriate when they are combined with other forms of assessment or are used to screen trainees. The success of any skill based assessment depends on finding a suitable balance between validity and reliability and between the ideal and the practical.



A modified Angoff procedure for an OSCE

The second picture and the picture showing an oral examination are from Microsoft Clipart.

Further reading

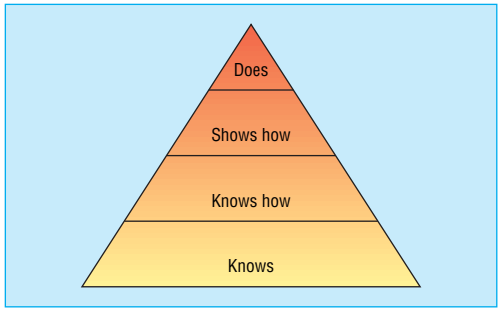
- Gorter S, Rethans JJ, Scherpbier A, van der Heijde D, Houben H, van der Linden S, et al. Developing case-specific checklists for standardized-patient-based assessments in internal medicine: a review of the literature. *Acad Med* 2000;75:1130-7.
- Hodges B, Regehr G, McNaughton N, Tiberius RG, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;74:1129-34.
- Kaufman DM, Mann KV, Muijtjens AMM, van der Vleuten CPM. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med* 2001;75:267-71.
- Newble DI, Dawson B, Dauphinee WD, Page G, Macdonald M, Swanson DB, et al. Guidelines for assessing clinical competence. *Teach Learn Med* 1994;6:213-20.
- Norcini JJ. The death of the long case? *BMJ* 2002;324:408-9.
- Reznick RK, Smee SM, Baumber JS, Cohen R, Rothman AI, Blackmore DE, et al. Guidelines for estimating the real cost of an objective structured clinical examination. *Acad Med* 1993;67:513-7.

11 Work based assessment

John J Norcini

In 1990 psychologist George Miller proposed a framework for assessing clinical competence. At the lowest level of the pyramid is knowledge (knows), followed by competence (knows how), performance (shows how), and action (does). In this framework, Miller distinguished between "action" and the lower levels. "Action" focuses on what occurs in practice rather than what happens in an artificial testing situation. Work based methods of assessment target this highest level of the pyramid and collect information about doctors' performance in their normal practice. Other common methods of assessment, such as multiple choice questions, simulation tests, and objective structured clinical examinations (OSCEs) target the lower levels of the pyramid. Underlying this distinction is the sensible but still unproved assumption that assessments of actual practice are a much better reflection of routine performance than assessments done under test conditions.

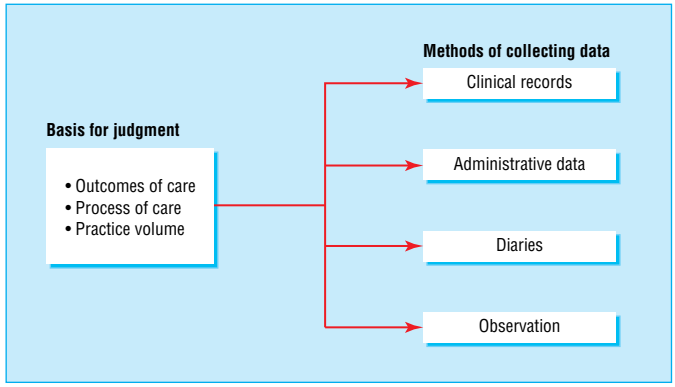
This article explains what is meant by work based assessment and presents a classification scheme for current methods



Miller's pyramid for assessing clinical competence

Methods

Although the focus of this article is on practising doctors, work based assessment methods apply to medical students and trainees as well. These methods can be classified in many ways, but this article classifies them in two dimensions. The first dimension describes the basis for making judgments about the quality of performance. The second dimension is concerned with how data are collected.



Classification for work based assessment methods

Basis for judgment

Outcomes

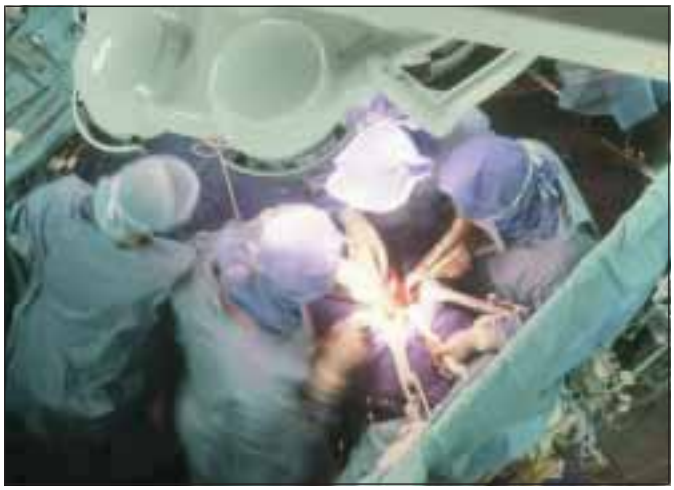
In judgments about the outcomes of their patients, the quality of a cardiologist, for example, might be judged by the mortality of his or her patients within 30 days of acute myocardial infarction. Historically, outcomes have been limited to mortality and morbidity, but in recent years the number of clinical end points has been expanded. Patients' satisfaction, functional status, cost effectiveness, and intermediate outcomes—for example, HbA_{1c} and lipid concentrations for diabetic patients—have gained acceptance.

Patients' outcomes are the best measures of the quality of doctors for the public, the patients, and doctors themselves. For the public, outcomes assessment is a measure of accountability that provides reassurance that the doctor is performing well in practice. For individual patients, it supplies a basis for deciding which doctor to see. For doctors, it offers reassurance that their assessment is tailored to their unique practice and based on real work performance.

Despite the fact that an assessment of outcomes is highly desirable, at least four substantial problems remain. These are attribution, complexity, case mix, and numbers.

Firstly, for a good judgment to be made about a doctor's performance, the patients' outcomes must be attributable solely to that doctor's actions. This is not realistic when care is delivered within systems and teams. Secondly, patients with the same condition will vary in complexity depending on the severity of their illness, the existence of comorbid conditions, and their ability to comply with the doctor's recommendations. Although statistical adjustments may tackle these problems, they are not completely effective. So differences in complexity directly influence outcomes and make it difficult to compare

Three aspects of doctors' performance can be assessed—patients' outcomes, process of care, and volume of practice



Care is delivered in teams, so judging a doctor's performance through outcomes is not realistic

doctors or set standards for their performance. Thirdly, unevenness exists in the case mix of different doctors, again making it difficult to compare performance or to set standards. Finally, to estimate well a doctor's routine performance, a sizeable number of patients are needed. This limits outcomes assessment to the most frequently occurring conditions.

Process of care

In judgments about the process of care that doctors provide, a general practitioner, for example, might be assessed on the basis of how many of his or her patients aged over 50 years have been screened for colorectal cancer. General process measures include screening, preventive services, diagnosis, management, prescribing, education of patients, and counselling. In addition, condition specific processes might also serve as the basis for making judgments about doctors—for example, whether diabetic patients have their HbA_{1c} monitored regularly and receive routine foot examinations.

Measures of process of care have substantial advantages over outcomes. Firstly, the process of care is more directly in the control of the doctor, so problems of attribution are greatly reduced. Secondly, the measures are less influenced by the complexity of patients' problems—for example, doctors continue to monitor HbA_{1c} regardless of the severity of the diabetes. Thirdly, some of the process measures, such as immunisation, should be offered to all patients of a particular type, reducing the problems of case mix.

The major disadvantage of process measures is that simply doing the right thing does not ensure the best outcomes for patients. That a physician regularly monitors HbA_{1c}, for example, does not guarantee that he or she will make the necessary changes in management. Furthermore, although process measures are less susceptible to the difficulties of attribution, complexity, and case mix, these factors still have an adverse influence.

Volume

Judgments about the number of times that doctors have engaged in a particular activity might include, for example, the number of times a surgeon performed a certain surgical procedure. The premise for this type of assessment is the large body of research showing that quality of care is associated with higher volume.

Data collection

Clinical practice records

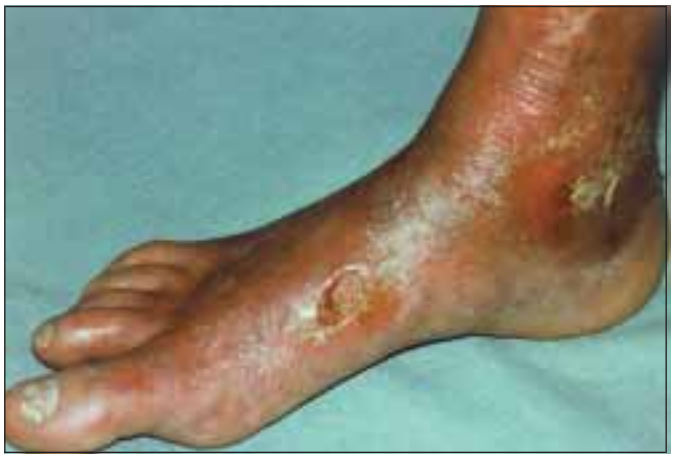
One of the best sources of information about outcomes, process, and volume is the clinical practice record. The external audit of these records is a valid and credible source of data. Two major problems exist, however, with clinical practice records.

Firstly, judgment can be made only on what is recorded —this may not be an accurate assessment of what was actually done in practice.

Secondly, abstracting records is expensive and time consuming and is made cumbersome by the fact that they are often incomplete or illegible.

Widespread adoption of the electronic medical record may be the ultimate solution, although this is some years away. Meanwhile, some groups rely on doctors to abstract their own records and submit them for evaluation. Coupled with an external audit of a sample of the participating doctors, this is a credible and feasible alternative.

Unevenness in case mix can reduce usefulness of using patients' outcomes as a measure of doctors' competence



Judgments on process of care might include foot examinations for diabetic patients

For a sound assessment of an individual doctor's process of care, a sizeable number of patients need to be included

Advantages of volume based assessment over assessment of outcomes and process

- Problems of attribution are reduced substantially
- Complexity is eliminated
- Case mix is not relevant

However, such assessment alone offers no assurance that the activity was conducted properly

Traditional medical records may give way to widespread use of electronic records, making data collection easier and quicker

Administrative databases

In some healthcare systems large computerised databases are often developed as part of the process of administering and reimbursing for health care. Data from these sources are accessible, inexpensive, and widely available. They can be used in the evaluation of some aspects of practice performance—such as cost effectiveness—and of medical errors. However, the lack of clinical information and the fact that the data are often collected for invoicing purposes makes them unsuitable as the only source of information.

Diaries

Doctors, especially trainees, may use diaries or logs to record the procedures they perform. Depending on the purpose of the diary, entries can be accompanied by a description of the doctor's role, the name of an observer, an indication of whether it was done properly, and a list of complications. This is a reasonable way to collect data on volume and an acceptable alternative to the abstraction of clinical practice records until medical records are kept electronically.

Observation

Data can be collected in many ways through practice observation, but to be consistent with Miller's definition of work based assessment, the observations need to be routine or covert to avoid an artificial test situation. They can be made in any number of ways and by any number of different observers. The most common forms of observation based assessment are ratings by supervisors, peers, and patients. Other examples of observation include visits by standardised patients (lay people trained to present patient problems realistically) to doctors in their surgeries and audiotapes or videotapes of consultations such as those used by the General Medical Council.

Portfolios

Doctors typically collect from various sources the practice data they think pertinent to their evaluation. A doctor's portfolio might contain data on outcomes, process, or volume, collected through clinical record audit, diaries, or assessments by patients and peers. It is important to specify what to include in portfolios as doctors will naturally present their best work, and the evaluation of it will not be useful for continuing quality improvement or quality assurance. In addition, if there is a desire to compare doctors or to provide them with feedback about their relative performance, then all portfolios must contain the same data collected in a similar fashion. Otherwise, there is no basis for legitimate comparison or benchmarking.

Further reading

- McKinley RK, Fraser RC, Baker R. Model for directly assessing and improving competence and performance in revalidation of clinicians. *BMJ* 2001;322:712-5.
- Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990:S63-7.
- Norcini JJ. Recertification in the United States. *BMJ* 1999;319:1183-5.
- Cunningham JPW, Hanna E, Turnbull J, Kaigas T, Norman GR. Defensible assessment of the competency of the practicing physician. *Acad Med* 1997;72:9-12.

Databases for clinical audit are becoming more available and may provide more useful information relating to clinical practice

Peer evaluation rating forms

Below are the aspects of competence assessed with the peer rating form developed by Ramsey and colleagues.* The form, given to 10 peers, provides reliable estimates of two overall dimensions of performance: cognitive and clinical skills, and professionalism.

Cognitive and clinical skills
- Medical knowledge
- Ambulatory care
- Management of complex problems
- Management of hospital inpatients
- Problem solving
- Overall clinical competence

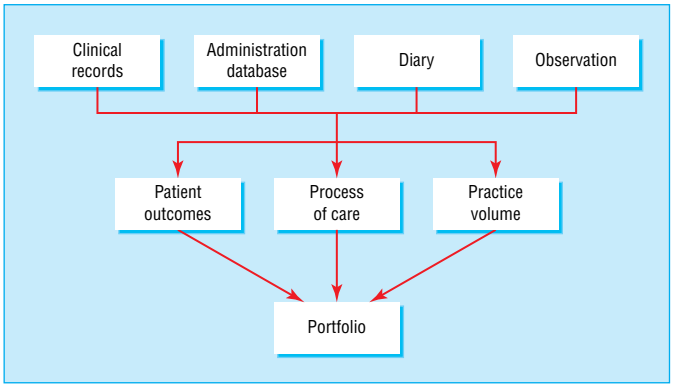
Professionalism
- Respect
- Integrity
- Psychosocial aspects of illness
- Compassion
- Responsibility

*Ramsey PG et al. Use of peer ratings to evaluate physician performance. *JAMA* 1993;269:1655-60

Patient rating form*

The form is given to 25 patients and gives a reliable estimate of a doctor's communication skills. Scores are on a five point scale—poor to excellent—and are related to validity measures. The patients must be balanced in terms of age, sex, and health status. Typical questions are:
- Tells you everything?
- Greets you warmly?
- Treats you as if you're on the same level?
- Lets you tell your story?
- Shows interest in you as a person?
- Warns you what is coming during the physical examination?
- Discusses options?
- Explains what you need to know?
- Uses words you can understand?

*Webster GD. *Final report of the patient satisfaction questionnaire study*. American Board of Internal Medicine, 1989



How portfolios are compiled

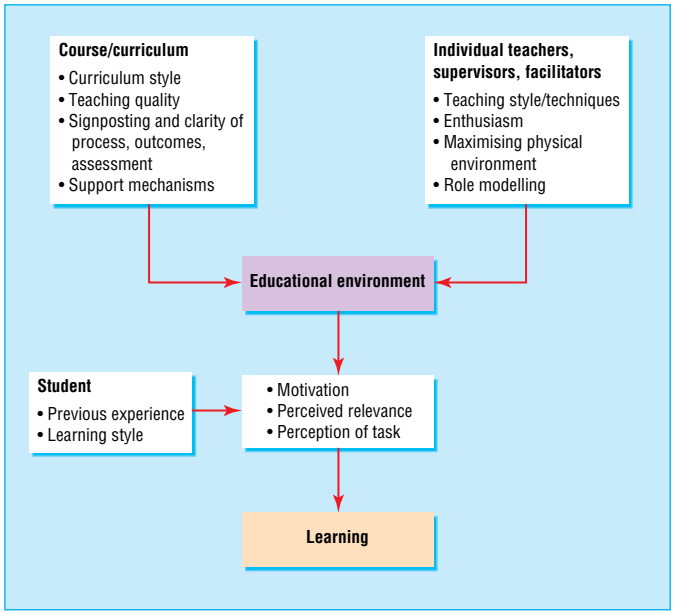
The photograph of a surgical team is from Philippe Plailly/Eurelios/SPL; the photographs illustrating case mix are from Photofusion, by David Tothill (left) and Pete Addis (right); the photograph of the foot is from Ray Clarke (FRPS) and Mervyn Goff (FRPS/SPL); and the medical records photograph is from Michael Donne/SPL.

12 Educational environment

Linda Hutchinson

A student might find a particular question threatening and intimidating in one context yet stimulating and challenging in a different context. What makes one learning context unpleasant and another pleasant?

Learning depends on several factors, but a crucial step is the engagement of the learner. This is affected by their motivation and perception of relevance. These, in turn, can be affected by learners' previous experiences and preferred learning styles and by the context and environment in which the learning is taking place. In adult learning theories, teaching is as much about setting the context or climate for learning as it is about imparting knowledge or sharing expertise.



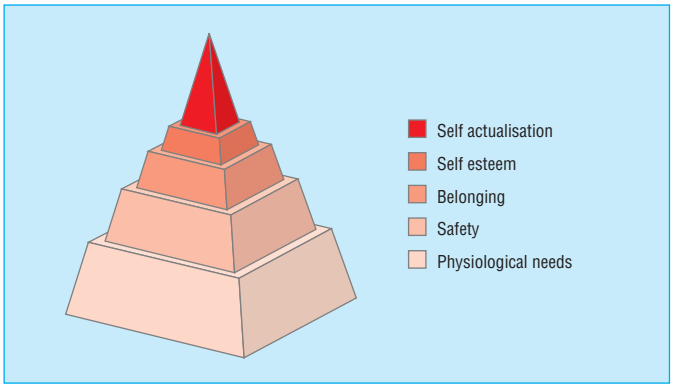
Many factors influence learning

Motivation

Motivation can be intrinsic (from the student) and extrinsic (from external factors). Assessments are usually a strong extrinsic motivator for learners. Individual learners' intrinsic motivation can be affected by previous experiences, by their desire to achieve, and the relevance of the learning to their future.

A teacher's role in motivation should not be underestimated. Enthusiasm for the subject, interest in the students' experiences, and clear direction (among other things) all help to keep students' attention and improve assimilation of information and understanding.

Even with good intrinsic motivation, however, external factors can demotivate and disillusion. Distractions, unhelpful attitudes of teachers, and physical discomfort will prompt learners to disengage. Maslow described a model to illustrate the building blocks of motivation. Each layer needs to be in place before the pinnacle of "self actualisation" is reached.



Maslow's hierarchy of needs for motivating learning. Adapted from Maslow A H. *Motivation and Personality*, New York: Harper and Row, 1954

Physiological needs

Although the need to be fed, watered, and comfortable seems trite, many teachers will have experienced, for example, the difficulties of running sessions in cold or overheated rooms, in long sessions without refreshments, in noisy rooms, in facilities with uncomfortable seating.

Physical factors can make it difficult for learners and teachers to relax and pay attention. Ensuring adequate breaks and being mindful of the physical environment are part of the teacher's role.

Safety

A teacher should aim to provide an environment in which learners feel safe to experiment, voice their concerns, identify their lack of knowledge, and stretch their limits. Safety can be compromised, for example, through humiliation, harassment, and threat of forced disclosure of personal details.

Teachers can create an atmosphere of respect by endorsing the learners' level of knowledge and gaps in knowledge as essential triggers to learning rather than reasons for ridicule.

Remembering names and involving the learners in setting ground rules are other examples of building mutual trust. Feedback on performance, a vital part of teaching, should be done constructively and with respect for the learner.

Case history: safe environment

Dr Holden claims to use interactive teaching techniques. She introduces the topic then points to a student in the audience and says, "tell me five causes of cyanosis." Each student will get asked questions in the session, and most spend their time worrying about when it is their turn to be "exposed."

After a course on teaching skills, she runs the same session. After the introduction, she tells the students to turn to their neighbour and together come up with some possible causes of cyanosis. After a couple of minutes of this "buzz group" activity, she asks for one suggestion from each group until no new suggestions are made. The suggestions are then discussed with reference to the aims of the session.

Belonging

Many factors help to give a student a sense of belonging in a group or team—for example, being a respected member, having one's voice heard and attended to, being given a useful role, and having colleagues with similar backgrounds, experiences, and goals.

Learners are motivated through inclusion and consultation. Their input to a course's objectives and structure should be sought, valued, and acted on. On clinical placements, staff should help to prevent medical students from feeling ignored, marginalised, or "in the way." Students should instead be valued as assets to a clinical unit or team.

Case history: sense of belonging

Five medical students arrive at a distant hospital for a four week attachment. They are met by a staff member, shown around the unit, canteen, library, and accommodation. They are given name badges in the style of the existing staff, and after a couple of days are given specific roles on the unit. These roles develop over the weeks. There is little set teaching time, but the students feel free to ask any staff member for more details at quieter times.

After the attachment, they meet up with friends who were placed elsewhere. Their experience was different. No one was expecting them when they arrived, and ward and clinic staff were unhelpful. There was a set teaching programme and an enthusiastic teacher, but the students were relieved that the hospital was near a town centre with good shops and nightlife.

Self esteem

Several of the points mentioned above feed directly into self esteem through making the learner feel valued. Praise, words of appreciation, and constructive rather than destructive criticism are important. It can take many positive moments to build self esteem, but just one unkind and thoughtless comment to destroy it.

Doctors are well used to their role in the doctor-patient relationship. Some find it hard to translate the same skills and attitudes to the teacher-student relationship. Their own experience of education or their own distractions, time pressures, and other stresses may be factors.

Case history: self esteem

A senior house officer (SHO) is making slow but steady progress. His confidence is growing and the level of supervision he requires is lessening. On one occasion, however, his case management is less than ideal, although the patient is not harmed or inconvenienced. The consultant feels exasperated and tells the SHO that everyone is carrying him and it still isn't working. The SHO subsequently reverts to seeking advice and permission for all decision making.

Self actualisation

If a teacher has attended to the above motivational factors, then they have sought to provide the ideal environment in which a learner can flourish.

An ethos that encourages intrinsic motivation without anxiety is conducive to a "deep" learning approach. However, there may be some who remain unable to respond to the education on offer. Teachers may need to consider whether the course (or that particular piece of study) is suitable for that student.

Relevance

Students' perception of the relevance of what they are being taught is a vital motivator for learning

The relevance of learning is closely linked to motivation: relevance for immediate needs, for future work, of getting a certificate or degree regardless of content. Learning for learning's sake is back in vogue in higher education after a move towards vocational or industrial preparation.

Certain courses in medical degrees have been notoriously poorly received by students. Faculty members need to explain to students why these courses are necessary and how they link to future practice. Allowing them to see for themselves, through early clinical exposure and experience, is likely to be helpful. Similarly, learning the basic medical sciences in the context of clinical situations is the basis for problem based learning.

A challenging problem is the trainee who is in a post because he or she needs to do it for certification, although it is of no perceived value to the trainee's future career direction. A balance needs to be negotiated between respect for the individual's needs and the expectation of a level of professional conduct.

If a teacher is asked to do a one-off session with learners they don't know, he or she should prepare—both before and at the start of the session—by determining what the learners know, want to know, and expect to learn. This involves and shows respect for the learners and encourages them to invest in the session

Teacher as role model

The teacher or facilitator is one of the most powerful variables in the educational environment. The teacher's actions, attitudes (as evidenced by tone of voice, comments made), enthusiasm, and interest in the subject will affect learners indirectly. The capacity for subliminal messages is enormous. Inappropriate

Case history: role models

Dr Jones is a well known "character" in the hospital. Medical students sitting in his clinic hear him talk disparagingly about nurses, patients, and the new fangled political correctness about getting informed consent that wastes doctors' time. Most students are appalled; but a few find him engaging—they view him as a "real" doctor, unlike the ethics or communication skills lecturers.

behaviour or expression by a staff member will be noticed; at worst the learners will want to emulate that behaviour, at best they will have been given tacit permission to do so.

It is easy to "learn" attitudes—including poor attitudes. Attitudes are learnt through observation of those in relative power or seniority. Teachers must therefore be aware of providing good role modelling in the presence of students

Maximising educational environment

Classroom, tutorials, seminars, lectures

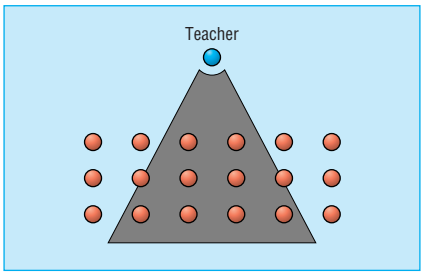
Room temperature, comfort of seating, background noise, and visual distractions are all factors of the environment that can affect concentration and motivation. Some are within the teacher's control, others not.

Respect for the learners and their needs, praise, encouragement of participation can all lead to a positive learning experience. Lack of threat to personal integrity and self esteem is essential, although challenges can be rewarding and enjoyable.

Small group teaching facilitates individual feedback, but the seating arrangement used will have an important effect on student participation. If, for example, students sit in traditional classroom rows, those on the edges will feel excluded. A circular format encourages interaction. It allows the teacher to sit alongside a talkative person, thus keeping them out of eye contact and reducing their input. A quiet student can be placed opposite to encourage participation through non-verbal means. Students can also work in unfacilitated groups on a topic, enabling them to work in teams and share the learning tasks.

Checklist to ensure good physical environment

- Is the room the right size?
- Is the temperature comfortable?
- Are there distractions (noise, visual distractions inside or outside)?
- Is the seating adequate, and how should it be arranged?
- Does the audiovisual equipment work?



Traditional teaching can leave some students excluded (that is, outside the "triangle of influence")

Clinical settings

In real life settings, the dual role of teacher and clinician can be complicated. The students will be closely observing the clinician, picking up hidden messages about clinical practice. They need to feel that there is no danger that they will unnecessarily distress or harm patients or their families. They also need to feel safe from humiliation. Making them feel welcomed and of value when they arrive at a new placement or post will aid their learning throughout.

Checklist for teaching in clinical settings

- Have patients and families given consent for students to be present?
- Do the staff know that teaching is planned and understand what their roles will be?
- Is there adequate space for all participants?
- How much time is available for teaching?
- How may the students be made to feel useful (for example, "pre-clerking" and presenting)?

Course and curriculum design

The designers of short and long courses should consider the relevance of the learning environment to the potential learners. Student representation on curriculum committees is one means of ensuring a more student centred course.

The aims, objectives, and assessments should be signposted well in advance of a course and should be demonstrably fair. The teaching methods should build on learners' experience, creating a collaborative environment. Disseminating the findings of course evaluations, followed by staff training, helps to identify and correct undesirable behaviour among faculty members. Evaluations should also include a means for reviewing the course's aims and objectives with the students.

In longer courses, student support systems and informal activities that build collective identity must be considered. Students who are having difficulties need to be identified early and given additional support.

Further reading

- Newble D, Cannon R. *A handbook for medical teachers.* 3rd ed. London: Kluwer Academic, 1994.
- Eraut M. *Developing professional knowledge and competence.* London: Falmer, 1994.
- Welsh I, Swann C. *Partners in learning: a guide to support and assessment in nurse education.* Abingdon: Radcliffe, 2002.
- Norman GR, Schmidt HG. The psychological basis of problem-based learning: a review of the evidence. *Acad Med* 1992;67:557-65.
- Dent JA, Harden RM. *A practical guide for medical teachers.* London: Churchill Livingstone, 2001.

13 Web based learning

Judy McKimm, Carol Jollie, Peter Cantillon

Many of us use the internet or the "web" (world wide web) as a source of information. In medical education, the web is increasingly used both as a learning tool to support formal programmes and as a means of delivering online learning programmes. What can educators do to ensure that the potential of the web is used effectively to support both their own learning and that of their students?

The technology

Much of the literature on web based learning shows that one of the main barriers to the effective use of teaching materials is the technology (for example, poor access, slow downloading) rather than the design of the learning materials themselves. Some of these issues are discussed later in the article, but it is vital that teachers take on expert help with technical issues in the planning, design, and delivery of web based learning programmes. Through programming and the use of "plug-ins" (programs that can be downloaded from the internet), designers can produce interactive course materials containing online activities (such as self assessments), animations, and simulations. These can improve learning and are often more enjoyable and meaningful for learners.

Distance learning

Two of the main developments in web based learning have been the adaptation of communication technology to support learning and the changes in distance learning strategies necessary for delivering online courses. Both aspects should be considered when designing or delivering web based learning programmes. Lessons can be learned by considering how distance education evolved.

Distance and open learning began with correspondence courses. The Open University in Britain is one of the best known examples of how university level education became accessible, through effective distance learning, to people who had neither the traditional qualifications nor the time to enter full time higher education.

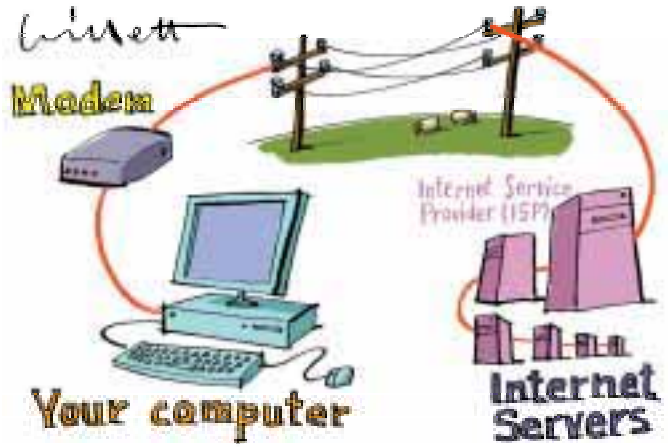
The secret of the Open University's success lies in clearly identifying students' needs; providing effective, local support; and combining conventionally taught components with the use of up to date multimedia resources, including books, course guides, videotapes, audiotapes, television, e-conferencing, and discussion groups.

What is web based learning?

Web based learning is often called online learning or e-learning because it includes online course content. Discussion forums via email, videoconferencing, and live lectures (videostreaming) are all possible through the web. Web based courses may also provide static pages such as printed course materials.

One of the values of using the web to access course materials is that web pages may contain hyperlinks to other parts of the web, thus enabling access to a vast amount of web based information.

A "virtual" learning environment (VLE) or managed learning environment (MLE) is an all in one teaching and learning software package. A VLE typically combines functions such as



Glossary

E-conferencing—Use of online presentations and discussion forums (in real time or stored as downloadable files on a website) to avoid the need for participants to travel

E-learning—Learning through electronic means, such as via the web (see world wide web), an intranet, or other multimedia materials

HTML (hypertext markup language)—The language used to create web pages. HTML files can also contain links to other types of files including wordprocessed files, spreadsheets, presentation slides, and other web pages

Hyperlinks—Links in web pages that enable the user to access another web page (either on the same or a different site) with just one mouse click

Internet—A global network of computers divided into subsets (for example, the web or email systems). Computers are linked to the internet via host computers, which link to other computers via dial up (for example, via a modem) and network connections

Internet service provider (ISP)—Home users usually access the internet through an internet service provider (such as AOL), which maintains a network of PCs permanently connected to the internet

Intranet—A network of computers that share information, usually within an organisation. Access normally requires a password and is limited to a defined range of users

Managed learning environment (MLE)—Usually has an integrated function, providing administrative tools, such as student records, and linking with other management information systems (MLS)

Search engines (such as Lycos, Google)—Can be used to help to find information

Videostreaming—The process by which video images are able to be stored and downloaded on the web. These might be in real time (such as a conference) or used asynchronously

Virtual learning environment (VLE)—A set of electronic teaching and learning tools. Principal components include systems that can map a curriculum, track student activity, and provide online student support and electronic communication

World wide web (web)—Use of the internet to present various types of information. Websites or home pages may be accessed with the aid of a browser program (such as Netscape Communicator or Microsoft Explorer). All such programmes use HTML

For additional information see www.learnthenet.com/english/section/intbas.html

discussion boards, chat rooms, online assessment, tracking of students' use of the web, and course administration. VLEs act as any other learning environment in that they distribute information to learners. VLEs can, for example, enable learners to collaborate on projects and share information. However, the focus of web based courses must always be on the learner—technology is not the issue, nor necessarily the answer.

Models of web based learning

Several approaches can be used to develop and deliver web based learning. These can be viewed as a continuum. At one end is "pure" distance learning (in which course material, assessment, and support is all delivered online, with no face to face contact between students and teachers). At the other end is an organisational intranet, which replicates printed course materials online to support what is essentially a traditional face to face course. However, websites that are just repositories of knowledge, without links to learning, communication, and assessment activities, are not learner centred and cannot be considered true web based learning courses.

In reality, most web based learning courses are a mixture of static and interactive materials, and most ensure that some individual face to face teaching for students is a key feature of the programme.

The individual learner

The first step in designing a web based course is to identify the learners' needs and whether the learners are to be considered as part of a group or as individual learners. The web can be a useful tool for bringing isolated learners together in "virtual" groups—for example, through a discussion forum. There are several online resources on how to design web based learning programmes (for example, at www.ltsn.ac.uk).

Questions to ask before starting a web based learning project

- What is the educational purpose of the web based learning project?
- What added value will online learning bring to the course or to the students?
- What resources and expertise on web based learning exist in the institution?
- Are colleagues and the institution aware of the planned course? (You need to avoid duplication of effort and be sure that the institution's computer system can support the course)
- Has the project taken account of existing teaching resources and ongoing maintenance costs after initial development?
- Have you allowed enough time to develop or redevelop materials?
- Have the particular design and student support requirements of web based learning courses been taken into account? If not, the e-learning starter guides on the LTSN website are a good resource (www.ltsn.ac.uk/genericcentre)

Incorporating web based learning into conventional programmes

Web based learning in an institution is often integrated with conventional, face to face teaching. This is normally done via an intranet, which is usually "password protected" and accessible only to registered users. Thus it is possible to protect the intellectual property of online material and to support confidential exchange of communication between students.

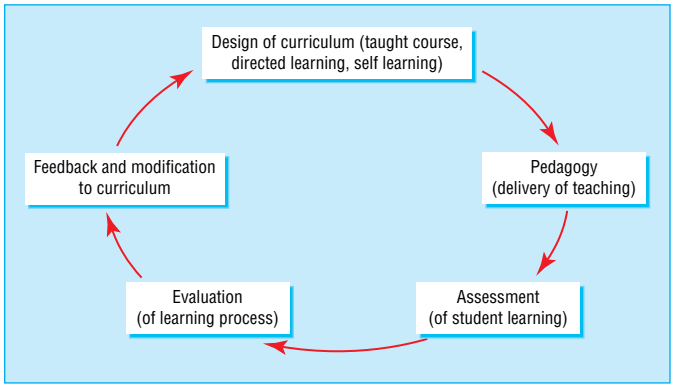
Medicine has many examples of online learning, in both the basic sciences and clinical teaching. As students are usually in large groups for basic science teaching, web based learning can

"Newer technologies such as computers and video conferencing are not necessarily better (or worse) for teaching or learning than older technologies … they are just different … The choice of technology should be driven by the needs of the learners and the context in which we are working, not by its novelty."

Bates AW. *Technology, open learning and distance education*. London: Routledge, 1995

Features of a typical web based course

- Course information, notice board, timetable
- Curriculum map
- Teaching materials such as slides, handouts, articles
- Communication via email and discussion boards
- Formative and summative assessments
- Student management tools (records, statistics, student tracking)
- Links to useful internal and external websites—for example, library, online databases, and journals



The learning cycle: useful to bear in mind when planning a web based course

Learning and teaching support network (www.ltsn.ac.uk)

- The learning and teaching support network (consisting of 24 subject centres) was set up to help staff in higher education to deliver programmes in their own subjects more effectively and to improve communication between teachers in different organisations and government agencies
- One subject centre covers medicine, dentistry, and veterinary science (LTSN 01)
- Although not strictly web based learning, LTSN 01 uses a combination of activities delivered via the web (such as an email discussion forum, copies of useful articles, research and funding programmes, and job vacancies) to support staff working in medical, dental, and veterinary education

be used to provide learning materials to complement conventional programmes and to enable self assessment—for example, access to anatomical sites and image banks for the teaching of pathology courses. Web based learning can be useful to support clinical teaching when learners are geographically dispersed—for example, to learn clinical skills through video demonstrations.

With web based learning, the material can be linked to libraries (for example, for ordering books or journals), online databases, and electronic journals. These functions are particularly useful for research and clinical activities

Assessment

With all types of learning, including web based learning, it is useful for students to receive constructive, timely, and relevant feedback on their progress. Online assessment is sometimes constrained by the medium in which it is operating. Computer marked assessments alone are not appropriate for marking or giving feedback on assignments such as essays or projects that require more than the mere reproduction of knowledge.

When planning online assessment it is important to determine what is to be assessed. If knowledge reproduction is being tested, objective questions (such as multiple choice or "true or false" questions) with instant or model answers can provide excellent feedback. Assessment of higher cognitive functions, such as analysis and synthesis, will require more complex tests. Automated marking may be difficult for such assessments, and the teacher is likely to have to do a substantial amount of work before he can add his or her comments to the student's record. Further guidance on how to design web based assessments for online courses can be found at www.ltsn.ac.uk and www.ltss.bris.ac.uk

Advantages and disadvantages of online assessment

Advantages
- Students can receive quick feedback on their performance
- Useful for self assessments—for example, multiple choice questions
- A convenient way for students to submit assessment from remote sites
- Computer marking is an efficient use of staff time

Disadvantages
- Most online assessment is limited to objective questions
- Security can be an issue
- Difficult to authenticate students' work
- Computer marked assessments tend to be knowledge based and measure surface learning

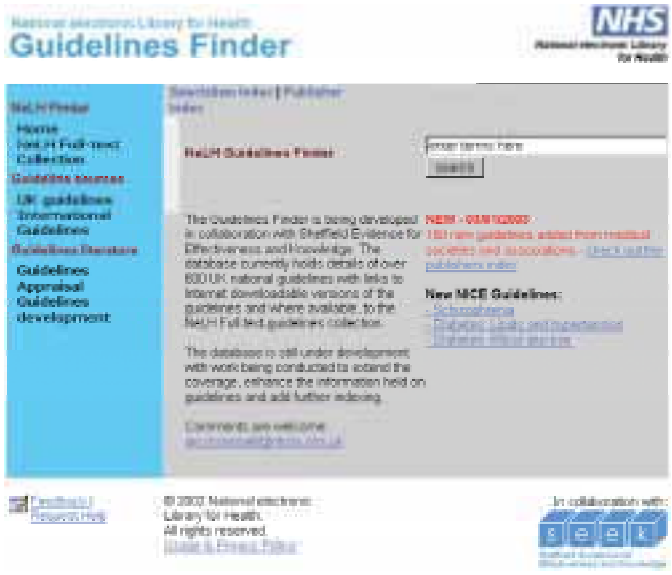
For and against web based learning

When designing web based programmes (as with any learning programme), the learners' needs and experience must be taken into account. Appropriate technology and reasonable computer skills are needed to get the best out of web based or online learning. Programmes and web pages can be designed to accommodate different technical specifications and versions of software. It is frustrating for learners, however, if they are trying to work on the internet with slow access or cannot download images and videos they need. On the other hand, web based programmes may, for example, encourage more independent and active learning and are often an efficient means of delivering course materials.

Advantages and disadvantages of web based learning

Advantages
- Ability to link resources in many different formats
- Can be an efficient way of delivering course materials
- Resources can be made available from any location and at any time
- Potential for widening access—for example, to part time, mature, or work based students
- Can encourage more independent and active learning
- Can provide a useful source of supplementary materials to conventional programmes

Disadvantages
- Access to appropriate computer equipment can be a problem for students
- Learners find it frustrating if they cannot access graphics, images, and video clips because of poor equipment
- The necessary infrastructure must be available and affordable
- Information can vary in quality and accuracy, so guidance and signposting is needed
- Students can feel isolated



Effective web teaching and learning

Course designers need to remember that younger students are more likely to be familiar with using the internet than older learners, who may feel less comfortable with a web based course. To get the best out of their learning experience, learners need basic computer skills, support, and guidance.

Teachers must design their courses to encourage effective web based learning rather than aimless "surfing." Programme design should therefore filter out poor information as well as signpost key information sources.

Many clinicians are beginning to use electronic patient records. This change means that doctors are becoming more adept at using computers and online resources to support their daily work and continuing professional development. Electronic media can facilitate access to evidence based

resources such as the Cochrane Library. These web based clinical support sites are excellent resources for postgraduate "on the job" learning.

Teachers should be encouraged, through training and support, to use the web and other information technologysystems in their teaching. They need examples and awareness of good practice, and standards should be set in relation to how teachers present information and manage the learning environment.

Conclusion

Web based learning offers huge opportunities for learning and access to a vast amount of knowledge and information. The role of teachers is to ensure that the learning environment provided takes account of learners' needs and ensures that they are effectively prepared and supported. Online learning has advantages, but web based learning should not always be viewed as the method of choice because barriers (such as inadequate equipment) can easily detract from student learning. The technology must therefore be applied appropriately and not used simply because it is available and new or because students and teachers have particular expectations of this means of course delivery.

Further reading

- Cook J. The role of virtual learning environments in UK medical education. *LTSN Bioscience Bulletin* 2002;5. http://bio.ltsn.ac.uk
- Forsyth I. *Teaching and learning materials and the internet.* 3rd ed. London: Kogan Page, 2001.
- Jolliffe A, Ritter J, Stevens D. *The online learning handbook: developing and using web based learning.* London: Kogan Page, 2001.
- World Federation for Medical Education. www.sund.ku.dk/WFME/WFME%20Guidelines/guidelines99.html (paper on using information technology in education, including web based learning)

14 Creating teaching materials

Richard Farrow

The nature and qualities of the teaching materials that you use can have a substantial effect on the educational experience of your students. Teaching materials can often distract learners rather than help them to learn. Common avoidable problems include overcrowded or illegible slides, irrelevant or badly prepared handouts, and incompatible multimedia equipment. It is important therefore to know how to create effective teaching materials.

Ground rules

Five basic principles apply to preparing teaching materials, irrespective of the type of material you choose: links, intelligibility, general style, highlighting, and targeting (LIGHT). You may sometimes decide to ignore one or more of these principles, but if you do, think carefully about what you are trying to achieve.

Links

Your teaching materials should have obvious and direct links to your talk, discussion, or presentation. Handouts are the main offenders in this category, and it is not unusual for handouts to have little in common with the talk. It is quite acceptable for the teaching materials to give some additional information, but this should not be excessive.

Intelligibility

The teaching material should be easy to understand and learn from. How this is achieved will depend on the medium used and the venue of the talk or presentation. Use simple language and avoid overlong sentences or statements. Diagrams can help to clarify a complex message. If you are using slides or overhead transparencies, the size of the print needs to be large enough to be read from the back of the auditorium. The font selected should be sans serif (for example, Arial).

Preparing overhead transparencies

Do
- Try to use typed rather than handwritten script
- Use a type size that is big enough to be read by the whole audience—for example, at least 20 points
- Make sure that the colour of your text works—for example, dark print on a pale background
- Limit each transparency to one idea or concept

Don't
- Use small print
- Use overhead transparencies packed with tables and figures
- Use light colours

General style

You should aim to use a consistent style throughout your teaching materials, particularly if you are giving a series of talks. Although it is tempting to use a variety of novel styles, consistency will allow learners to concentrate on the meaning and relevance of what you are trying to communicate.

Uniformity in the teaching materials will help learners to focus on content rather than style

Highlighting

Highlighted information helps to emphasise important issues or pivotal points in a developing argument. Methods of highlighting include changing the colour of text or underlining words or phrases. This also applies to videotapes and audiotapes, where changing your tone of voice can be used to emphasise key points.

It is easy to **overdo** highlighting by **emphasising** virtually every point that you make. This reduces the **usefulness** of the technique and **hides** the really pivotal shifts in a **morass** of *highlighted text.*

Targeting

It is important that both the type of educational event (for example, presentation, seminar, discussion) and the teaching materials that supplement it are targeted at what your students need to learn. Targeting therefore requires an awareness of what knowledge and skills your students already have. This can

Target your talk at learners' needs—don't just pull out the slides or overheads from a previous talk

be difficult to judge, but it is worth spending time finding out about your expected audience. It becomes easier if you are doing a series of talks with the same group as you can get feedback from the learners to help you plan more effectively.

Types of teaching materials

Black, green, or white boards
These are ideal for brainstorming sessions and small group work. If you are doing the writing, try not to talk at the same time as it is difficult for your learners to hear you if you have your back to them. Remember the LIGHT principles, and try to put concepts, not an essay, on the board. Make sure that everyone has finished copying information before you rub the board clean. Using different colours can add emphasis and highlight your important messages.

Lecture notes
Ensure that any handouts are produced to a high quality. Photocopies of handwritten notes (and frequently photocopied elderly pages) look scrappy and tend not to be valued. Give handouts to the learners at the beginning of the talk as copying down information is not a good use of their limited "face to face" time. Use headings and diagrams to make the handouts intelligible.

Overhead projector
The technical equipment for displaying overhead transparencies is widely available and reliable. It is a good backup resource, and for critical presentations it is comforting to know that, if all else fails, you have transparencies in your bag. Presentations using an overhead projector have the advantage that they allow you to face your audience while pointing out features on the transparency.

Correct preparation following the LIGHT principles is vital. Ensure that the transparencies will fit the projector—most will display A4 size, but some are smaller, so check in advance. The absolute minimum height for text on transparencies is 5 mm, although using larger text and fewer words usually produces a more effective educational tool. A good rule of thumb is to use a type size of at least 20 points. Several simple transparencies are usually better than one complicated one.

It is fairly straightforward to design your transparency on a computer then print it using a colour printer. Avoid using yellow, orange, and red, as these colours are difficult to see. Insead, use dark text on a light background. You can write and draw directly on to the transparencies with felt tipped pens. Use permanent markers to avoid smudging, and place a sheet of ruled paper underneath so that the writing is evenly spaced. You can also use a photocopier to copy print on to a transparency, but remember that you may need to enlarge it to make the text readable.

If you are likely to use a transparency again it is worth storing it carefully in dust free covers. One commonly used method is to store transparencies in clear plastic sleeves that can be filed in a ring binder. When showing transparencies, do not overuse the technique of covering the transparency and revealing a little at a time—many learners find this irritating.

35 mm slides
The need for 35 mm slides has decreased substantially with the advent of computer programs such as Microsoft's PowerPoint. However, multimedia projectors and computers are expensive and not available in all locations, whereas most educational institutions have a slide projector. Making your own slides can be difficult, so get help from the local illustration department or

Types and uses of teaching materials

Boards, flip charts—Small groups, problem based learning tutorials, workshops
Lecture notes—Small and large groups; help to improve interactivity
Overhead projector—Small and large groups, workshops, and interactive sessions
35 mm slides and PowerPoint—Generally large groups and lecture formats
Videos—Good for clinical teaching in larger groups (use film of patients); also for teaching communication skills and practical skills (students can keep films for self appraisal)
Life and plastic models—Anatomy teaching in small groups or for self directed learning
Computer assisted learning packages—Small groups with a tutor; large groups in computer laboratories; self directed learning
Skills centres and simulators—Small groups learning clinical skills

Leave spaces in the handout for your learners to record the results of interactive parts of your talk—this ensures that the handout the learners take away has more value than the one they were given. Also, leave spaces for exercises to be completed later, thus linking self directed learning with face to face learning

Paper copies of transparencies and slides can make useful handouts—your learners can then add clarifying statements or diagrams to their own copy of the presentation

Number your slides so that if a projectionist is loading them or the carousel is dropped they can be quickly reordered

a commercial company. Ensure that the text is large enough to see when projected and that the slides are marked so that they are loaded in the projector correctly. Dual projection is rarely done well and rarely necessary unless you are using visual images (for example, *x* ray films, clinical photographs) with accompanying text. If you use dual projection make sure that each of the slides is labelled for the correct projector.

Computer generated slides

The ability to make computer generated slides (for example, PowerPoint) has transformed the way that many people create teaching materials and has greatly reduced the use of 35 mm slides. Try not to get seduced by the technology, however, and remember that it is just another educational tool. Having tried all of the colours and slide layouts available, many experienced lecturers now prefer simple formats that are easy to read and in which the medium does not get in the way of the message.

However, the computer package has many useful tools—diagrams and "clip art" can help to conceptualise difficult problems. Video clips can be inserted into a presentation, but be certain that they are there to illustrate a point and not simply to show off your own technological skills. Use advanced formats for PowerPoint presentations only if you are well practised and comfortable with the medium.

Ensure that the computer you are planning to use is compatible with the multimedia projector. Similarly, if you have stored your presentation on a CD or floppy disk (or any one of the other portable storage formats), make sure that this is supported at the venue. The latest version of the presentation software can give you access to many features that may not work on the computer provided at the teaching venue, so a wise precaution is to save your presentation as an older version of the software.

Further reading

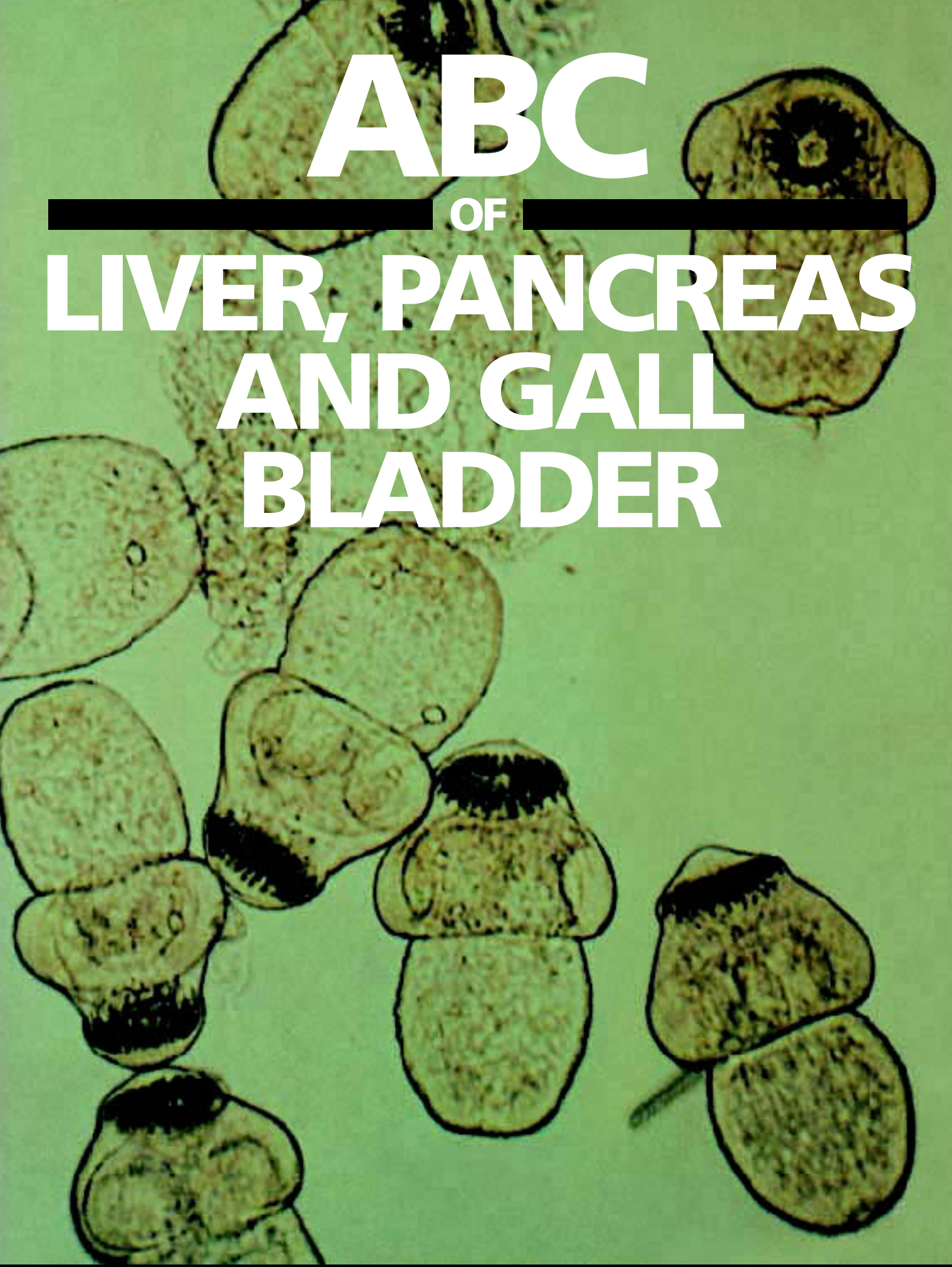
- Cannon R, Newble D. *A handbook for teachers in universities and colleges.* 4th ed. London: Kogan Page, 1999.
- Newble DI, Cannon R. *A handbook for medical teachers.* 4th ed. Dordrecht, Netherlands: Kluwer Academic, 2001.
- Kemp JE, Dayton DK. *Planning and producing instructional media.* 5th ed. New York: Harper and Row, 1985.
- Hartley J. *Designing instructional text.* 3rd ed. London: Kogan Page, 1994.

Ground rules for slide preparation (35 mm or PowerPoint)

- Use a clear font that is easily readable
- Use a type size of 20 points or greater
- Use a light text on a dark background for slides (in contrast with OHP transparencies)
- Use short sentences and small tables
- Restrict the overall number of words on each slide to about 40 or fewer
- Avoid patterned backgrounds—they are extremely distracting
- Limit the number of colours on your slides to a maximum of three
- Use highlighting to emphasise items in lists
- Use animation and sound effects sparingly

Index

ABC OF LIVER, PANCREAS AND GALL BLADDER



BMJ Books

ABC OF
LIVER, PANCREAS AND GALL BLADDER

This Page Intentionally Left Blank

ABC OF LIVER, PANCREAS AND GALL BLADDER

Edited by

I J BECKINGHAM

Consultant Hepatobiliary and Laparoscopic Surgeon, Queen's Medical Centre, Nottingham

BMJ Books

BMJ Books is an imprint of the BMJ Publishing Group

First published in 2001
by BMJ Books, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 7279 1531 2

Cover design by Marritt Associates, Harrow, Middlesex
Typeset by FiSH Books and BMJ Electronic Production

Printed and bound in Spain by GraphyCems, Navarra

Contents

This Page Intentionally Left Blank

Contributors

I J Beckingham
Consultant Hepatobiliary and Laparoscopic Surgeon, Queen's Medical Centre, Nottingham

P C Bornman
Professor of Surgery, University of Cape Town, South Africa

J E J Krige
Associate Professor of Surgery, Groote Schuur Hospital, Cape Town, South Africa

J P A Lodge
Consultant Hepatobiliary and Transplant Surgeon, St James Hospital, Leeds

K R Prasad
Senior Transplant Fellow, St James Hospital, Leeds

S D Ryder
Consultant Hepatologist, Queen's Medical Centre, Nottingham

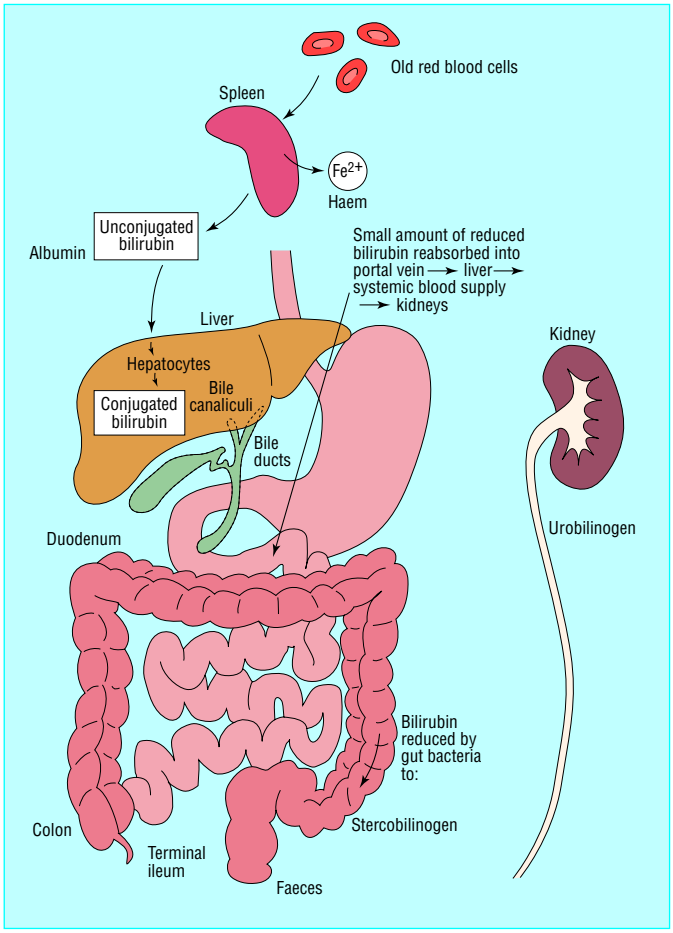
This Page Intentionally Left Blank

Preface

Diseases of the Liver, Pancreas and biliary system affect a substantial proportion of the worlds population and involve doctors and health care workers across many disciplines. Many of these diseases produce great misery and distress and are economically important requiring much time off work. The aim of this series was to provide an overview of these diseases and enable the busy clinician to keep abreast of advances in diagnosis and management of not only the common but also the rarer, but none the less important, conditions.

This Page Intentionally Left Blank

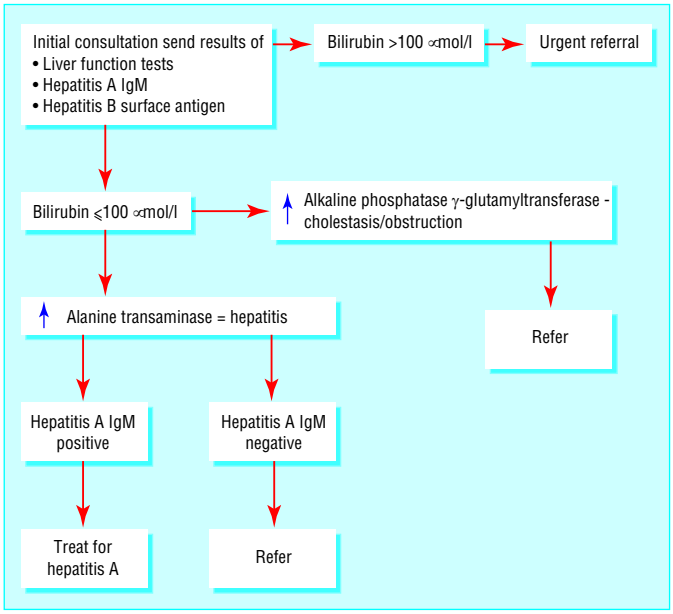
J


Old red blood cells
Spleen
Fe^{2+}
Haem
Unconjugated bilirubin
Albumin
Small amount of reduced bilirubin reabsorbed into portal vein → liver → systemic blood supply → kidneys
Liver
Kidney
Hepatocytes
Bile canaliculi
Conjugated bilirubin
Bile ducts
Duodenum
Urobilinogen
Bilirubin reduced by gut bacteria to:
Stercobilinogen
Colon
Terminal ileum
Faeces

“ ”

γ

′ γ


Initial consultation send results of
• Liver function tests
• Hepatitis A IgM
• Hepatitis B surface antigen
Bilirubin >100 ∝mol/l
Urgent referral
Bilirubin ⩽100 ∝mol/l
Alkaline phosphatase γ-glutamyltransferase - cholestasis/obstruction
Refer
Alanine transaminase = hepatitis
Hepatitis A IgM positive
Hepatitis A IgM negative
Treat for hepatitis A
Refer

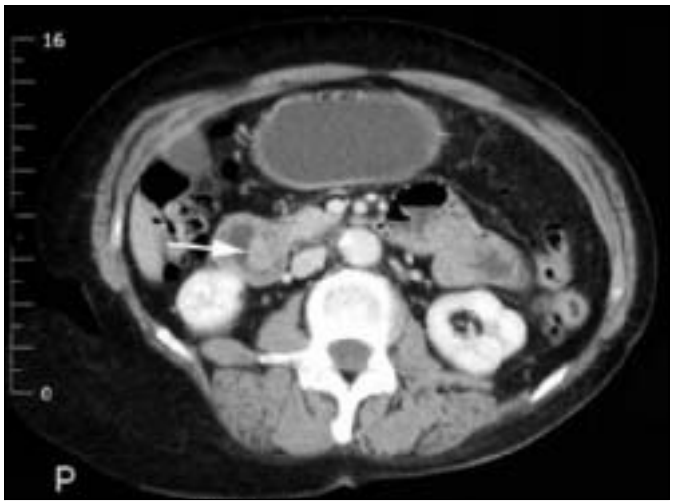
-
-
-

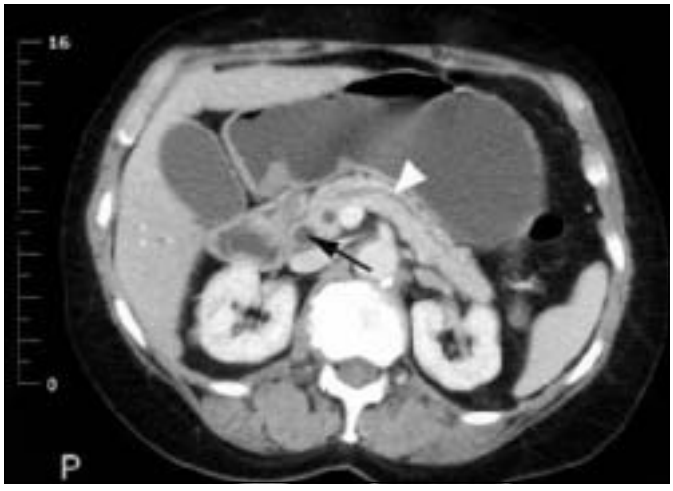
-
-

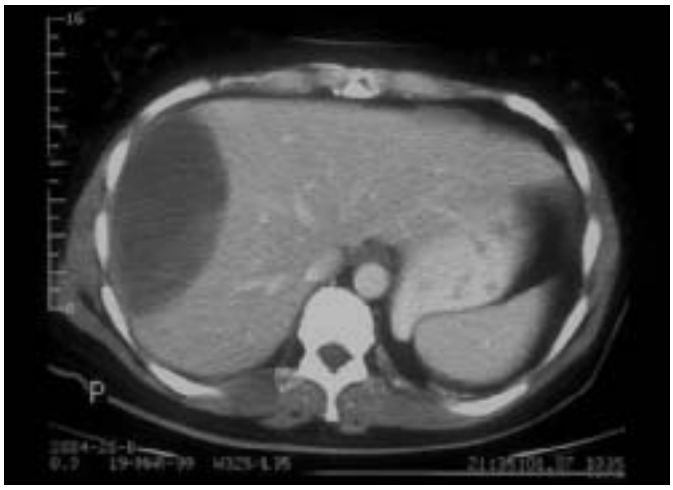
-
-

-
-

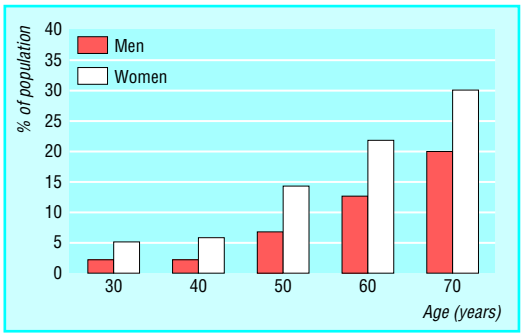
α







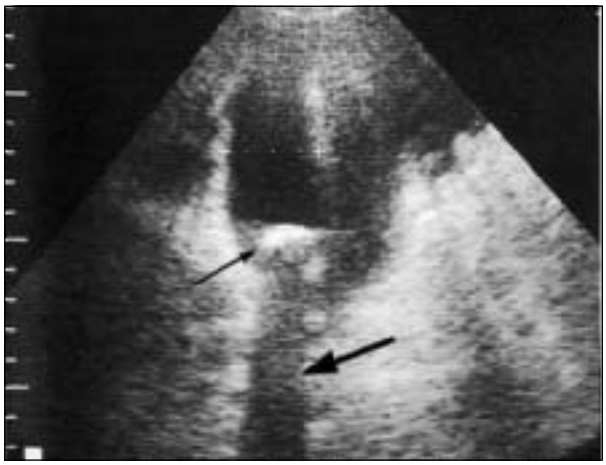
-
-
  γ
-
-
-



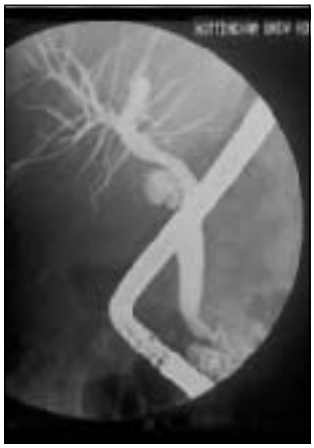
Escherichia coli

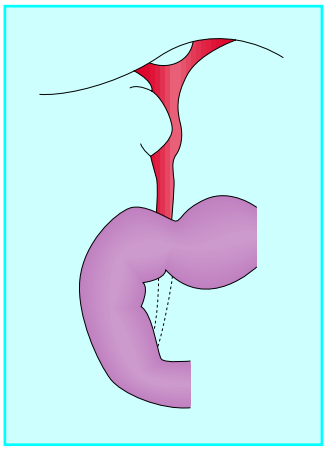
Klebsiella spp,

Ascaris lumbricoides *Opisthorchis senensis*



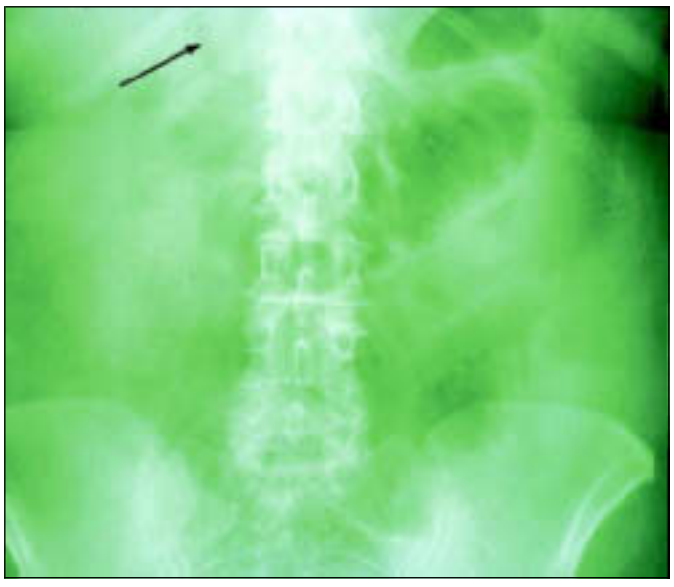
J



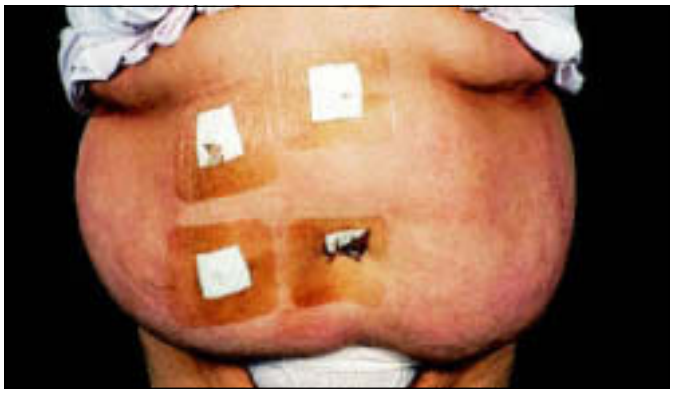


-
-
-

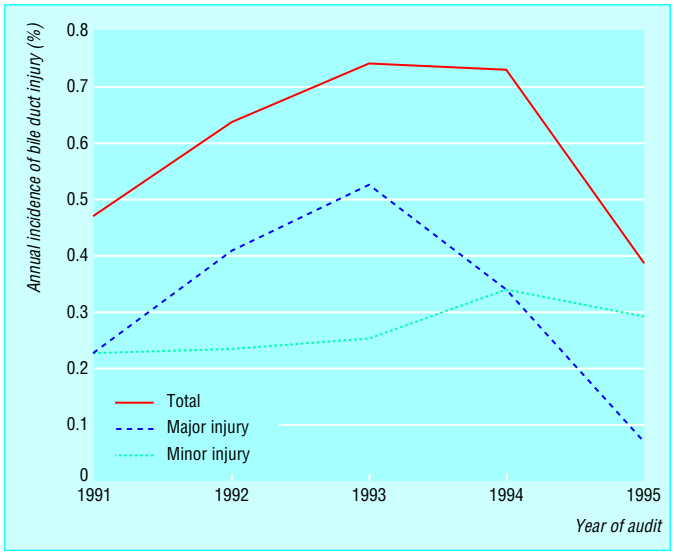
“ ”



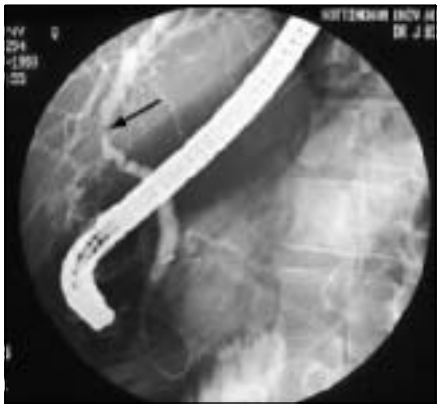
—

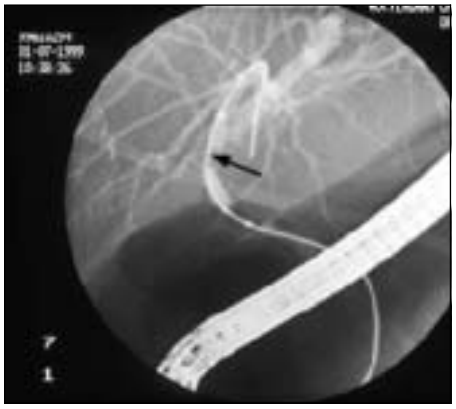


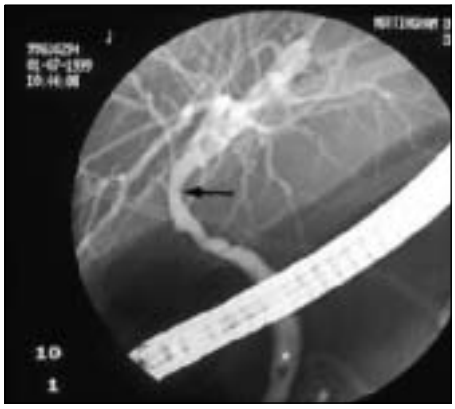
-
-
-
- —

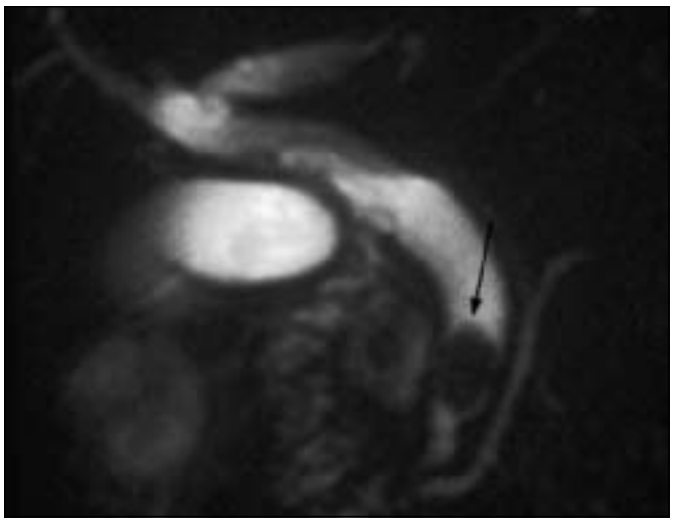

Annual incidence of bile duct injury (%)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Total
Major injury
Minor injury
1991
1992
1993
1994
1995
Year of audit

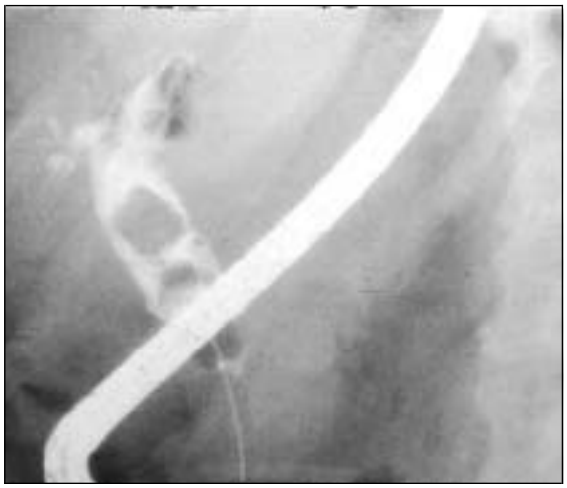
Br J Surg











Surgery of the liver and biliary tract

Am J Surg

Surg Endosc

γ

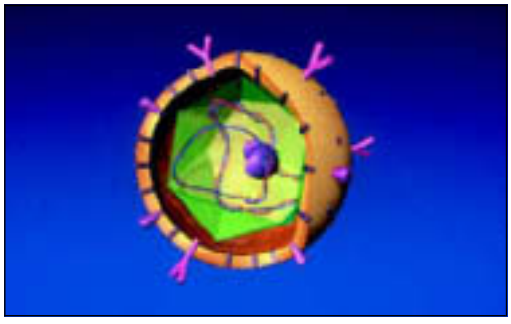
“ ”

γ

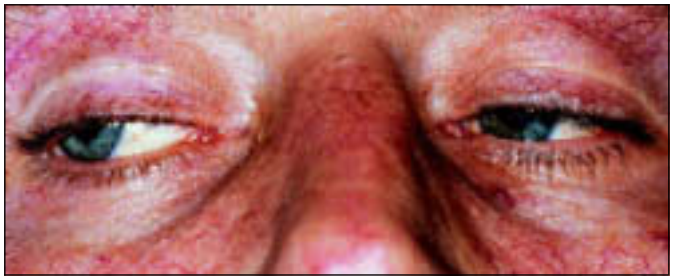
-
-
-
-
-
-
-

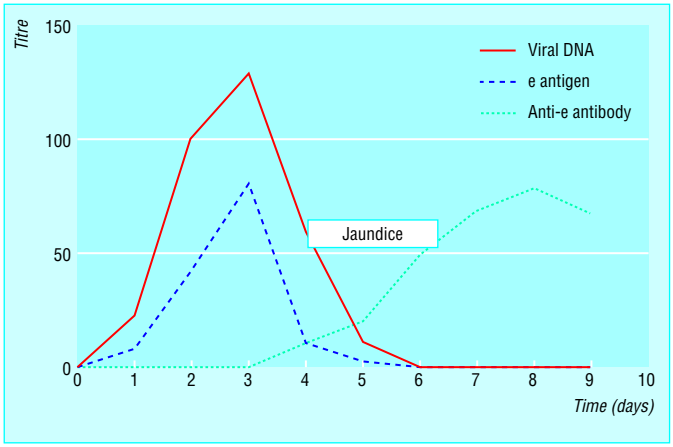
-
-
-

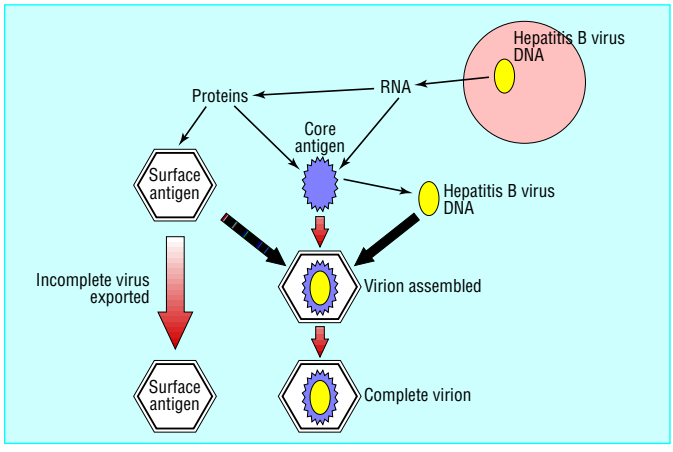
×

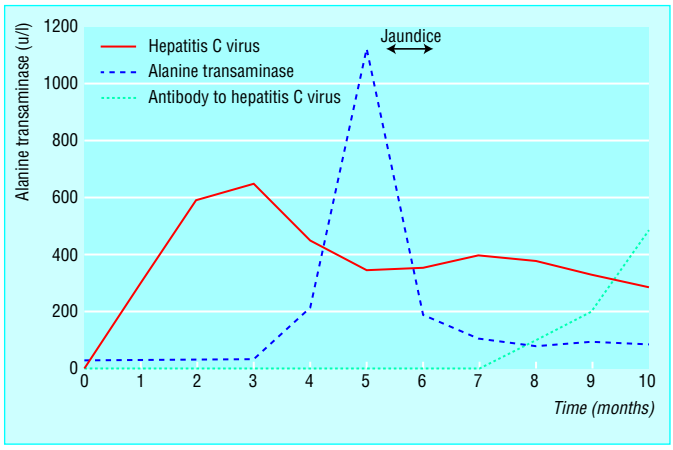


—

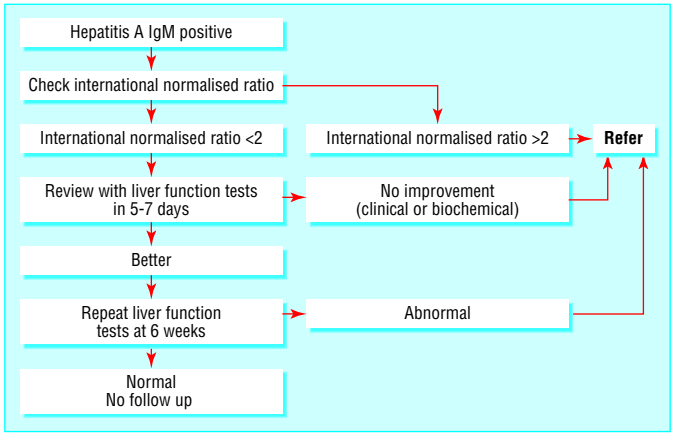



Titre
150
100
50
0
0
1
2
3
4
5
6
7
8
9
10
Time (days)
Viral DNA
e antigen
Anti-e antibody
Jaundice

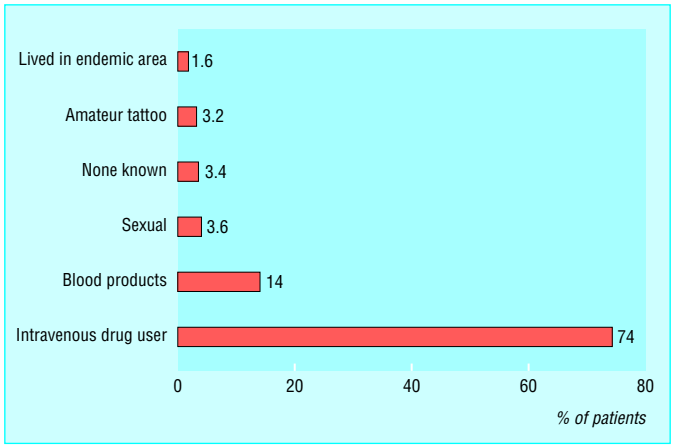

Hepatitis B virus DNA
RNA
Proteins
Core antigen
Surface antigen
Hepatitis B virus DNA
Incomplete virus exported
Virion assembled
Surface antigen
Complete virion

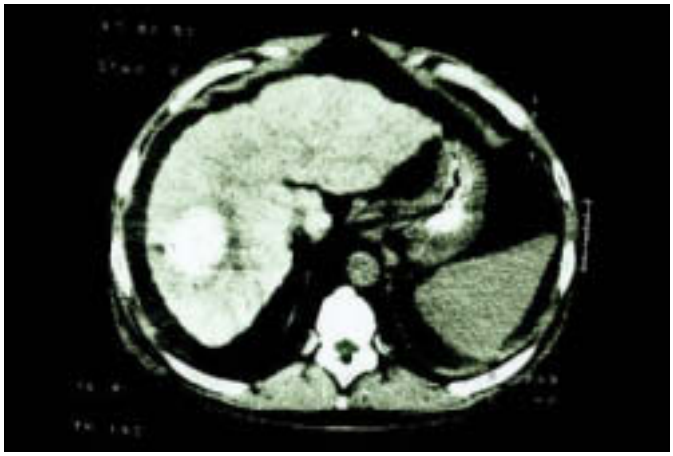

Alanine transaminase (u/l)
1200
1000
800
600
400
200
0
Hepatitis C virus
Alanine transaminase
Antibody to hepatitis C virus
Jaundice
0 1 2 3 4 5 6 7 8 9 10
Time (months)

-
-
-
-
-
-

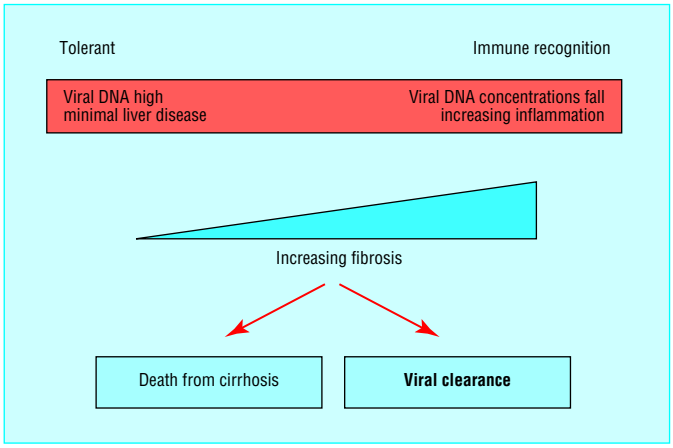

Hepatitis A IgM positive
Check international normalised ratio
International normalised ratio <2
International normalised ratio >2
Refer
Review with liver function tests in 5-7 days
No improvement (clinical or biochemical)
Better
Repeat liver function tests at 6 weeks
Abnormal
Normal
No follow up

-
-
-
-
-

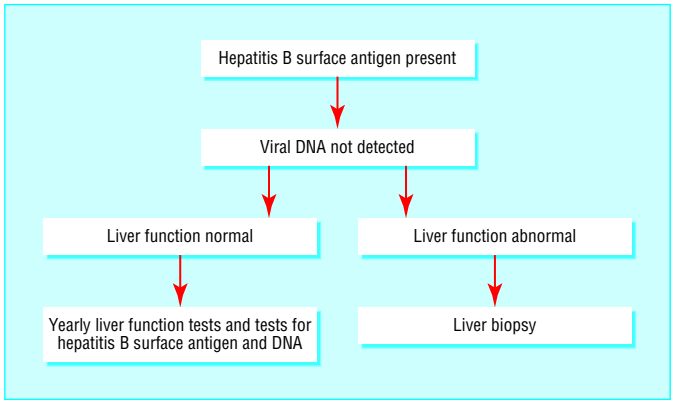

Lived in endemic area
1.6
Amateur tattoo
3.2
None known
3.4
Sexual
3.6
Blood products
14
Intravenous drug user
74
0
20
40
60
80
% of patients

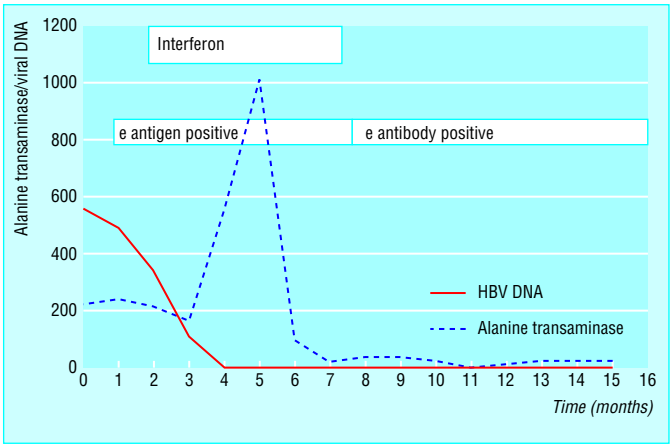


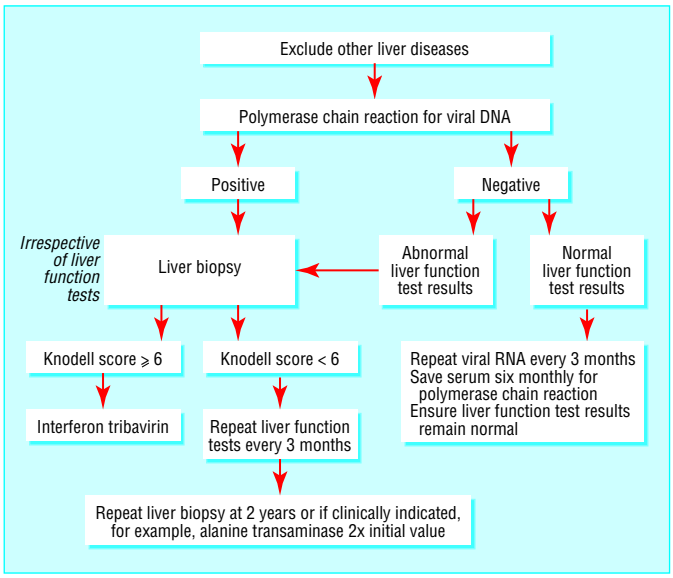
γ

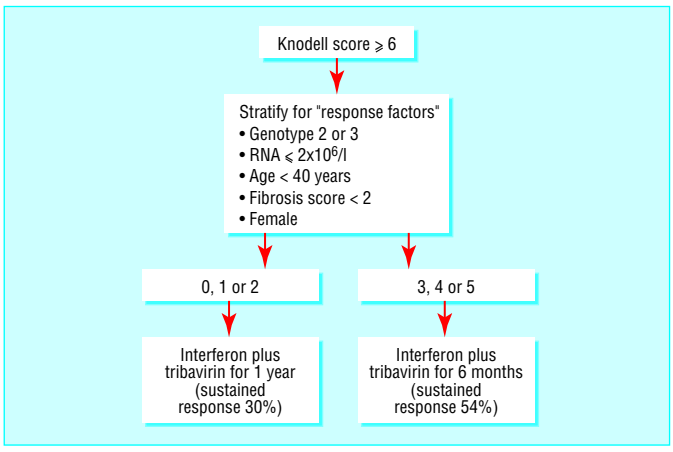

Tolerant
Immune recognition
Viral DNA high
minimal liver disease
Viral DNA concentrations fall
increasing inflammation
Increasing fibrosis
Death from cirrhosis
Viral clearance

“ ”


Hepatitis B surface antigen present
Viral DNA not detected
Liver function normal
Liver function abnormal
Yearly liver function tests and tests for hepatitis B surface antigen and DNA
Liver biopsy


Alanine transaminase/viral DNA
1200
1000
800
600
400
200
0
Interferon
e antigen positive
e antibody positive
HBV DNA
Alanine transaminase
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Time (months)

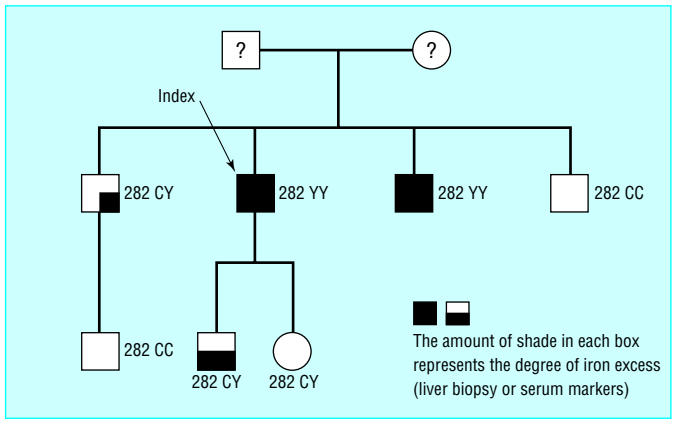




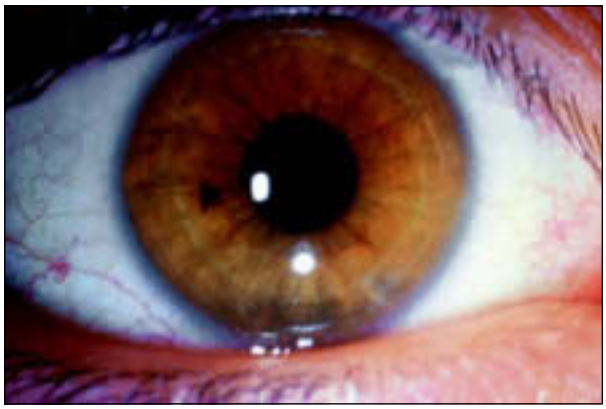
Prog Med Virol

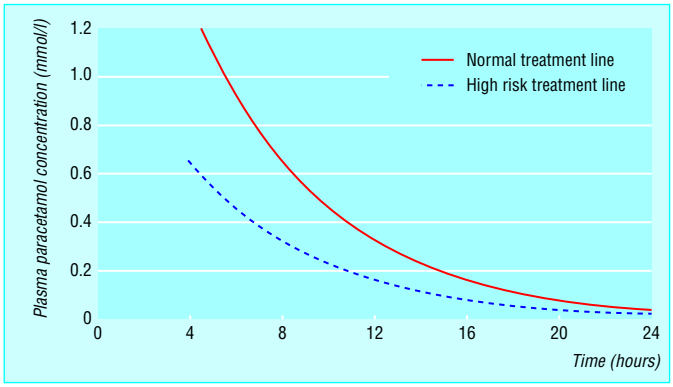
N Engl J Med

Hepatology


?
?
Index
282 CY
282 YY
282 YY
282 CC
282 CC
282 CY
282 CY
The amount of shade in each box represents the degree of iron excess (liver biopsy or serum markers)

-
-
-
-
-
-

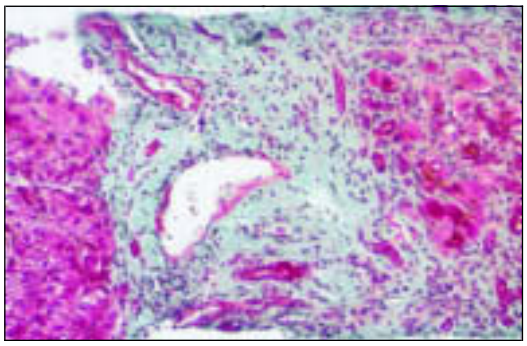




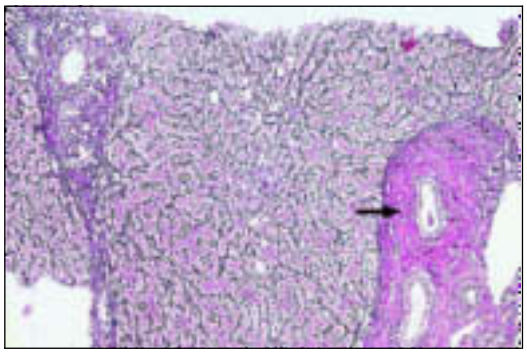
British National Formulary

-
-
-
-
-
-
-

-
-
-
-

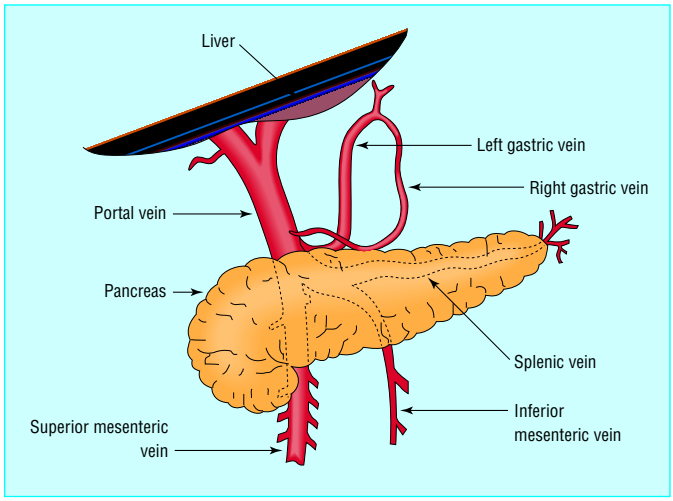


γ



“ ”

-
-
-
-
-



Prehepatic (portal vein obstruction)

-
-
-
-

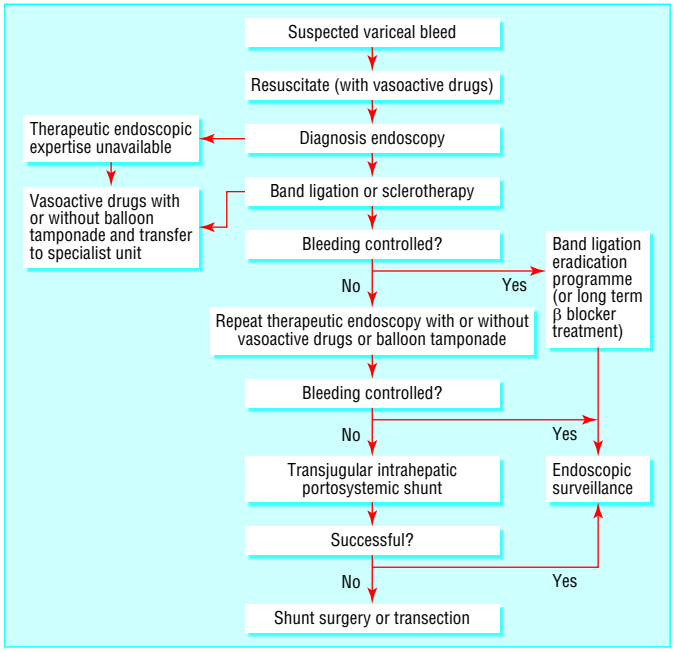
Hepatic

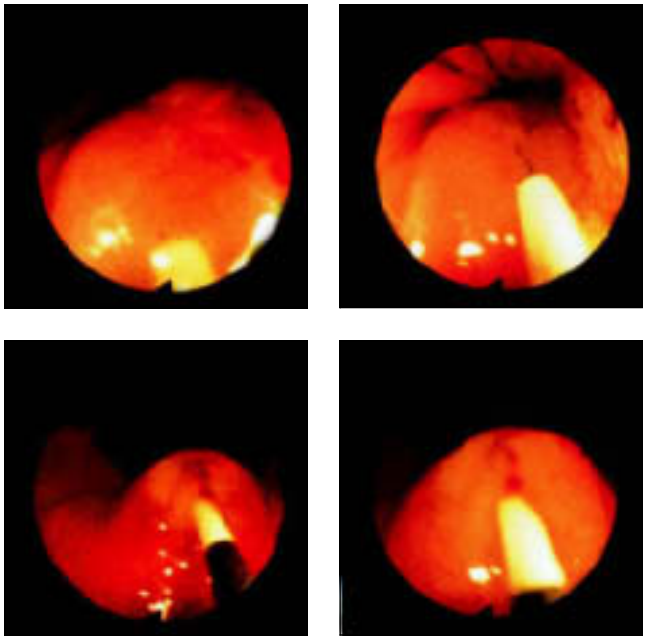
-
-
-
-
-

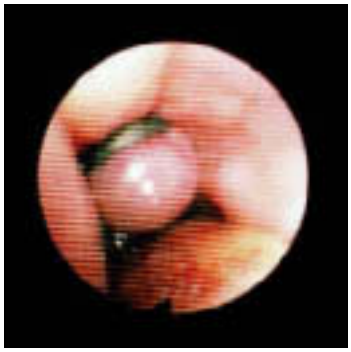
Posthepatic

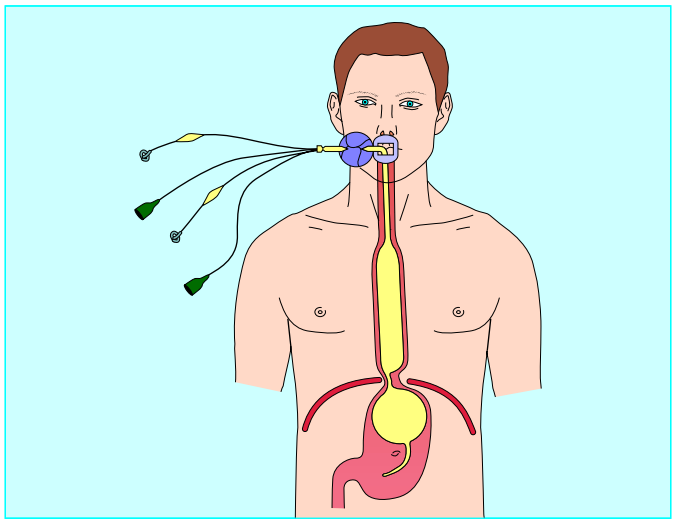
-
-

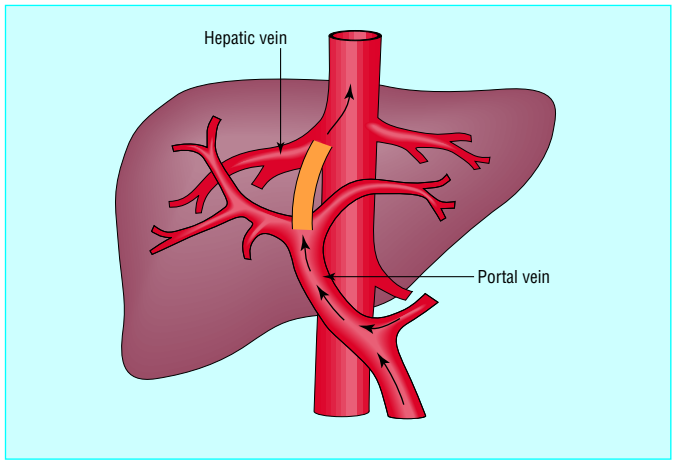
-
-


Suspected variceal bleed
Resuscitate (with vasoactive drugs)
Therapeutic endoscopic expertise unavailable
Diagnosis endoscopy
Vasoactive drugs with or without balloon tamponade and transfer to specialist unit
Band ligation or sclerotherapy
Bleeding controlled?
Band ligation eradication programme (or long term β blocker treatment)
No
Yes
Repeat therapeutic endoscopy with or without vasoactive drugs or balloon tamponade
Bleeding controlled?
No
Yes
Transjugular intrahepatic portosystemic shunt
Endoscopic surveillance
Successful?
No
Yes
Shunt surgery or transection

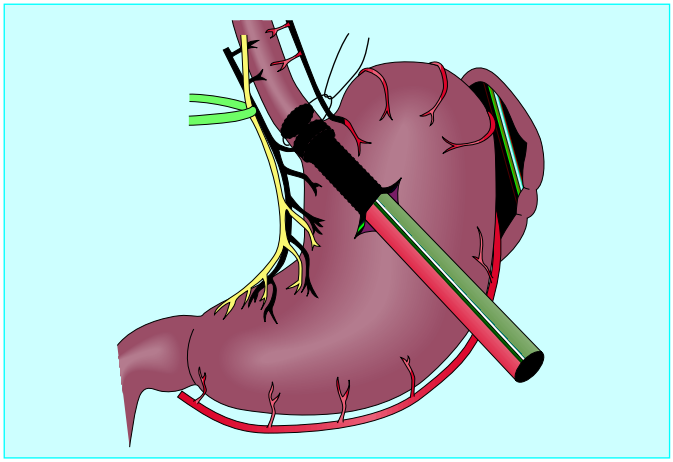







Hepatic vein
Portal vein

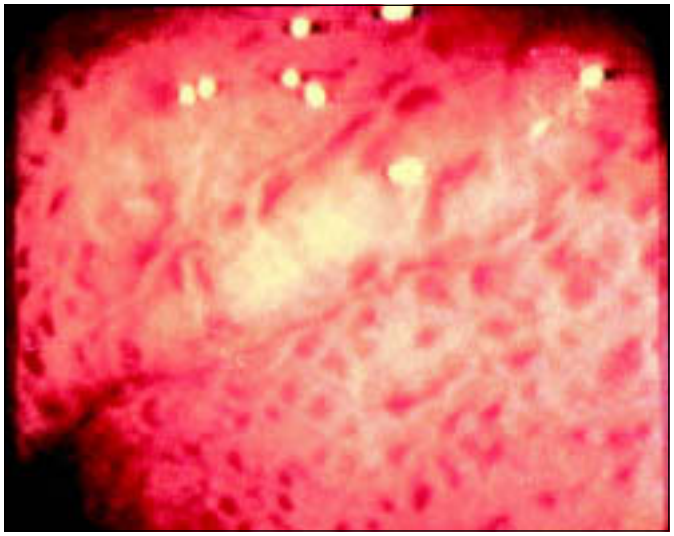
-
-
-
-

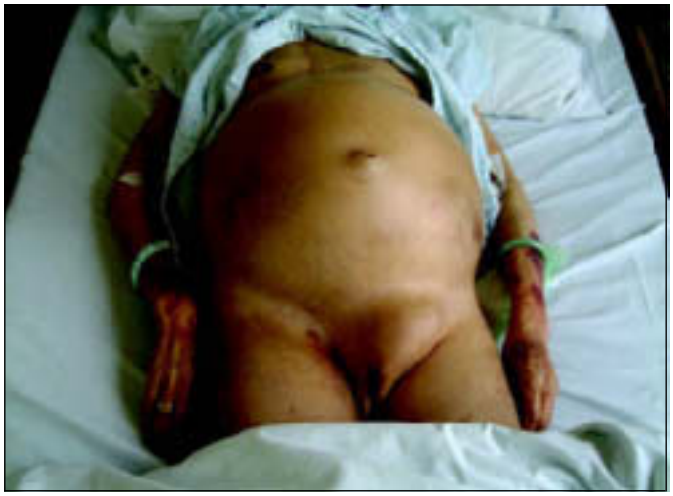


Surgery of the liver and biliary tract

Diseases of the liver and biliary system

N Engl J Med





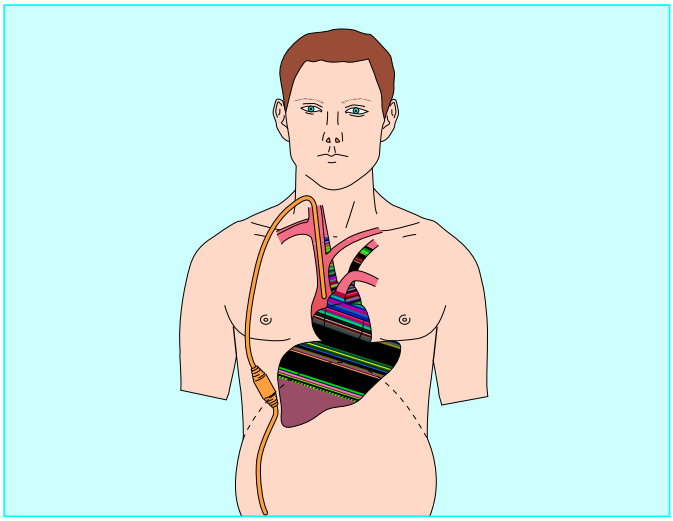
Tests to consider ordering

High gradient (≥ 11 g/l)

Low gradient (<11g/l)

—

“ ”



-
-
-

-
-

-
-

-
-
-

-

-
-
-

-
-
-
-

-
-
-
-

-
-
-
-

-
-
-

-
-

-

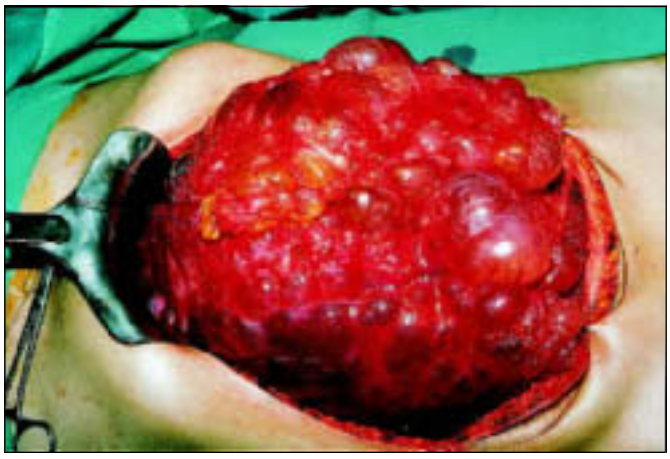
Surgery of the liver and biliary tract

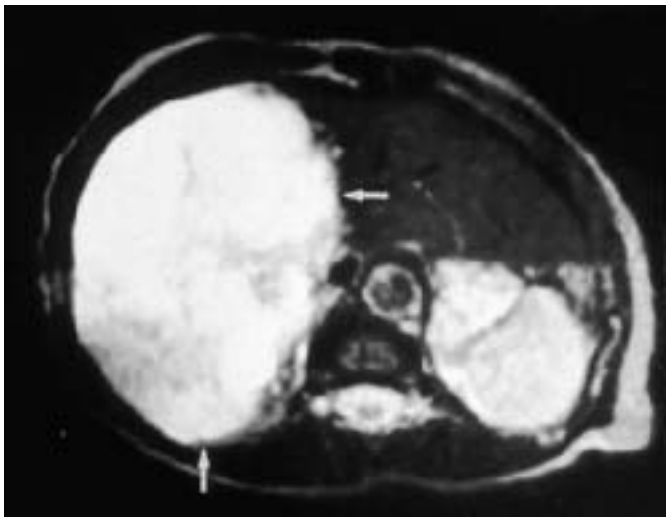
-

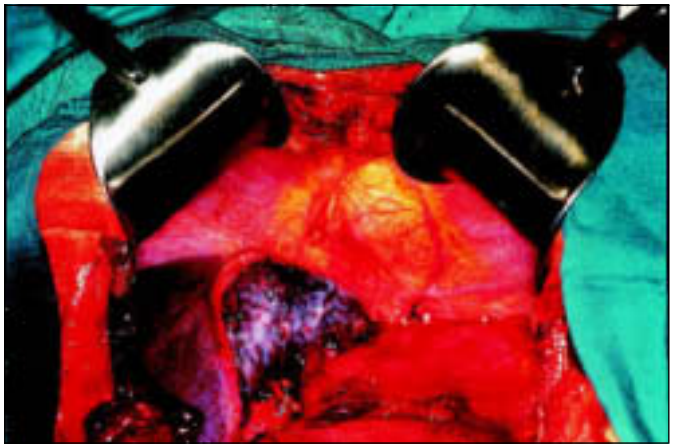
-
-
-

×

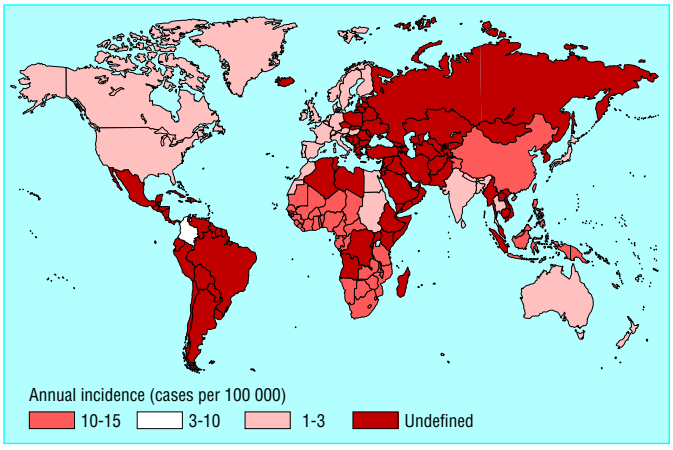
–



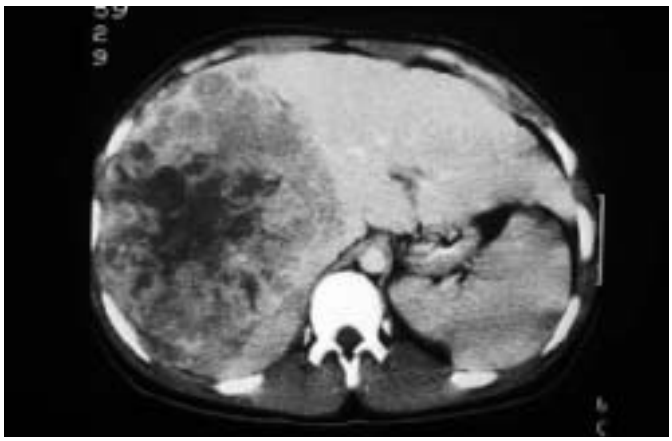




–

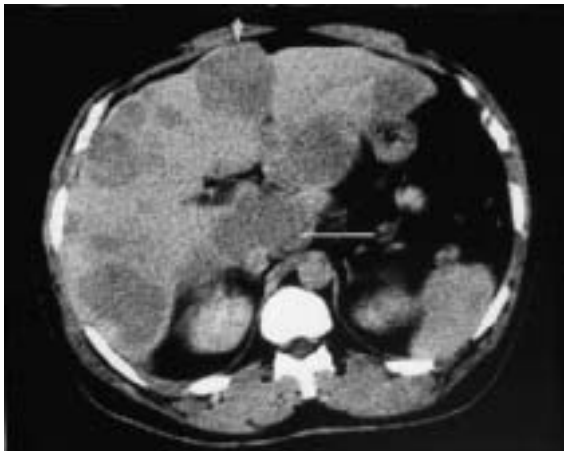


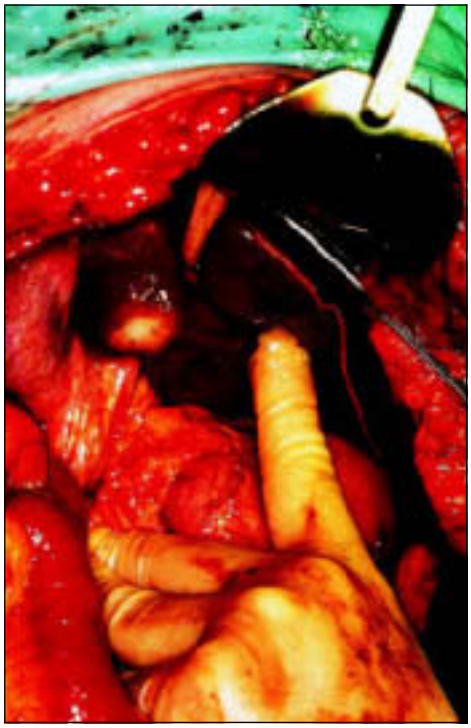
Aspergillus flavus,

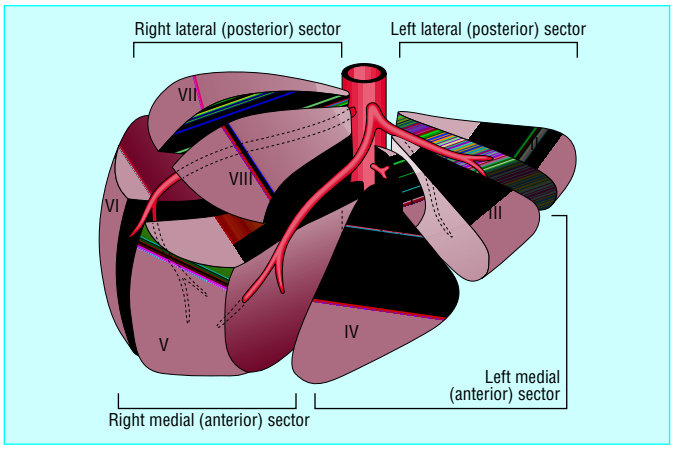


α

α





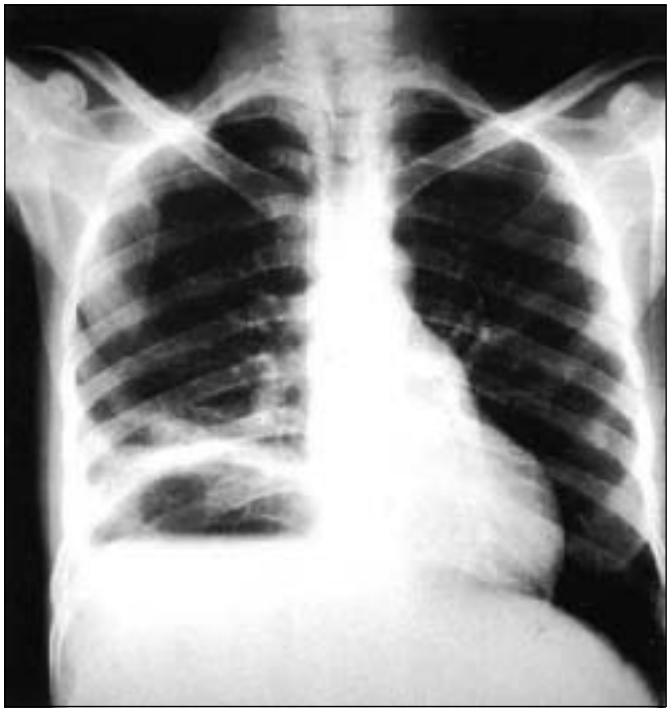

Right lateral (posterior) sector
Left lateral (posterior) sector
VII
VIII
VI
III
IV
V
Left medial
(anterior) sector
Right medial (anterior) sector

-
-
-
-
-

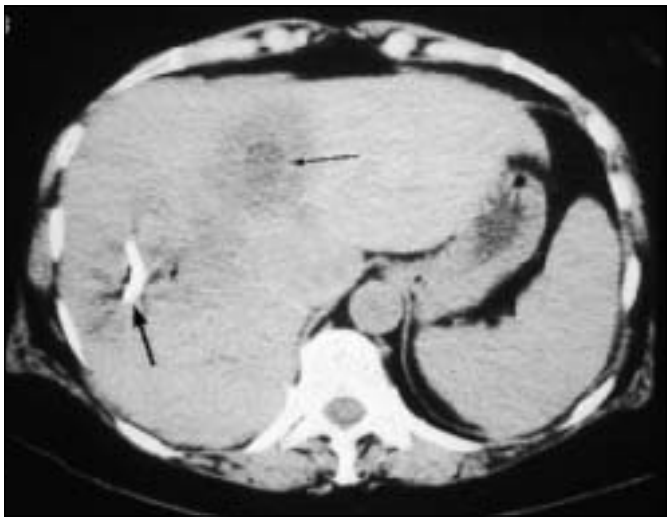
- *Surgery of the liver and biliary tract.*
- *Modern operative techniques in liver surgery.*
- *Br J Surg*

Escherichia coli, Klebsiella pneumoniae,

Streptococcus milleri



E coli, K pneumoniae,



—

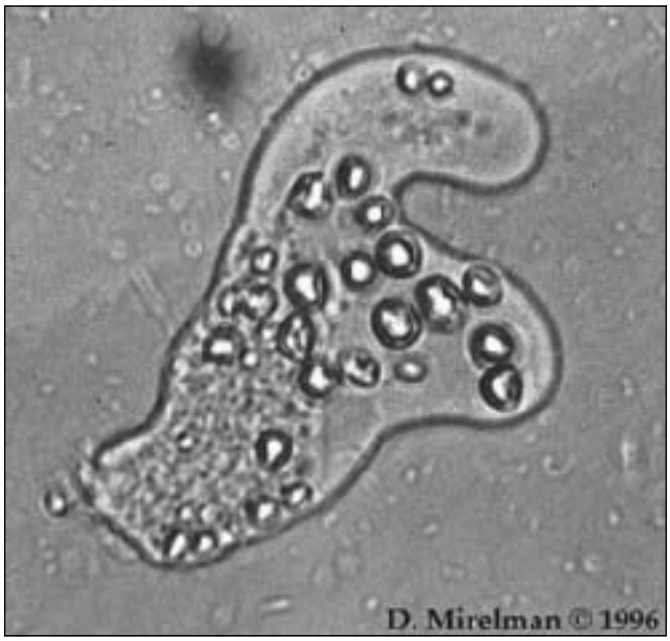
—

⩾

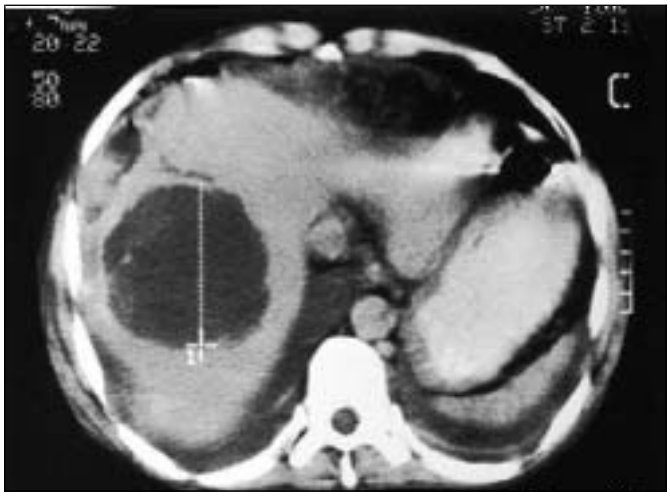
Entamoeba histolytica

“

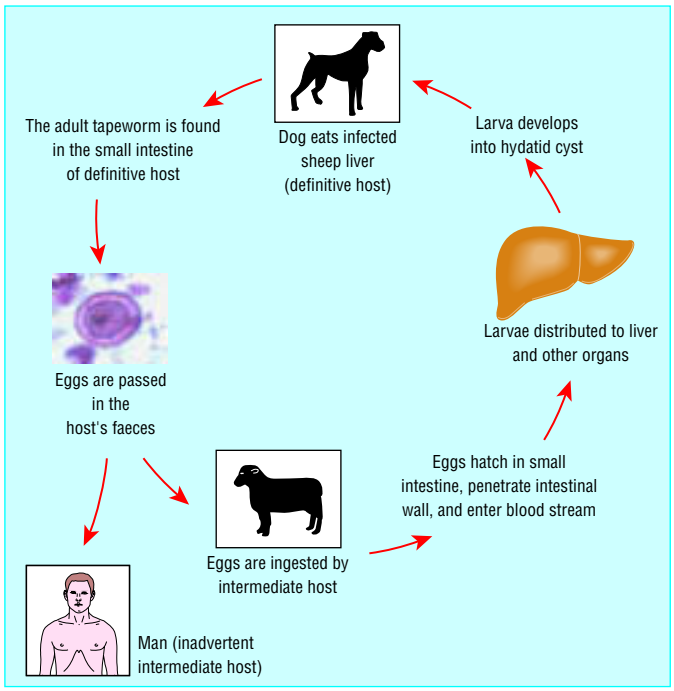
”



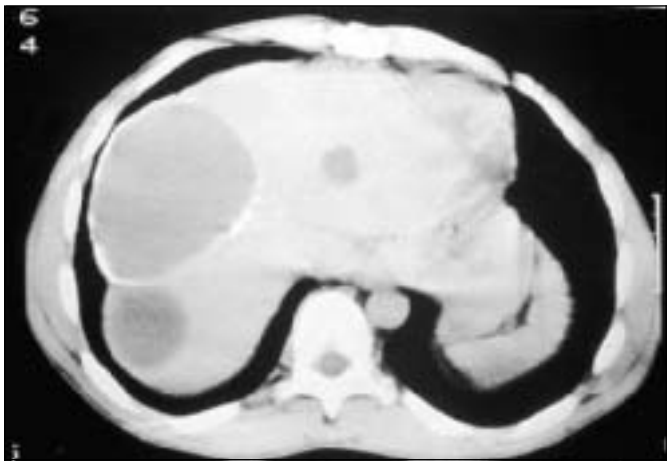
°



Echinococcus granulosus

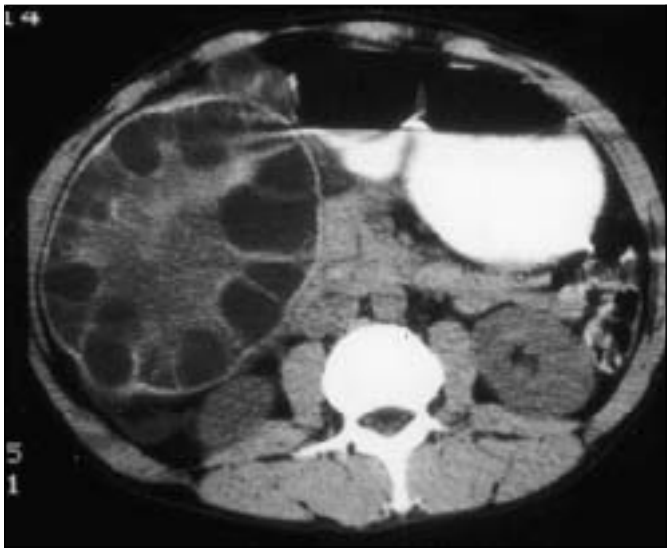


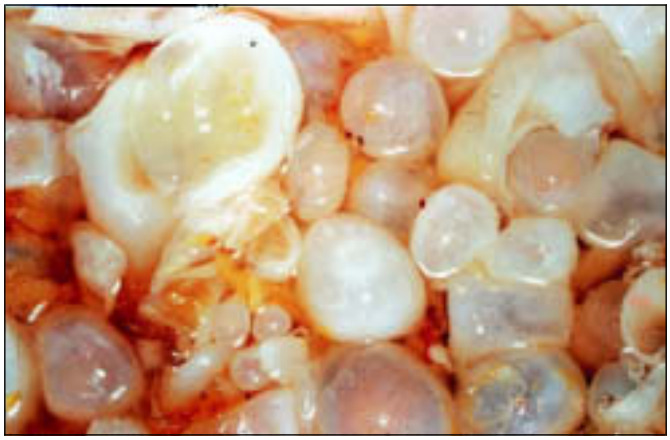
Echinococcus granulosus



γ

“ ”





Diagnosis and management of liver disease

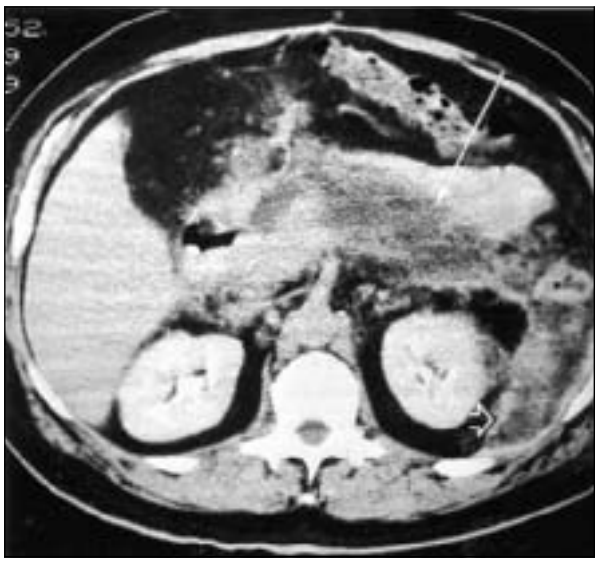
Diagnosis and management of liver disease

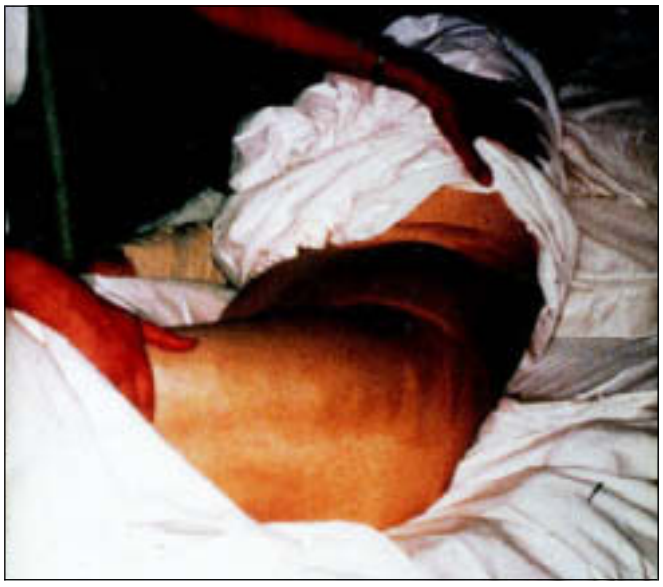
Current surgical therapy

-
-
-
-
-

- }
-
-
-
-

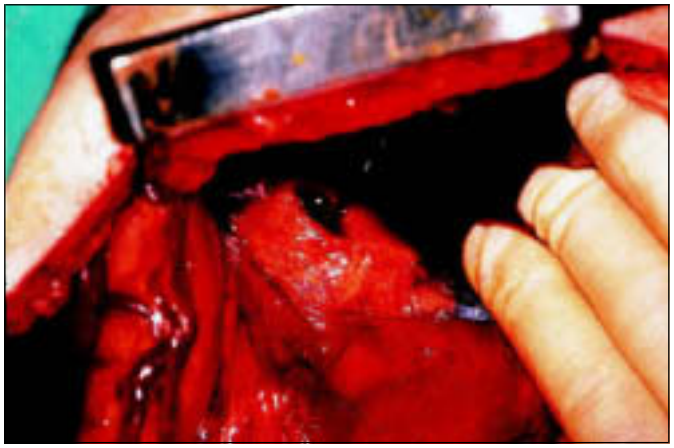
Ascaris lumbricoides





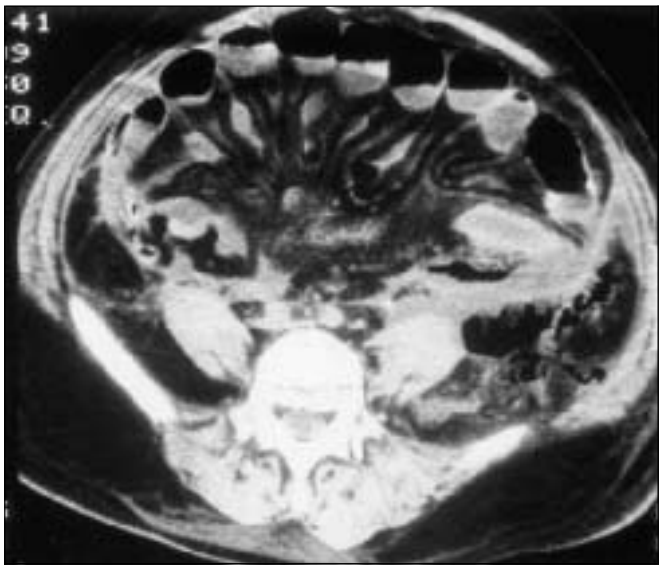
-
-
-
-
-
-

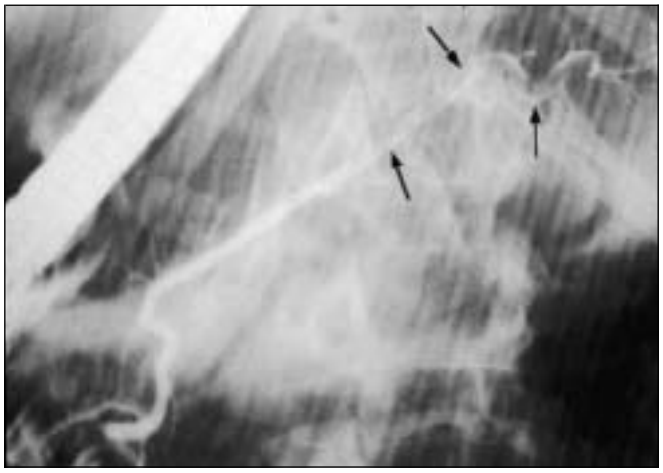
-
-
-



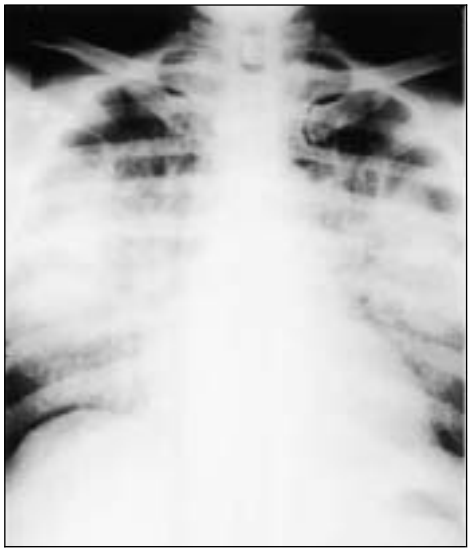
-
- ×
-
-
-
-
-
-

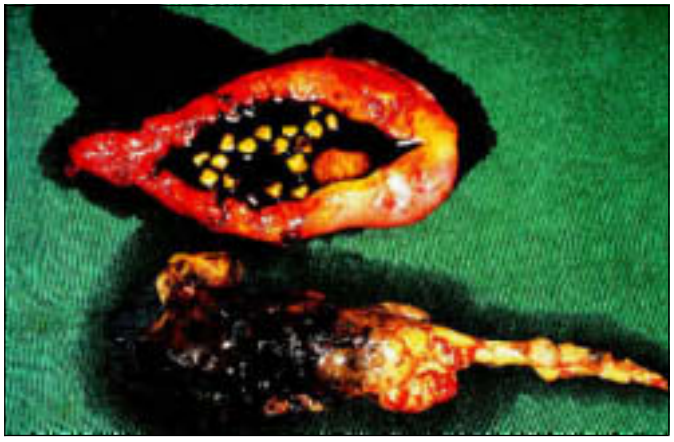
≥


41
9
0
0



Ascaris lumbricoides





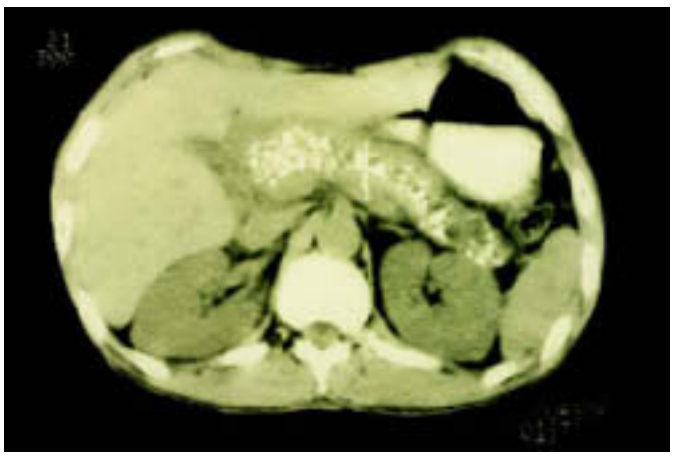
“ ”

Gut

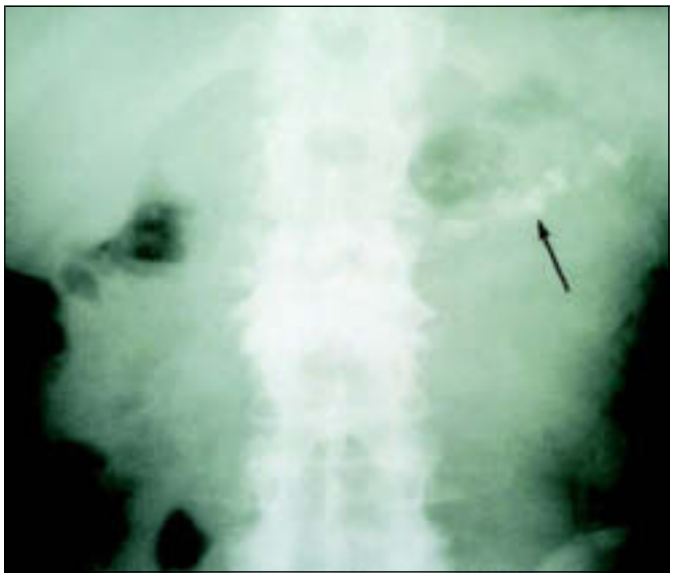
Eur J Gastroenterol Hepatol

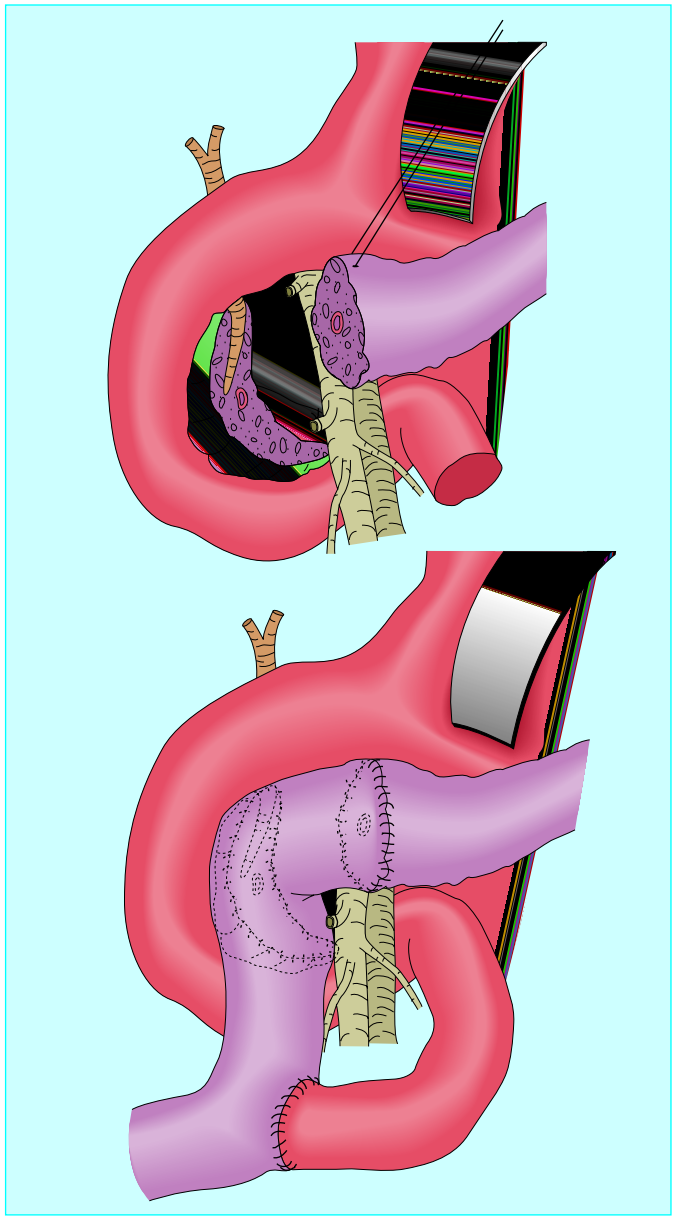
Surgery of the pancreas.

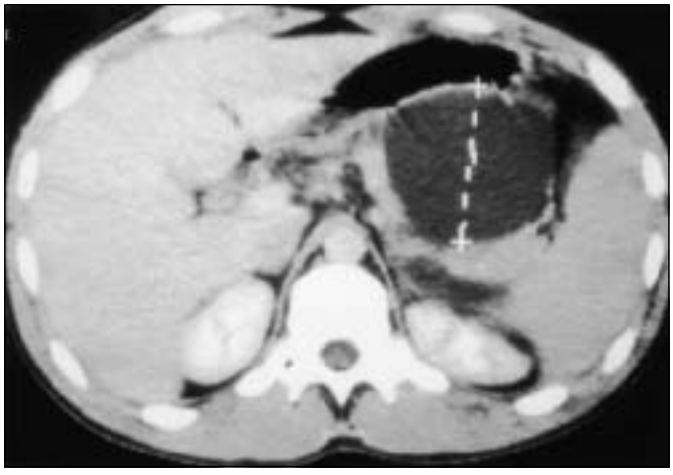
-
-
-
-
- “ ”
-

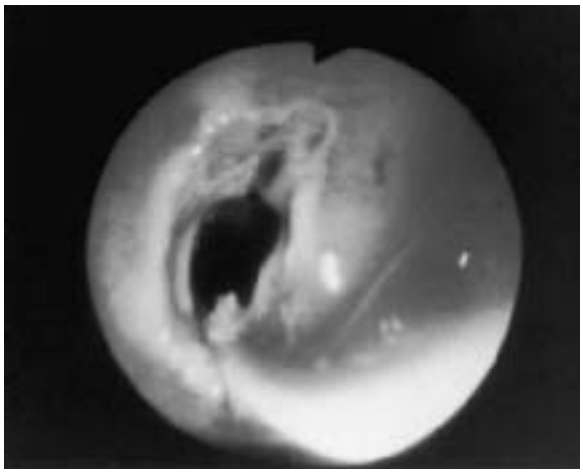


“ ”







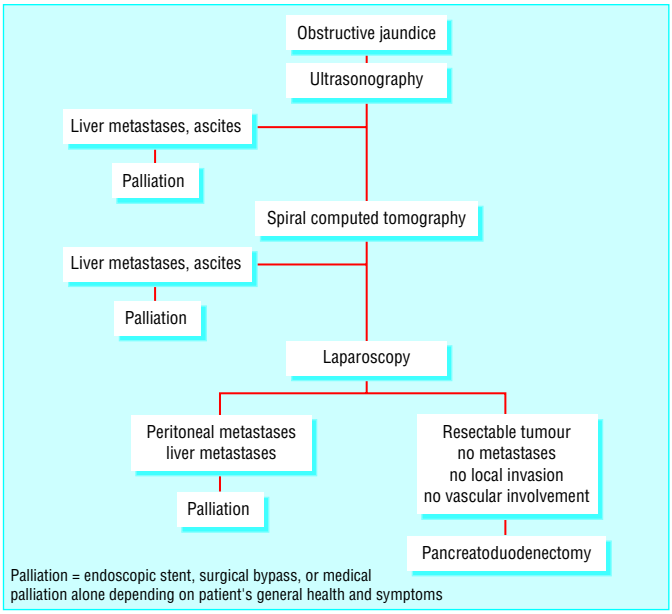


-
-
-
-
-

Br J Surg

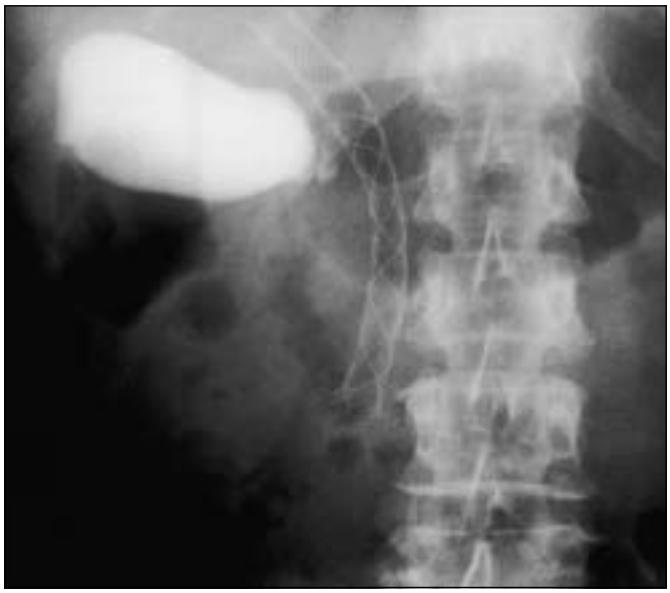
Hepatobiliary and pancreatic disease. The team approach to management

Diseases of the gut and pancreas

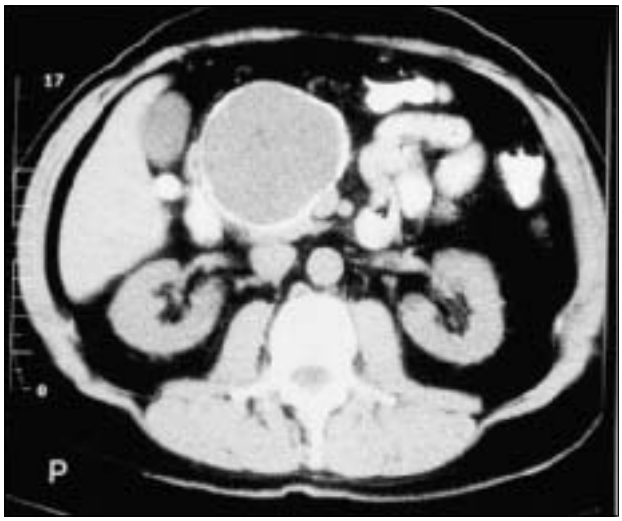


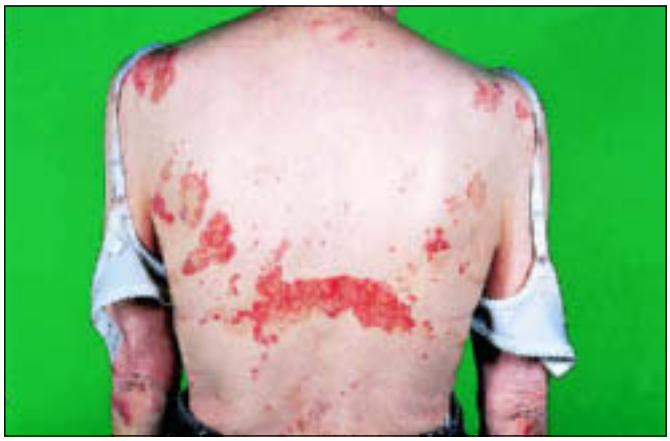
Resectable tumours

Locally advanced disease



Metastatic disease



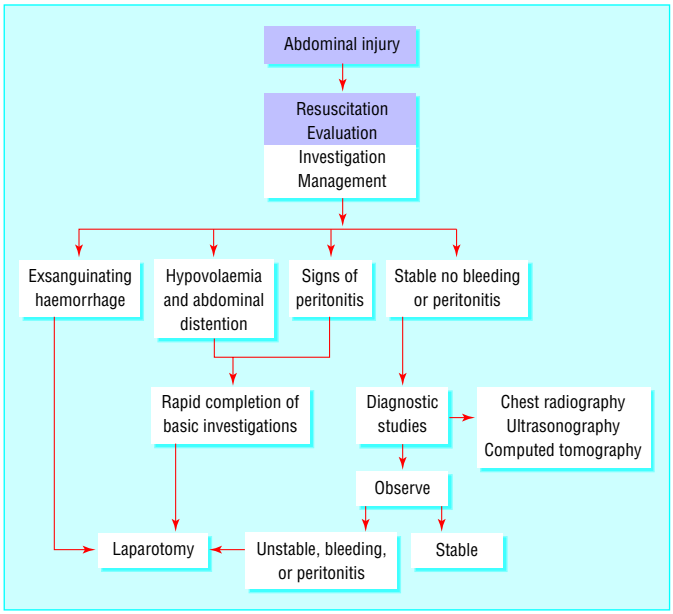


-
-
-
-
-
-

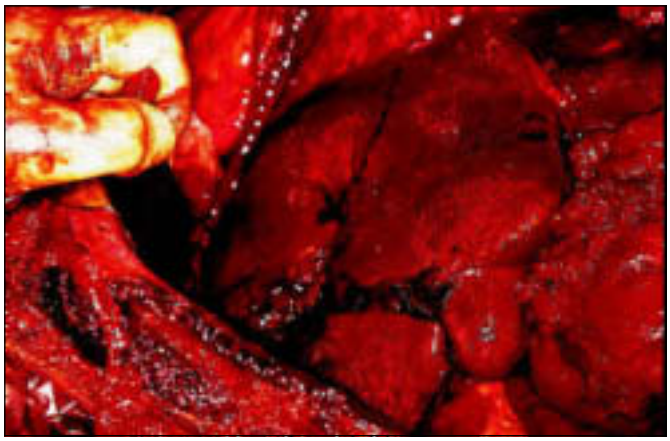
Surgery of the pancreas

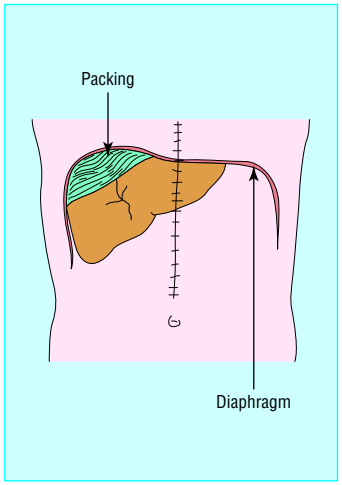
Hepatobiliary and pancreatic disease—the team approach to management

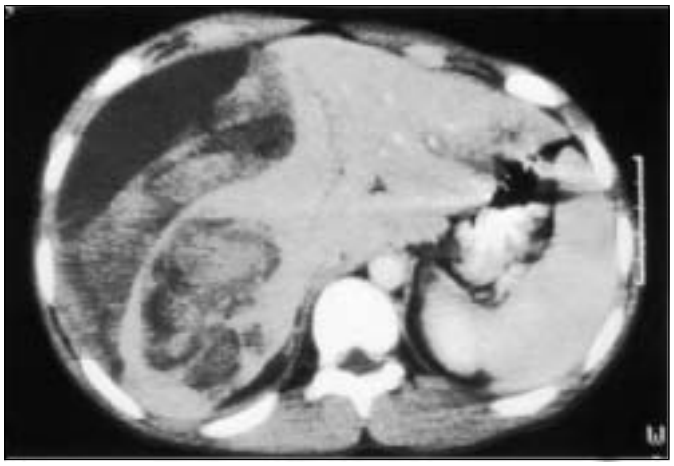
Practical gastrointestinal endoscopy

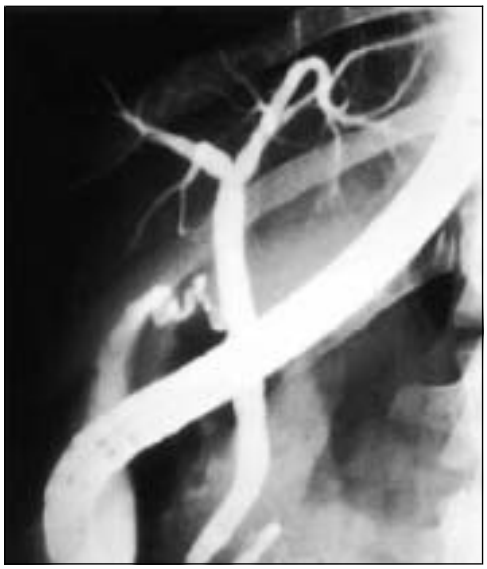

Abdominal injury
Resuscitation
Evaluation
Investigation
Management
Exsanguinating haemorrhage
Hypovolaemia and abdominal distention
Signs of peritonitis
Stable no bleeding or peritonitis
Rapid completion of basic investigations
Diagnostic studies
Chest radiography
Ultrasonography
Computed tomography
Observe
Laparotomy
Unstable, bleeding, or peritonitis
Stable

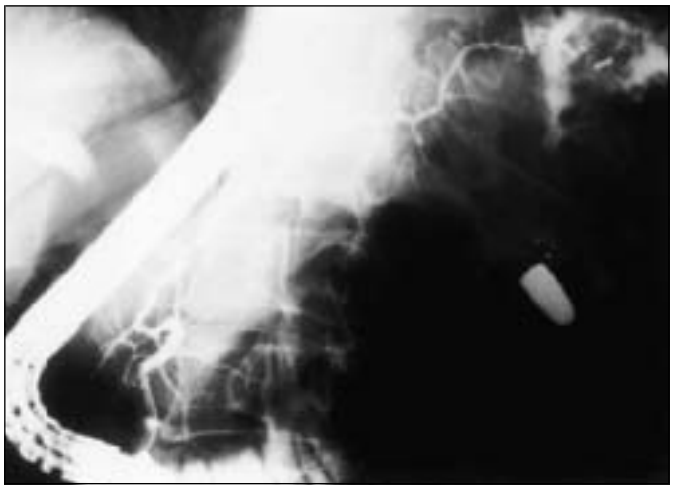
Operative approach











-
-
-
-
-
-

Hepatobiliary and pancreatic disease

Am J Surg

Hepatobiliary and pancreatic disease

Hepatectomy in the organ recipient is the most difficult part of the operation as the patient is at risk of developing a serious haemorrhage due to a combination of portal hypertension, defective clotting, and fibrinolysis. Improvements in surgical technique and anaesthesia have resulted in large reductions in blood loss, and the average requirement for transfusion is now four units of blood. At reimplantation, the suprahepatic and infrahepatic inferior vena cava and the portal vein are anastomosed and the organ is reperfused with blood. This is followed by reconstruction of the hepatic artery and bile duct.

Postoperative management

Patients are usually managed in an intensive care unit for the first 12-24 hours after surgery. Enteral feeding is restarted as early as possible, and liver function tests are done daily. Immunosuppressive protocols usually include a combination of cyclosporin or tacrolimus together with azathioprine or mycophenolate mofetil and prednisolone. The dose of steroids is rapidly tapered off, and they can often be stopped after two to three months. The doses of cyclosporin or tacrolimus are reduced gradually during the first year (during which pregnancy should be avoided) and continued at much lower levels for life.

Acute rejection occurs in about half of patients, but this is easily treated in most cases with extra steroids or by altering the drug regimen. Despite routine use of prophylactic treatment against bacterial, viral, and fungal pathogens, infections remain a major cause of morbidity. The side effects of the drugs are usually well controlled before the patient leaves hospital about two weeks after surgery.

At discharge, patients need to be familiarised with the drug regimen and side effects and educated about the warning signs of rejection and infection. Patients are usually followed up weekly for the first three months and then at gradually increasing intervals thereafter.

Results

The five year survival is 60-90%, depending on the primary disease and the clinical state of the patient before transplantation. The newer antiviral drugs plus the preoperative and postoperative adjuvant therapies for malignancies should lead to further improvements in survival. Although alcoholic liver disease remains a controversial indication for transplantation, carefully selected patients do well.

After successful transplantation patients have a greatly improved lifestyle and are often able to return to work and normal social activities. However, some patients experience medical and social problems. Drug compliance is one of the biggest problems after all types of organ transplantation. Poor compliance leads to chronic rejection and loss of the graft.

An extensive network of support services is available to help liver transplant patients. These include the transplant team, referring physician, general practitioner, social services, and local liver patient support groups. Shared care protocols operate in most regions, with most patients cared for primarily by their general practitioner and a gastroenterologist at their local hospital. The mainstay of follow up is regular liver function tests to detect any dysfunction of the transplant. Regular discussion of concerns with the transplant team is essential, and many problems can be sorted out by telephone.

Paediatric liver transplantation

In children, the most common indication for liver transplantation is biliary atresia, often after failure to respond to a portoenterostomy. Most children who need a liver transplant are young (under 3 years) and small (< 20 kg). Size matched donors are in short supply, and reduced size ("cut down") and

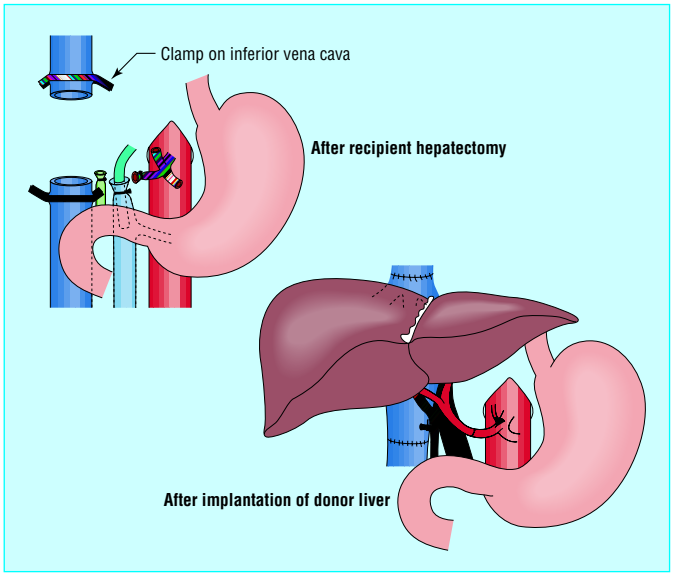


Figure 14.2 Implantation of liver transplant after hepatectomy

Table 14.1 Side effects of immunosuppresive drugs

Drug	Side effect	Monitoring
Cyclosporin	Neurotoxicity, nephrotoxicity, hypertension, hirsutism, gum hyperplasia, diabetes	Drug concentrations
Tacrolimus	Nephrotoxicity, neurotoxicity, hair loss, hypertension, diabetes	Drug concentrations
Azathioprine	Leucopenia, hair loss	White blood cell count
Mycophenolate mofetil	Gastrointestinal upset, leucopenia	White blood cell count and gastrointestinal symptoms
Steroids	Osteoporosis, diabetes, cushingoid face, hypertension	Symptoms
General	Infections, malignancy	Liver and renal function tests, regular follow up, and high index of suspicion

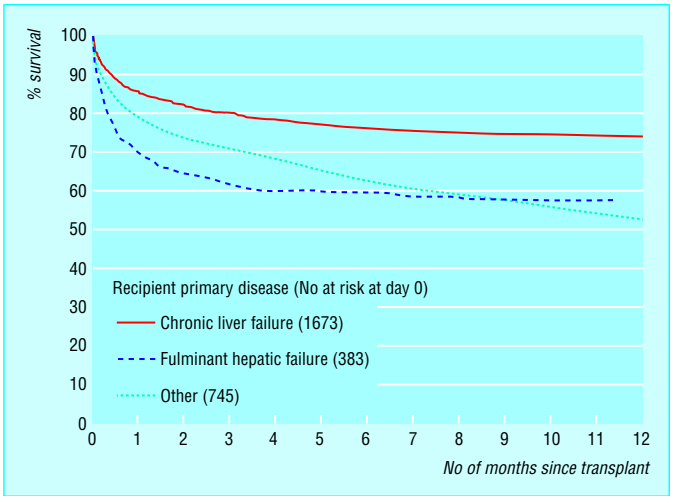


Figure 14.3 One year survival after first liver transplant according to primary disease, United Kingdom 1985-94

split (where one liver is split between two recipients) liver techniques have been used to overcome this problem. Donation of the left lobe of the liver by a living relative is also possible.

The shortage of child liver donors has been partly resolved by using smaller sections of adult livers, usually the left lobe

Pancreatic transplantation

The goals of transplantation of the pancreas are to eliminate the morbidity associated with labile blood glucose concentrations, stabilise or improve secondary diabetic complications, and improve the quality of life of patients with diabetes mellitus by restoring normal glucose metabolism. The stabilisation of diabetic control, the avoidance of exogenous insulin, and the ability to return to a normal diet for the first time since childhood are indisputable benefits of this procedure.

The selection of recipients for pancreatic transplantation is crucial. The magnitude of the surgery and need for long term immunosuppression means that whole organ transplantation is currently reserved for patients with end stage disease. Recipients are typically young (<50 years) with type 1 diabetes and end stage renal disease but without untreatable peripheral vascular or coronary artery disease. Simultaneous transplantation of the pancreas and kidney is the commonest procedure. Separate transplantation of the pancreas after kidney transplantation increases the chances of getting a good HLA matching for the kidney and allows a kidney to be donated by a living relative. The presence of immunosuppression at the time of implantation of the pancreas is also advantageous. The transplanted pancreas is usually placed in the pelvis and anastamosed to the iliac vessels, with the pancreatic duct anastomosed to the bladder or a loop of small bowel.

First year mortality is 3-10% in large units, with most deaths due to overwhelming sepsis. Transplant survival is 86% for the kidney and 70% for the pancreas. Successful transplantation greatly improves quality of life, and most patients are fully rehabilitated. Glucose homoeostasis seems to be excellent after pancreatic transplantation. Patients can stop exogenous insulin treatment and have normal glycated haemoglobin concentrations and glucose tolerance test results within three months of transplantation.

The long term effect on diabetic complications will not be known for several years, but recent results are encouraging. Evidence that diabetic nephropathy does not recur in the kidney transplant is accumulating, but there is no evidence for amelioration of established glomerular lesions in native kidneys. Improvements in autonomic and peripheral neuropathy have been documented. Further studies are needed to examine the potential for reducing the rate of progression of retinopathy and macrovascular disease.

Isolated pancreatic islet transplantation

A more logical approach is to attempt to prevent the development of the irreversible complications of diabetes by improving blood glucose metabolism at an early stage. Transplantation of pancreatic islet cells has been studied extensively as an alternative to whole organ grafting and has several theoretical and practical advantages. Pancreatic islets can be isolated by using collagenase digestion to separate the endocrine from the exocrine tissues and purified by density gradient separation. Some difficulties remain, particularly with the purification stage. The islets are injected into the recipient liver via the portal vein or by subcapsular injection into the kidney or spleen. Rejection of the islets remains a problem, and the success rates of this type of transplantation have been poor in the clinical setting.

Table 14.2 Types of pancreatic transplantation

Type	Indication
Simultaneous pancreas and kidney transplant (SPK)	Diabetic renal failure
Pancreas after kidney transplant (PAK)	After successful kidney transplant
Pancreas transplant alone (PTA)	Prerenal failure, unstable diabetic control, severe neuropathy
Segmental (transplantation of pancreatic tail)	Applicable to live donation
Multivisceral (pancreas transplanted with liver and sometimes small bowel)	For example, extensive abdominal tumour
Isolated pancreatic islets	The future solution?

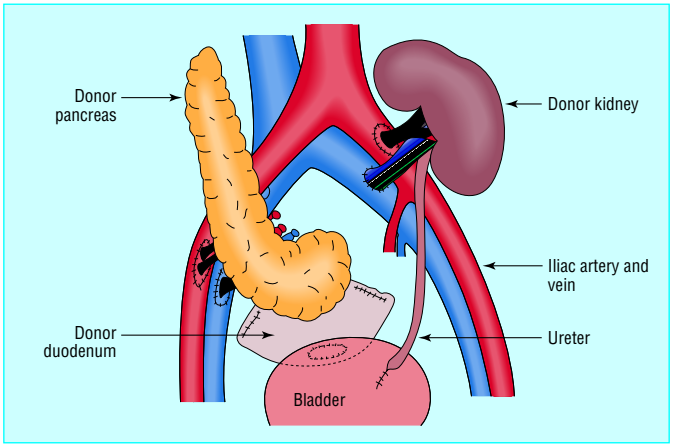


Figure 14.4 Simultaneous transplantation of pancreas and kidney with bladder drainage

Summary points

- Hepatitis C cirrhosis is the commonest worldwide indication for liver transplantation
- Alcoholic liver disease remains a controversial indication for liver transplantation but carefully selected patients do well
- Patients with chronic liver disease and signs of decompensation should be assessed for transplantation before they become critically ill
- Drug compliance is an important problem, with poor compliance leading to chronic rejection and graft loss
- Paracetamol overdose is the commonest cause of acute liver failure in the United Kingdom and accounts for 5% of all liver transplants in Britain
- Pancreas transplantation is most commonly performed for patients with end stage diabetes mellitus and renal failure

The photo of donor liver was obtained from J L Martha/Science Photo Library.

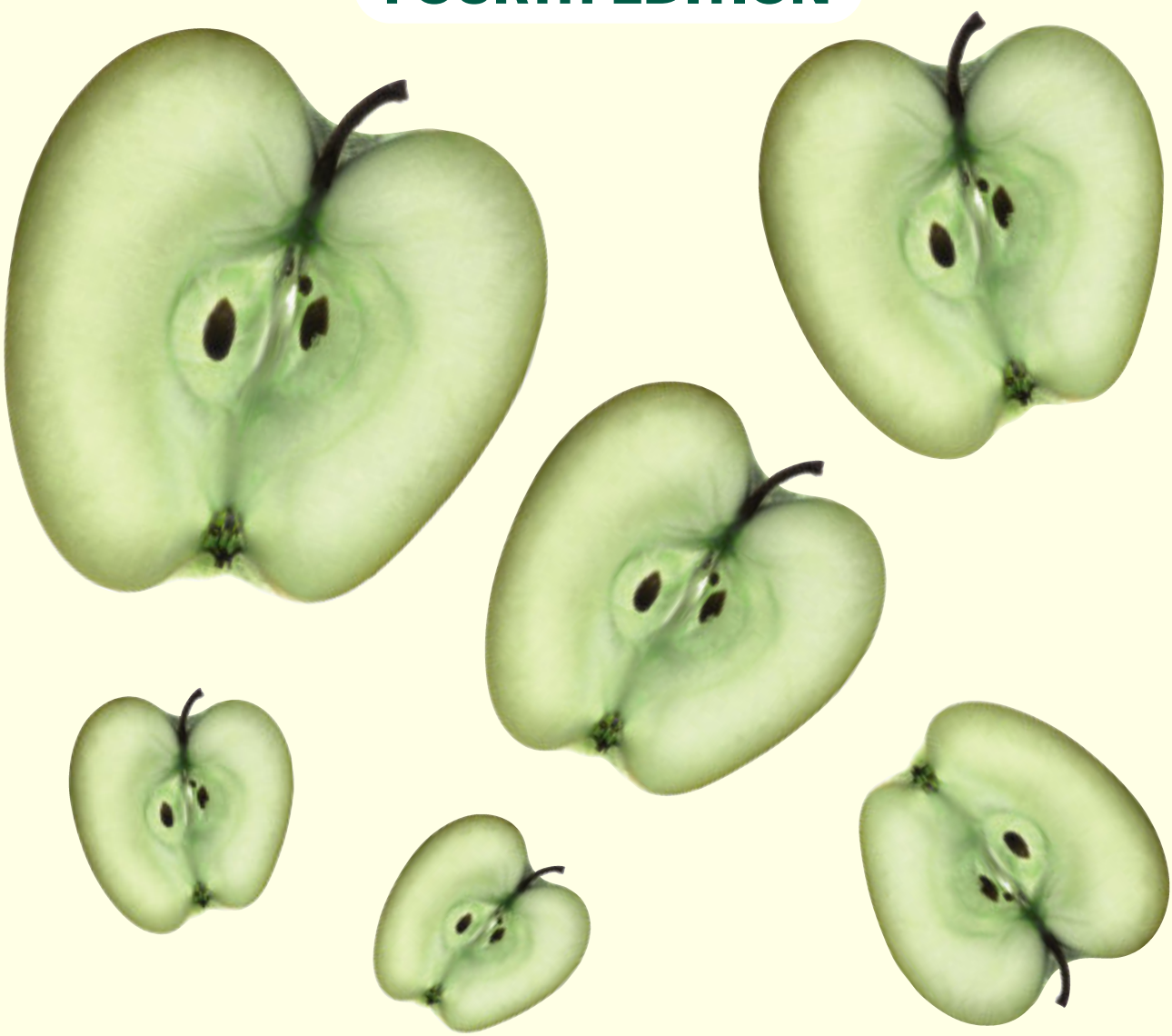
This Page Intentionally Left Blank

Index

Index

ABC OF NUTRITION

FOURTH EDITION



Written by A Stewart Truswell

BMJ Books

ABC OF NUTRITION

Fourth Edition

ABC OF NUTRITION

Fourth Edition

A STEWART TRUSWELL
Emeritus Professor of Human Nutrition,
University of Sydney, Australia

with contributions from
PATRICK G WALL
CIARA E O'REILLY
the late CHRISTOPHER R PENNINGTON
NIGEL REYNOLDS

BMJ Books

ЗНАНАР

First published in 1986
by BMJ Books, BMA House, Tavistock Square,
London WC1H 9JR
www.bmjbooks.com

First edition 1986
Second edition 1992
Third edition 1999
Fourth edition 2003

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0 7279 1664 5

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed and bound in Spain by Graphycems, Navarra

Cover shows halved apple, with permission
from Gusto productions/Science Photo Library

Contents

Contributors

Ciara E O'Reilly PhD
Technical Executive, Food Safety Authority of Ireland, Dublin, Ireland

Christopher R Pennington MD, FRCPEd
Late Professor of Gastroenterology,
Ninewells Hospital and Medical School, Dundee, Scotland

Nigel Reynolds MB, ChB, MRCP
Medicine and Cardiovascular Group, Department of Digestive Diseases and Clinical Nutrition, Ninewells Hospital and Medical School, Dundee, Scotland

A Stewart Truswell AO, MD, DSc, FRCP, FRACP
Emeritus Professor of Human Nutrition, University of Sydney, Australia

Patrick G Wall MB, BCh, BAO, MRCVS, MFPMM
Chief Executive, Food Safety Authority of Ireland, Dublin, Ireland

Preface

Preface to 3rd edition

Nutrition is one of those subjects which comes up every day in general practice—or should do—yet in most undergraduate medical schools it is crowded out by the big clinical specialities and high technology procedures. It is for subjects like nutrition that the British Medical Journal's ABC series is extremely useful.

This book was started when Dr Stephen Lock, previous editor of the BMJ asked me to write a series of weekly articles for an imagined general practitioner, in an unfashionable provincial town who had been taught almost no nutrition at medical school. They now felt the need to use nutrition in the practice, but could spare only 15 to 20 minutes a week to read about it.

The brief was that the writing must be practical and relevant; about half the page was to be for tables, figures, photographs or boxes (that is, not text) and these have to tell part of the story. The writing was to "come down off the fence", to make up its mind on the balance of evidence and state it plainly. The first edition had no references but some reviewers asked for them and now in the era of evidence-based medicine some well chosen references seem indispensable when writing about nutrition.

Nutritional concepts, of course, are not as tightly evidence-based as information about drugs because randomised controlled trials, so routine for drug therapy, are rare for nutrition.

This book does not deal with all aspects of human nutrition, only those that are useful in everyday medical practice. The latest fads and controversies are not here either. This is the ABC of Nutrition, not the XYZ.

A Stewart Truswell
1999

Preface to 4th edition

When the first edition of this ABC was written in 1985 there was no "evidence-based medicine", no human genome, no BSE or nvCJD, no epidemic of obesity and associated type II diabetes; there were no statins to lower plasma cholesterol and no genetically modified foods. *Helicobacter pylori* had just been discovered. The role of folate in neural tube defects had not been established, or raised plasma homocysteine as a risk factor for heart disease. The Barker hypothesis had not been propounded. These recent discoveries and ideas affect nutritional practice and they appear or influence what is in this new edition.

A Stewart Truswell
2003

1 Reducing the risk of coronary heart disease

For some doctors in affluent countries the first question about prevention of coronary heart disease (CHD) nowadays is whether to write a prescription for one of the statins (simvastatin, pravastatin, fluvastatin, atorvastatin, etc) which inhibit an early step of cholesterol biosynthesis in the body (see p 7). Tables are available to show whether the 5- or 10-year risk justifies the cost of long term statin medication, but the relation of diet and CHD is still of primary importance for the majority of people. **What we eat is bound up with the aetiology of CHD**. Many people do not know their current plasma cholesterol, many coronary deaths occur before medical help and most countries cannot afford these expensive drugs.

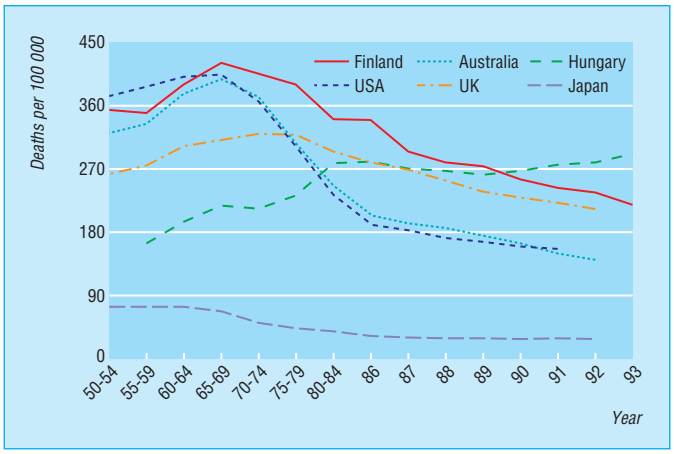
Coronary heart disease is the largest single cause of death in Britain and the disease that causes most premature deaths, but it is only one-seventh as common in industrial Japan and rare in the masses in most developing countries. Its incidence must be environmentally determined because **immigrant groups** soon take on the incidence rate of their new country and there have been large changes in mortality over time. Coronary heart disease was uncommon everywhere before 1925 and then increased steadily in Western countries until the 1970s, except for a dip during the Second World War. Age-standardised mortality rates from coronary heart disease in the United States of America and Australia **started to decline** from 1966 and have reduced by more than 70%. In Britain rates are higher in Scotland and Ireland than in England, and higher in the north of England than the south. They have been declining since 1979 and have fallen by about 25%. Most EU countries have shown similar recent modest reductions of coronary mortality, but in the countries of eastern Europe coronary mortalities have risen. They have, however, recently fallen in Poland and the Czech Republic.

Coronary heart disease is a **multifactorial disease**, but diet is probably the fundamental environmental factor. The pathological basis is **atherosclerosis**, which takes years to develop. **Thrombosis** superimposed on an atherosclerotic plaque, which takes hours, usually precipitates a clinical event. Then whether the patient dies suddenly, has a classic **myocardial infarct**, develops **angina**, or has asymptomatic electrocardiographic changes depends on the state of the myocardium. Each of these three processes is affected by somewhat different components in the diet.

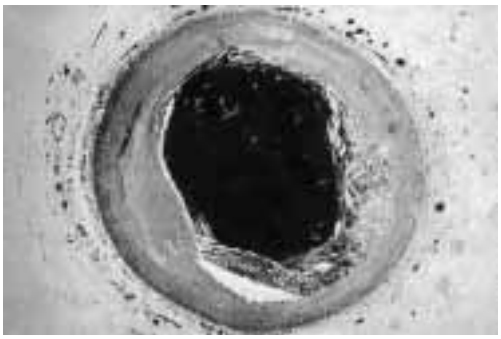
The characteristic material that accumulates in atherosclerosis is **cholesterol ester**. This and other lipids in the plaque, such as yellow carotenoid pigments, come from the blood where they are carried on low density lipoprotein (LDL). In animals, including primates, atheroma can be produced by raising plasma cholesterol concentrations with high animal fat diets. Much of this cholesterol is present in modified macrophages that have the histological appearance of foam cells. Experimental pathology studies indicate that these cells only take up large amounts of LDL if *it has been oxidised*.[2] This oxidation probably occurs within the artery wall.

People with genetically raised LDL-cholesterol (*familial hypercholesterolaemia*) tend to have premature coronary heart disease. This is accelerated even more in homozygotes who have plasma cholesterols four times normal and all develop clinical coronary heart disease before they are 20.

Thousands of papers have been written on diet and CHD. Since early in the century scientists have suggested links



Coronary heart disease death rates in six countries, for men aged 25-74, 1950-83. (Adapted from *Heart and Stroke Facts* published by the National Heart Foundation of Australia, from WHO data.) CHD mortality in USA and Australia started to fall 10 years before any decline in UK coronary deaths and fell more profoundly. Smoking rates and medical treatments cannot explain these phenomena. They may have been due to dietary changes (increased polyunsaturated and decreased saturated fatty acids)[1]



Photomicrograph of coronary artery with atherosclerosis

Evidence linking diet and CHD

This comes from:
- animal experiments
- pathology studies
- genetic polymorphisms
- epidemiology: ecological and cohort/prospective studies
- randomised controlled trials with dietary changes.

The strongest body of evidence comes from cohort studies which demonstrate environmental factors that are either associated with increased subsequent risk of CHD events (risk factors) or decreased subsequent risk (protective factors).

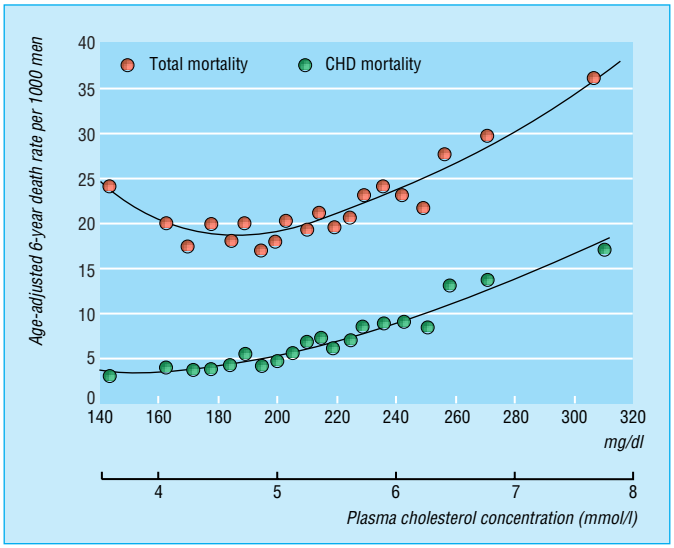
between a series of dietary components and CHD. Some of these were subsequently found to be unconnected or of little importance, for example sucrose, soft water, milk. The latest component to be associated is in the news, but this does not mean that the older components have been disproved—just that well-established facts are not newsworthy.

Risk factors

Over 50 prospective (cohort) studies in more than 600 000 subjects in 21 countries have reported on risk factors associated with or protective against CHD. The three best established risk factors are: raised plasma total and LDL-cholesterol, cigarette smoking, and high blood pressure.[3]

Two step reasoning

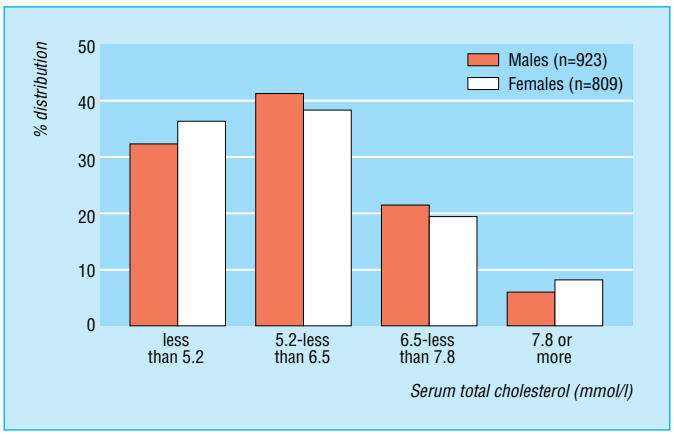
High plasma LDL- (and total) cholesterol is firmly established as a major risk factor for CHD, both from cohort study epidemiology and from randomised controlled trials with statins. In turn, how diet affects LDL-cholesterol concentration can be—and has been—demonstrated in controlled human dietary experiments, in which one dietary component is changed in the experimental period, with control periods on either side or in parallel.



Within-population relation between plasma cholesterol and CHD and total mortality based on 6-year follow up of 350 000 US men. (Adapted from Martin *et al.*[4]) The increased total mortality at (only) the lowest cholesterol concentration is thought to reflect acute and chronic illnesses (which often lower plasma cholesterol)

Plasma total and low density lipoprotein cholesterol (LDL-cholesterol)

About three quarters of plasma total cholesterol is normally in LDL-cholesterol and the higher the total cholesterol the higher the percentage of LDL-cholesterol because HDL-cholesterol rarely exceeds 2 mmol/l (and never exceeds 3). The **mean** plasma total cholesterol of healthy adults ranges widely in different communities, from 2.6 mmol/l (Papua New Guinea highlanders) to 7.2 mmol/l (in east Finland some years ago). Only in countries whose average total cholesterol exceeds 5.2 mmol/l (200 mg/dl)—as in Britain—is coronary heart disease common.



Percentage distribution of serum total cholesterol in British adults by sex (Adapted from Gregory *et al.*[5])

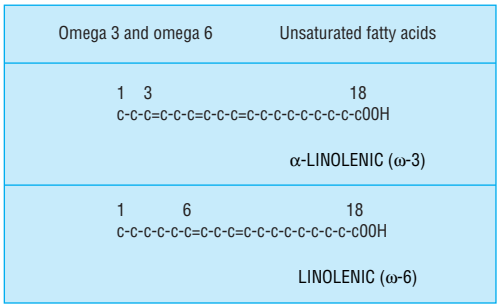
Dietary components that affect plasma LDL-cholesterol: type of fat

The major influence is the type of fat. Fats in the diet are mostly in the form of triglycerides (triacylglycerols): three fatty acids joined to glycerol. The most abundant fatty acid(s) determine(s) the effect. *Saturated fatty acids* raise LDL-cholesterol; these are mostly 12:0 (lauric), 14:0 (myristic), and 16:0 (palmitic). Palmitic may be less potent but is the most abundant of these saturated fatty acids in foods. 18:0 (stearic) has little or no cholesterol-raising effect.

Monounsaturated fatty acids—the main one is 18:1 (oleic)—in the natural *cis* configuration have an intermediate effect on LDL-cholesterol: lower than on saturated fatty acids, not as low as on linoleic.

Polyunsaturated fatty acids (PUFA), (with two or more double bonds) lower LDL-cholesterol. The most abundant of these in foods is 18:2 (linoleic) which belongs to the ω-6 (omega-6 or n minus 6, n−6) family of polyunsaturated fatty acids (first double bond, numbering from the non-carboxylic acid end is at 6th carbon). The omega-3 (ω-3) series of PUFAs are less abundant in most foods 18:3, ω-3, α-linolenic occurs in plants and some vegetable oils. 20:5, ω-3, eicosapentaenoic acid (EPA) and 22:6, ω-3, docosahexaenoic acid (DHA) are mostly obtained from fatty fish and fish oils. The cholesterol-lowering effect of ω-3 PUFAs is less important than their other properties (p 6).

In unsaturated fatty acids the double bond is normally in the *cis* configuration and the carbon chain bends at the double



bond. If the configuration is *trans*, straight at the double bond, the fatty acid behaves biologically like a saturated fatty acid. The usual *trans* fatty acid is 18:1 *trans* (elaidic) acid, found in foods produced by hydrogenation in making older-type hard margarines.

Dietary cholesterol and phytosterols

Cholesterol is only found in animal foods. Dietary cholesterol has less plasma cholesterol-raising effect than saturated fats. This is because about half the plasma cholesterol comes from the diet and half is biosynthesised in the liver from acetate. When more cholesterol is absorbed it tends to switch off this endogenous synthesis.

Plant oils also contain sterols, but these are **phytosterols**, for example, β-sitosterol, campesterol, brassicasterol. These typically have one or two more extra carbons on the side chain of the cholesterol molecule. They interfere competitively with cholesterol absorption and are poorly absorbed themselves. Phytosterols in vegetable oils (200-500 mg/100 g) add a little to their cholesterol-lowering effect. They are also present in nuts and seeds. Some premium PUFA margarines (introduced 1999) are enriched with concentrated natural phytosterols (or-stanols) to enhance cholesterol lowering.

Overweight and obesity

Overweight people tend to have raised plasma triglycerides and to a lesser extent total and LDL-cholesterol. Weight reduction by diet and/or exercise will usually reduce their cholesterol. Overweight, especially abdominal visceral adiposity, is itself a direct risk factor for CHD.

Dietary fibre

The effect of dietary fibre depends on the type. Wheat fibre (bran or wholemeal breads) does not lower plasma cholesterol but viscous ("soluble") types, pectin and guar and oat fibre, in large intakes, produce moderate cholesterol reductions. Although wheat fibre does not lower plasma cholesterol cohort studies consistently show less subsequent CHD in people who eat more wheat fibre and whole grain foods.[7]

Vegetable protein

Most vegetable foods are low in protein. Soya is an exception. When soya protein replaces animal protein in the diet there has usually been a reduction of plasma total and LDL-cholesterol. Although many human trials have been carried out, the mechanism has been elusive.

Coffee[9]

Coffee contains small amounts of diterpenes (lipids), cafestol and kahweol—not caffeine—that raise plasma total and LDL-cholesterol. Several cups a day of boiled, plunger or espresso coffee can raise the cholesterol but filtered or instant coffee does not—the diterpenes have been removed from the beverage.

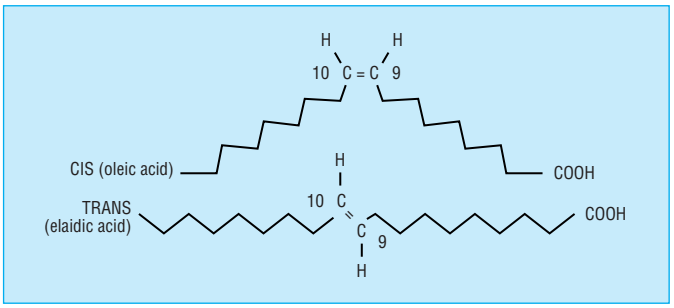
Mechanisms for LDL-cholesterol lowering

Many complex experiments have been done to elucidate how different fatty acids affect LDL-cholesterol. The main mechanism appears to be by effect on the number and activity of the LDL-receptors in cell membranes. Saturated fatty acids downregulate these receptors, so less cholesterol is taken up from the plasma; unsaturated fatty acids have the opposite effect. In overweight people there is increased secretion of very low density lipoprotein (VLDL) from the liver.

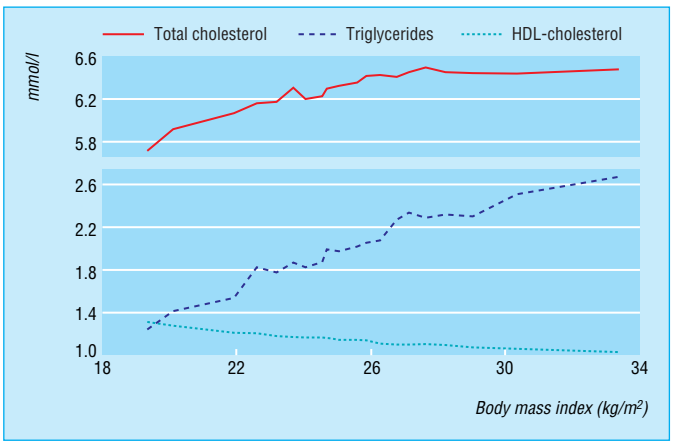
Effect of dietary fatty acids on plasma LDL-cholesterol

• Up to 10:0 (MCTs)	0
• 12:0 (lauric)	↑
• 14:0 (myristic)	↑↑
• 16:0 (palmitic)	↑
• 18:0 (stearic)	(↑)
• 18:1 *cis* (oleic)	(↓)
• 18:1 *trans*	↑↑
• 18:2 6-*cis* (linoleic)	↓
• Other polyunsaturates	(↓)

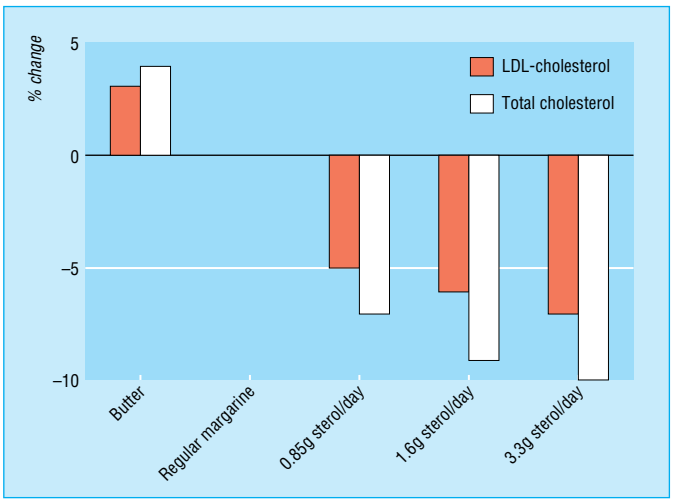
MCTs = medium chain triglycerides



Cis unsaturated fatty acids are bent at the double bond(s), *trans* fatty acids are not



The relation between body mass index (weight/height^2) and total cholesterol, HDL-cholesterol and triglycerides (all in mmol/l). (Adapted from Thelle *et al.*[6])



Plasma LDL and total cholesterol change over 3.5 weeks (double-blind, controlled trial) in 100 healthy human subjects who took in turn (randomised) butter, standard PUFA margarine or this enriched with different amounts of phytosterols. 20 g/day of the commercial product provides 1.6 g phytosterols[8]

Large amounts of viscous (soluble) dietary fibre increase viscosity in the lower small intestine and reduce reabsorption of bile acids, so producing negative sterol balance, hence increased cholesterol→bile acids (cholestyramine effect). The mechanism for the potent plasma cholesterol-raising effect of coffee lipids has not yet been worked out (plasma aminotransferase goes up too); no animal model has been found.

Plasma high density lipoprotein cholesterol (HDL-cholesterol)

HDL-cholesterol is a potent protective factor in communities with high LDL- and total cholesterols.[2] It appears to act by mobilising cholesterol from deposits in peripheral tissues, including arteries, and transporting it to the liver for disposal ("reverse cholesterol transport"). Levels of plasma HDL-cholesterol do not explain the big differences of coronary disease incidence between countries; its concentration is often lower in countries with little coronary heart disease. But in countries with a high incidence of CHD and high plasma-LDL-cholesterol, individuals with above average HDL-cholesterol have a lower risk of the disease. HDL-cholesterols are higher in women (related to oestrogen activity), a major reason why coronary disease usually affects women at older ages than men.

Low HDL-cholesterols are often associated with raised plasma triglycerides and the latter metabolic dysfunction may compound the risk of coronary disease. HDL-cholesterols tend to be lower in overweight people, in those with diabetes, and in those who smoke. They may be reduced by a high carbohydrate (that is, low fat) diet. They are raised by alcohol consumption, by moderate or heavy exercise, by reduction of body weight, and by high fat diets.

Increased HDL concentration is the clearest reason why moderate alcohol consumption is associated epidemiologically with reduced risk of CHD. Note that above two drinks per day, total mortality goes up because of other diseases and accidents associated with alcohol.

When someone changes from a typical Western diet to a low fat (therefore high carbohydrate) diet LDL-cholesterol goes down, (good!) because percentage saturated fat was reduced, but HDL-cholesterol goes down as well (may not be so good). If instead the fat intake is maintained but saturated fat is replaced by polyunsaturated and monounsaturated fats, LDL also goes down but with little or no reduction of HDL-cholesterol. Changing fat type like this should give a lower risk of coronary disease but reducing total fat intake is better for the management of overweight.

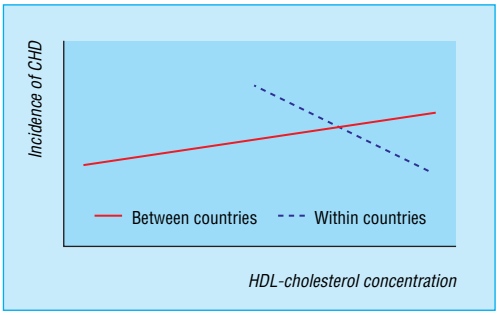
Plasma triglycerides

If a patient has raised plasma triglycerides the first question is whether they had been fasting when the blood was taken. The next question is whether the hypertriglyceridaemia is a pointer to other risk factors that tend to be associated with it: high plasma cholesterol, overweight, lack of exercise, glucose intolerance, low-HDL-cholesterol or other metabolic disease (renal disease, hypothyroidism). A common cause of increased plasma triglycerides is excessive alcohol indulgence the evening before blood was taken.

Risk factors for coronary heart disease

- High plasma total cholesterol
- High plasma LDL-cholesterol
- Low plasma HDL-cholesterol
- High plasma triglycerides
- High blood pressure
- (Cigarette smoking)
- Obesity; high intra-abdominal fat
- Diabetes mellitus
- (Lack of exercise)
- Increased plasma coagulation factors
- Increased platelet adhesiveness
- High plasma homocysteine
- Increased Lp(a)
- (Apo E4 genotype)

Factors in parentheses are not influenced by diet.



Relation of HDL-cholesterol to incidence of CHD. (Adapted from Knuiman and West[10])

Alcohol intake, coronary heart disease (CHD), and total mortality*

	Mortality-relative risk		
Stated alcohol consumption	**From CHD**	**From accidents**	**Total**
Non-drinkers	1.00	1.00	1.00
1/day	0.79	0.98	0.84
2/day	0.80	0.95	0.93
3/day	0.83	1.32	1.02
4/day	0.74	1.22	1.08
5/day	0.85	1.22	1.22
6+/day	0.92	1.73	1.38

* 12-year follow up of cohort of 276 802 US men by stated alcohol habits at entry. Reduced risk of CHD brought down total mortality at 1 and 2 drinks/day but not above
Reproduced from Boffeta and Garfinkel[11]

The management of hypertriglyceridaemia consists of looking for and dealing with any of the common associations. The non-pharmacological treatment is more exercise, fewer calories (weight reduction), and less alcohol. Reduced carbohydrate is not advised; it implies an increased fat intake which can only increase lipaemia during the day. People with exaggerated postprandial lipaemia appear to have an increased risk of coronary heart disease. Fish oil (for example, Maxepa) is a nutritional supplement with a powerful plasma triglyceride-lowering effect and regular consumption of fatty fish also lowers plasma triglycerides.

Other risk factors

High blood pressure is discussed in chapter 2; **overweight and inactivity** in chapter 11.

Increased levels of two of the **coagulation factors**, Factor VII and fibrinogen, have been clear in some prospective studies (they were not assayed in most studies).[13] **Factor VII** activity is increased during alimentary lipaemia after a fatty meal and is persistent in people with hypertriglyceridaemia. Plasma **fibrinogen** is raised in people who smoke and in obesity; it is reduced by alcohol consumption.

Antioxidants

The LDL oxidation hypothesis of atherogenesis predicts that if LDL carries more lipid-soluble antioxidants they should provide some protection against CHD. The principal antioxidant in LDL is α-tocopherol, vitamin E (average 7 tocopherol molecules per LDL particle). Its concentration can be raised by intake of vitamin E supplements. *In vitro* (outside the body) extra vitamin E delays the oxidation of LDL (by copper). In two large prospective studies, one in US nurses, the other in health professionals, those with high intakes of vitamin E experienced less subsequent CHD. But these high intakes of vitamin E were achieved by taking supplements, and people who regularly take vitamin supplements are likely to have more health conscious lifestyles than the average citizen.

Five large randomised controlled prevention trials, in Western populations, with acronyms ATBC,[14] GISSI,[15] HOPE, PPP, and CHAOS involving 56 000 subjects have now been reported. There was no reduction of cardiovascular disease or mortality. LDL contains smaller amounts of carotenoids, which are also lipid-soluble antioxidants. But supplements of β-carotene have also not prevented CHD in large randomised controlled trials.[14]

Polyunsaturated fatty acids, 18:2, 20:5 and 22:6 are more susceptible to peroxidation *in vitro* than saturated or monounsaturated acids but in the whole body there is a lot of evidence that PUFA intake is **negatively** associated with CHD.[16]

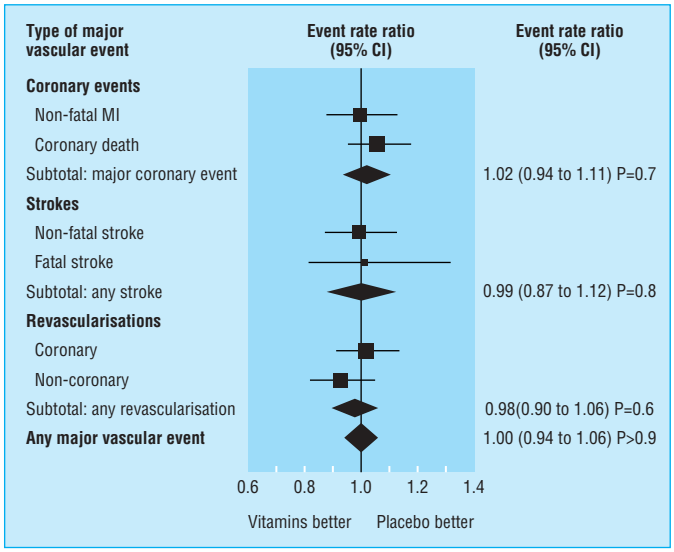
Plasma homocysteine

In the inborn error of metabolism homocystinuria, plasma homocysteine is so high that it spills into the urine and vascular diseases are among the complications. Then during the 1990s evidence accumulated (many case-control studies and several prospective studies) that lesser degrees of elevated plasma homocysteine (above 16 μmol/l total homocysteine, tHcy) are a largely independent risk factor for CHD. They also increase the risk of cerebral and peripheral arterial diseases and even venous thrombosis.[18] Raised plasma homocysteine appears to both damage the endothelium and increase liability to thrombosis.

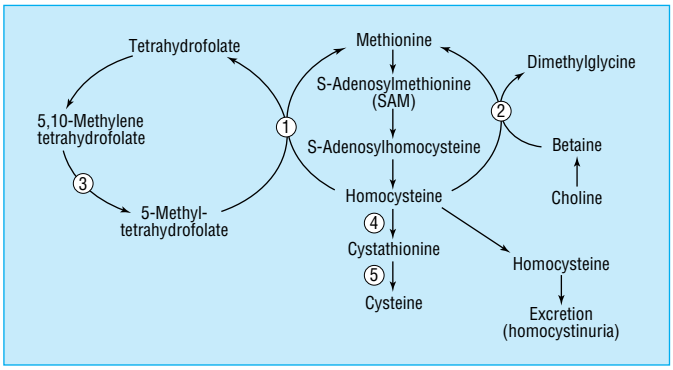
Homocysteine is an intermediary metabolite of the essential amino acid, methionine (it is methionine minus its terminal methyl group). Folic acid is co-factor for the enzyme in a pathway that re-methylates homocysteine back to methionine.

Plasma triglycerides

- Triglycerides in the blood after overnight fast are mainly in VLDL (very low density lipoprotein), synthesised in the liver, hence endogenous. Triglycerides in casual blood samples taken during the day may be mainly in chylomicrons, after a fatty meal, and hence exogenous.
- In prospective studies, raised fasting triglycerides have often shown up as a risk factor for coronary heart disease in single-factor analysis. But hypertriglyceridaemia is likely to be associated with raised plasma cholesterol, or overweight/obesity, or glucose intolerance, or lack of exercise or low HDL-cholesterol. When these are controlled, increased triglycerides is certainly not as strong a risk factor as hypercholesterolaemia but it has emerged in some studies as an independent coronary risk factor, more often in women.[12]



No significant benefit from vitamins C and E and β-carotene in MRC/BHF secondary prevention trial in over 20 000 subjects[17]



Homocysteine metabolism in humans. Enzymes [vitamins involved]: 1. *N*-5-methyltetrahydrofolate:homocysteine methyltransferase (methionine synthase) [folate, vitamin B-12]; 2. betaine:homocysteine methyltransferase; 3. methylene-tetrahydrofolate reductase (MTHFR) [folate]; 4. cystathione beta-synthase [vitamin B-6]; 5. gamma-cystathionase [vitamin B-6]

In apparently well-nourished people folic acid lowers elevated homocysteine by about a quarter.[19] A dose of 0.5 mg or even 200 μg folic acid is effective. Plasma homocysteine is also increased in mild vitamin B-12 deficiency. Folic acid may be a safe, inexpensive way of reducing vascular disease. Randomised controlled trials are under way.

Dangerous arrhythmias

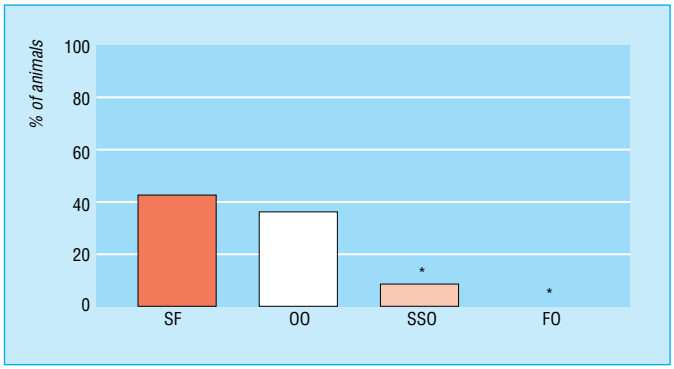
Dangerous arrhythmia is one of the two major causes of death in CHD. Over half the deaths occur before the arrival of paramedical or medical help. Then in the ambulance or coronary care unit the treatment of ventricular fibrillation saves lives. Developments in nutrition research are showing, with animal experiments, that electrical instability of ischaemic myocardium is influenced by the fatty acid pattern of the diet and hence of myocardial membranes. In rats or marmoset monkeys fed polyunsaturated oils, fewer animals had sustained ventricular arrhythmia when a coronary artery was tied, than in animals that had been fed on saturated fat or (monounsaturated) olive oil.[20] The fish oil group were more resistant to arrhythmia than the sunflower oil group (ω-6 linoleic acid). Canola oil containing linolenic acid (18:3, ω-3), the plant ω-3 fatty acid, also appears to reduce arrhythmias. Kang and Leaf have studied the mechanism of the fatty acid effect with cultured, neonatal, rat ventricular myocytes whose spontaneous contractions are recorded by a microscope and video camera. Eicosapentaenoic acid (20:5, ω-3) and the plant oil ω-3 acid, 18:3 (linolenic) as well as linoleic acid (18:2, ω-6) prevent tachyrhythmia induced by a variety of chemicals known to produce fatal ventricular fibrillation in humans. It appears that polyunsaturated fatty acids act by binding to sodium channel proteins in the membrane and altering their electrical charge.[21]

The reduction of deaths outside hospital has been a striking feature in countries where coronary death rates have reduced. This may be explained, at least partly, by an anti-arrhythmic effect of increased ω-6 polyunsaturated fat intake (national fish intakes have not increased).

Platelet function and thrombosis

In patients with symptomatic CHD tests of platelet function have usually indicated activation. Available tests of platelet function are not on lists of risk factors predicting coronary disease; they are *in vitro* tests and are inevitably indirect. However platelet activation is of course a central phenomenon in myocardial infarction or recurrent angina, so that any diet that reduces platelet aggregation should reduce the risk of coronary disease.

Following up an observation that the rarity of coronary disease in Greenland Eskimos might be due to their heavy consumption of marine fat, it was discovered that eicosapentaenoic acid (20:5, ω-3) or EPA, a principal fatty acid of fish oil, displaces arachidonic acid (20:4, ω-6) in platelets, so that when stimulated they produce an inactive thromboxane TXA_3 instead of the active TXA_2 derived from arachidonic acid. EPA is only present in traces in the body fat of land animals and is absent from vegetable oils. In human experiments fish oil also reduced the levels of PAI-I, plasminogen activator inhibitor-1. Fish oil is therefore a pharmaceutical alternative (for example Maxepa) to aspirin to reduce the tendency to thrombosis. Results have been mixed in trials with fish oils to see if they delay restenosis after coronary angioplasty.



Total mortality from irreversible ventricular fibrillation during ischaemia or reperfusion in rats fed a saturated fat (SF), olive oil (OO), sunflower seed oil (SSO), or fish oil (FO) diet for 12 weeks from 18 weeks of age. *Significantly different from SF, $P < 0.05$. (Adapted from McLennan *et al.*[20])

Coronary deaths per 100 000 in men in three Australasian cities using standardised MONICA criteria

	1984	1993	Change
Coronary deaths before patient in hospital			
Auckland (NZ)	188	141	47
Newcastle (NSW)	186	102	84
Perth (WA)	128	78	50
Coronary deaths after patient in hospital			
Auckland (NZ)	57	24	33
Newcastle (NSW)	78	28	50
Perth (WA)	41	30	11

Reproduced from Beaglehole R *et al.*[22] with permission from Oxford University Press

Effects of fish oil

↑ EPA and DHA in plasma and red cells
↓ Arrhythmias in ischaemic myocardium
↓ Platelet aggregation
↓ PA1-1, ↓ fibrinogen, ↓ TPA
↓ Fibrinolysis
↑ Bleeding time
↓ Fasting plasma VLDL and triglycerides
↓ Postprandial lipaemia

DHA = docosahexaenoic acid (22:6, ω-3), TPA = tissue plasminogen activator

More on diets and platelet function

- Several prospective studies (in countries with intermediate fish intake) and a secondary prevention trial in Cardiff [23] suggest that a modest intake of fatty fish (for example sardines, herring, mackerel, or salmon) two or three times a week may help to prevent coronary heart disease. The EPA in this amount of fish is less than that needed (at least 2 g of EPA per day) to inhibit platelet aggregation.
- ω-6 polyunsaturated oils also appear to have an inhibiting effect on platelet function. They are less active but people eat more plant seed oils than fish oil.
- Heavy alcohol ingestion exerts an inhibitory effect on platelet function, which is reversible on abstinence.

Dietary components associated directly with coronary disease in cohort epidemiological studies

Most of the many prospective studies involving coronary heart disease have not measured diet. It is much more complex and expensive to estimate all the different foods, and thence to compute all the nutrients, than to measure blood pressure or plasma lipids. Of all the parts of a total diet there have been most reports of **alcohol** intake. It is simpler to include in a questionnaire than to tackle the intricacies of type of fat intake.

In the minority of prospective studies that did report on foods or food components, most have used food frequency questionnaires (chapter 12), which are easier to handle than open-ended dietary records. Another method, occasionally used, is to measure objective biomarkers of food intake such as plasma fatty acid pattern. Interpretation of associations in the table must allow for uncertainties in assessing usual food intake, and confounding between different food components and with lifestyle. Vitamin E findings have not been confirmed in randomised controlled trials.

Dietary components directly related to CHD

	Component	No. of studies
	Alcohol ↓	29/38
	Fruits and vegetables ↓	many
	Cereal fibre ↓	several
	Saturated fatty acids ↑	4/4
	ω-6 PUFA ↓	6/12
	Trans fatty acids ↑	2/3
	α-linolenic (ω-3) ↓	2/3
	Fish ↓	5/11
	Coffee ↑ or 0	14
	Nuts ↓	3/3
	Vitamin E ↓	3/3
	Folate ↓	3/3
And	0 for eggs (2/2) and iron intake (7/9)	
	↑ Increased risk	↓ Protective

Adding a statin to the diet

Treatment with statins lowers raised plasma cholesterol by average 20% and LDL-cholesterol 25%, without lowering HDL-cholesterol, and reduces subsequent CHD events significantly. Statin treatment has also been shown to reduce CHD events by about 24% in people who had survived a myocardial infarction and had average plasma cholesterols of around 5.4 mmol/l.[26]

Note that a statin is prescribed (as the manufacturers state) as an **adjunct to diet** and normally after a proper trial of a cholesterol lowering diet. The dietary principles described in this chapter lower plasma cholesterol by different mechanisms from the HMG COA reductase inhibition by statins. Parts of diets used to protect against CHD do not act by lowering LDL-cholesterol, for example, only by diet and exercise can overweight be treated.

Statins are very expensive at present, either for the patient or the health service, and we do not yet know if there might be long-term complications. Put very simply the indications for adding a statin to diet are for patients with:

- existing clinical CHD
- two or more coronary risk factors and high plasma cholesterol
- no or one coronary risk factor and very high plasma cholesterol.

In assessing the plasma cholesterol, LDL-cholesterol should be used or total cholesterol/HDL-cholesterol (after repeat measurements in a good laboratory). Risk factors are diabetes, hypertension, smoking, strong family history.

Randomised controlled trials (RCTs) with diet or nutrients

- **Reduced saturated, increased ω-6 PUFA diets**

8 RCTs in UK, USA, Finland and Norway, published 1965-1992. Total 17 529 subjects. In intervention groups plasma cholesterol fell. Combined result CHD events 81% of control ($P < 0.05$) and total mortality 95%.[24]

- **Lyon "Mediterranean" diet**[25]

Intervention group used a canola margarine, rich in linolenic acid (18:3, ω-3): they ate more bread, fruit, legumes, fish, less meat and butter but showed no fall in plasma cholesterol. CHD events were significantly reduced but the mechanism and dietary components responsible are not clear.

- **Fish and fish oil**

One secondary prevention RCT with fish (DART)[22] and another with fish oil (GISSI)[15] reduced CHD events significantly.

- **Vitamin E and β-carotene** have both been ineffective in several RCTs.

The dietary prescription (consistent with NCEP[27] and DOH[28]

Total fat

Reduction is not essential for improving plasma lipids but should reduce coagulation factors and daytime plasma triglycerides and contribute to weight reduction.

Saturated fatty acids

Principally 14:0, 16:0 and 12:0 should be substantially reduced from around 15% of dietary energy in many Western diets to 8-10%.

Polyunsaturated fatty acids

Mainly linoleic acid (18:2, ω-6): they should be about 7% of dietary energy (present British level), up to 10%. Omega-3

polyunsaturated fatty acids should be increased, both 20:5 and 22:6 from seafoods and 18:3 from canola (rapeseed) oil, etc.

Monounsaturated fatty acids
Ideal intake if total fat 30%, saturated 10% and polyunsaturated 8% would be 12% of total dietary energy.

Trans *fatty acids*
With the help of margarine manufacturers these have been reduced. The Department of Health recommends no more than 2% of dietary energy. Avoid older hard margarines.

Dietary cholesterol
This boils down to the question of egg yolks. Eggs are a nutritious, inexpensive and convenient food. The Department of Health recommends for the general population no rise in cholesterol intake.

Salt (NaCl)
Restriction to under 6 g/day is advised for the general population (100 mmol Na). It is more important for coronary patients.

Fish
The Department of Health recommends at least twice a week, preferably fatty fish. It should not be fried in saturated fat.

Fibre
Eat plenty of high fibre and whole grain cereal foods, including oatmeal.

Vegetables and fruit
These are low in fat, and contain pectin and other fibres, flavonoids and other antioxidants, and they contain folate. Expert Committees in Britain and the USA recommend five servings of different vegetables and fruit per day (400 g/day average weight).[28]

Soy products
(Not salty soy sauce) recommended.

Fatty acid patterns of fats, oils, and some meats (as % total fat in the food)

	Saturated					
	C4-12	C14-18 (myristic, palmitic, stearic)	Mono-unsaturated	Linoleic	Other poly-unsaturated	P:S*
Butter, cream, milk	13	48	30	2	1	0.05
Cocoa butter	—	61	36	3	—	0.05
Beef	—	48	48	2	1	0.06
Coconut oil	58	31	8	2	—	0.1
Bacon and pork	—	42	50	7	1	0.2
Palm oil (used in ice cream)	—	45	45	9	—	0.2
Margarine (old style, hard)	3	37	33**	12	1	0.3
Chicken	—	34	45	18	2	0.6
Olive oil	—	14	73	11	1	0.9
Groundnut oil	—	15	53	30	1	2.1
Fish oil	—	23	27	7	43†	2.2
Margarine, polyunsaturated	2	21	22	52	1	2.3
Corn (maize) oil	—	14	24	53	2	3.9
Soya bean oil	—	14	24	53	7	4.3
Canola oil	—	7	63	20	10‡	4.3
Sunflower seed oil	—	12	33	58	—	4.8
Flaxseed oil	—	9	18	16	57‡	8.0
Safflower oil	—	9	14	77	—	8.5

* These are in ascending order of ratio of polyunsaturated to saturated fats, but this is not the only consideration in choosing dietary fats and oils
† Includes varying amounts of 20:5 ω-3 and 22:6 ω-3 (depending on species).
‡ α Linolenic acid
** Includes variable amounts of *trans* 18:1

Alcohol
In moderation, one or two drinks per day is beneficial for middle-aged people at risk of CHD but cannot be recommended for the general population because of the greater danger of accidents in younger people and of all the medical complications of excessive intake.

Coffee
Should be instant or filtered.

References

1 Truswell AS. Cholesterol controversy. *BMJ* 1992; **304**: 912-13.
2 Steinberg D, Parthasarathy S, Carew TE, Khoo JC, Witztum JL. Beyond cholesterol: modifications of low-density lipoprotein that increase its atherogenicity. *N Engl J Med* 1989; **320**: 915-24.
3 Keys A, Aravanis C, Blackburn H *et al. Seven countries: a multivariate analysis of death and coronary heart disease.* Cambridge, Massachusetts: Harvard University Press, 1980.
4 Martin MJ, Hulley SB, Browner WS, Kuller LH, Wentworth D. Serum cholesterol, blood pressure and mortality implications from a cohort of 361 662 men. *Lancet* 1986; **ii**: 933-6.
5 Gregory J, Foster K, Tyler H, Wiseman M. *The dietary and nutritional survey of British adults.* London: HMSO 1990: 266.
6 Thelle DS, Shaper AG, Whitehead TP, Bullock DG, Ashby D, Patel J. Blood lipids in middle-aged British men. *Br Heart J* 1983; **49**: 205-13.
7 Truswell AS. Cereal grains and coronary heart disease. *Eur J Clin Nutr* 2002; **56**: 1-14.
8 Hendricks HFJ, Westrate JA, Van Vliet T, Meijer GW. Spreads enriched with three different levels of vegetable oil sterols and the degree of cholesterol lowering in normocholesterolaemic and mildly hypercholesterolaemic subjects. *Eur J Clin Nutr* 1999; **53**: 319-27.
9 Urgert R, Meybom S, Kuilman M *et al.* Comparison of effect of cafetiere and filtered coffee on serum concentrations of liver aminotransferases and lipids: six month randomised controlled trial. *BMJ* 1996; **314**: 1362-6.
10 Knuiman JT, West CA. Differences in HDL cholesterol between populations: no paradox? *Lancet* 1983; **i**: 296.
11 Boffeta P, Garfinkel L. Alcohol drinking and mortality among men enrolled in an American Cancer Society prospective study. *Epidemiology* 1990; **1**: 342-8.
12 Tunstall-Pedoe H, Woodward M, Tavendale R, Brook RA, McClusky MK. Comparison of the prediction by 27 different factors of coronary heart disease and death in men and women of the Scottish heart health study: cohort study. *BMJ* 1997; **315**: 722-9.
13 Miller GJ. Postprandial lipid metabolism and thrombosis. *Proc Nutr Soc* 1997; **56**: 739-44.
14 Rapola JM, Virtamo J, Ripatti S *et al.* Randomised trial of α-tocopherol and β-carotene supplements on incidence of major coronary events in men with previous myocardial infarction. *Lancet* 1997; **349**: 1715-20.
15 GISSI-Prevenzione Investigators. Dietary supplement with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: results of the GISSI-Prevenzione trial. *Lancet* 1999; **354**: 447-55.

16 Hu FB, Stampfer MJ, Manson J *et al.* Dietary fat intake and the risk of coronary heart disease in women. *N Engl J Med* 1997; **337**: 1491-9.
17 Heart Protection Study Group. MRC/BHF Heart Protection Study of antioxidant vitamins supplementation in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002; **360**: 23-33.
18 Bouskey CJ, Beresford SAA, Omenn GS, Motulsky AG. A quantitative assessment of plasma homocysteine as a risk factor for vascular disease. Probable benefits of increasing folic acid intake. *JAMA* 1995; **274**: 1049-57.
19 Homocysteine Lowering Trialists' Collaboration. Lowering blood homocysteine with folic acid based supplements: meta-analysis of randomised trials. *BMJ* 1998; **316**: 894-8.
20 McLennan PL. Relative effects of dietary saturated, monounsaturated and polyunsaturated fatty acids on cardiac arrhythmias in rats. *Am J Clin Nutr* 1993; **57**: 207-12.
21 Kang JX, Leaf A. Antiarrhythmic effect of polyunsaturated fatty acids. Recent studies. *Circulation* 1996; **94**: 1774-80.
22 Beaglehole R, Stewart AW, Jackson R. Declining rates of coronary disease in New Zealand and Australia. *Am J Epidemiol* 1997; **145**: 707-13.
23 Burr ML, Fehily AM, Gilbert JF *et al.* Effects of changes in fat, fish and fibre intakes on death and myocardial reinfarction: Diet and Reinfarction Trial (DART). *Lancet* 1989; **ii**: 757-61.
24 Truswell AS. Review of dietary intervention studies: effect on coronary events and on total mortality. *Aust NZ J Med* 1994; **24**: 98-106.
25 de Lorgeril M, Renaud, Mamalle N *et al.* Mediterranean alpha-linolenic acid-rich diet in secondary prevention of coronary heart disease. *Lancet* 1994; **343**: 1454-9.
26 Sacks FM, Pfeffer MA, Moye LA *et al.* The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *N Engl J Med* 1996; **335**: 1001-9.
27 Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel of Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA* 2001; **285**: 2486-97.
28 Department of Health. Nutritional Aspects of Cardiovascular Disease. *Report on the Cardiovascular Review Group, Committee on Medical Aspects of Food Policy*. London: HMSO, 1994.
29 National Heart Forum. *At least five a day. Strategies to increase vegetable and fruit consumption*. London: The Stationery Office, 1997.

2 Diet and blood pressure

Essential hypertension is a multifactorial disease. It is common in older people not only in urban and industrialised areas but also in a quiet Hebridean island, in tropical Africa, where Albert Schweizer used to work, and in an isolated Solomon Islands' tribe minimally influenced by Western ways, which cooks in sea water.[1]

Salt (sodium)

To what extent is essential hypertension related to an unnecessarily high intake of salt?

Hypertension is not an inevitable accompaniment of ageing. Evidence showed that hypertension did not occur in a few isolated communities, such as Yanomamo Indians (in the Amazon), Kalahari Bushmen (Botswana),[2] and remote Pacific islanders. These people typically had no access to salt and their urinary sodiums (reflecting salt intake) were under 30 mmol/day.

Salt and blood pressure history

Salt is the best known of the dietary factors affecting blood pressure. It has been hypothesised for the longest time, first by Ambard and Beaujard in 1904. Then in 1922 Allen first documented reduction of blood pressure by sodium restriction.

In the Intersalt study,[3] 10 000 people were examined by standardised methods in 52 different communities in 30 countries around the world. The rise in blood pressure with age was significantly related to 24-hour urinary sodium.

Within communities correlation between individuals' blood pressure and sodium intake (or excretion) is difficult to see. This is partly because of large day-to-day swings in people's sodium intake,[4] partly because people should only be compared in the same age group, and also because not all individuals are sensitive to salt—this can be demonstrated by a week of 12 g NaCl, followed by a very low salt diet.[5] However, in the dietary and nutritional survey of British adults blood pressure was found to correlate with 24-hour urinary sodium, reflecting salt intake.[6] A cohort study in 2436 Finnish men and women found that those who started with high 24-hour urine sodiums had more cardiovascular and total mortality over the following 8 years.[7]

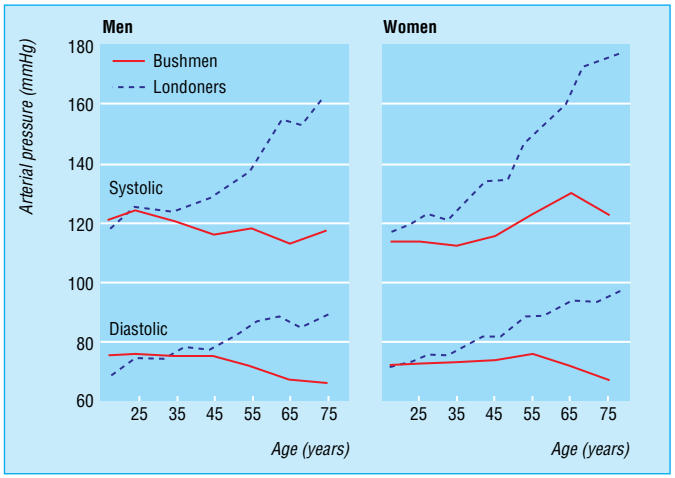
The requirement for sodium in health is usually under 25 mmol Na/day (equivalent to 1.5 g NaCl).[9] Normal kidneys can shut down sodium excretion almost to zero and sweat loss is reduced in people on low salt intakes or adapted to hot climates. Human milk contains only 7 mmol Na/litre, so young infants' sodium intake per megajoule is only about one-sixth that of their parents'!

Salt intakes in Britain are around 9 g NaCl (150 mmol Na) per day and in parts of Asia considerably higher, over 250 mmol Na/day. To prove that our unnecessarily high intakes of salt contribute to the development of essential hypertension, blood pressures of a group of adults eating only their sodium requirement (25 mmol Na/day) would have to be compared over many years with another group, similar in all respects, eating the usual 150 mmol sodium/day. Such a human trial is probably impossible, so a trial in chimpanzees, who have 98% the same DNA as humans and half our life span, is important.

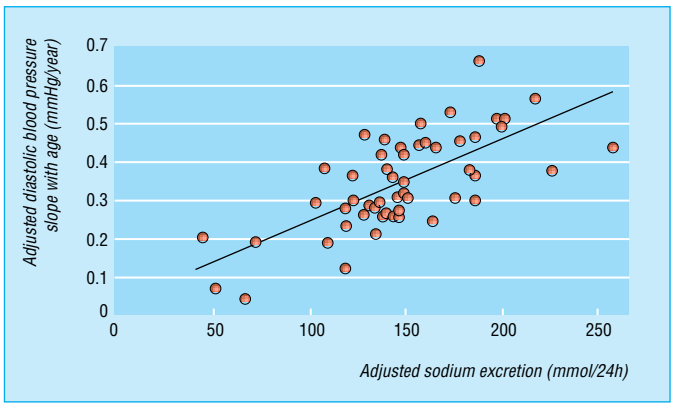
Causal factors in essential hypertension

- **Genetic**—several mechanisms
- **Tension from anxiety**—via increased sympathetic tone or circulating catecholamines
- **Dietary—positive correlation—negative correlation**

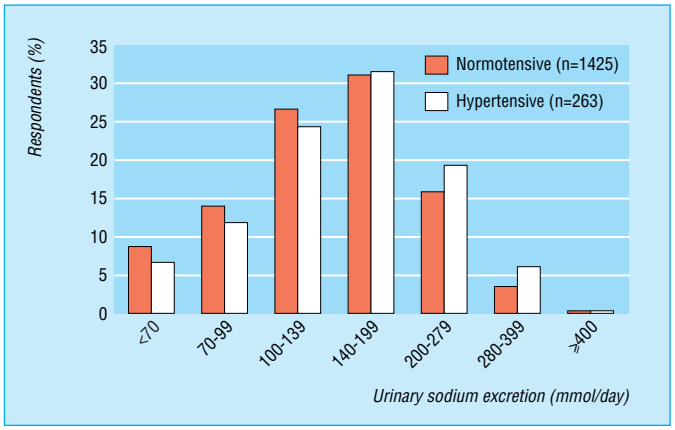
Overweight and obesity	Potassium
Sodium (salt) intake	?Calcium
Alcohol	



Blood pressure with age of 152 Bushmen, hunter gatherers (aged 15-83 years) in NW Botswana (continuous lines) compared with standard figures from London measured in 1954. (Adapted from Truswell *et al.*[2])



Cross-centre plots of diastolic blood pressure slope with age and median sodium excretion; $P < 0.001$. (Adapted from Intersalt study[3]). For an additional 100 mmol Na/day, the increase of BP over 30 years (25 to 55) was 10 systolic/6 diastolic mmHg greater[8]



Distribution of normotensive and hypertensive respondents by urinary sodium excretion rate in the Dietary and Nutritional Survey of British Adults, 1986-87. (Adapted from Beard *et al.*[6])

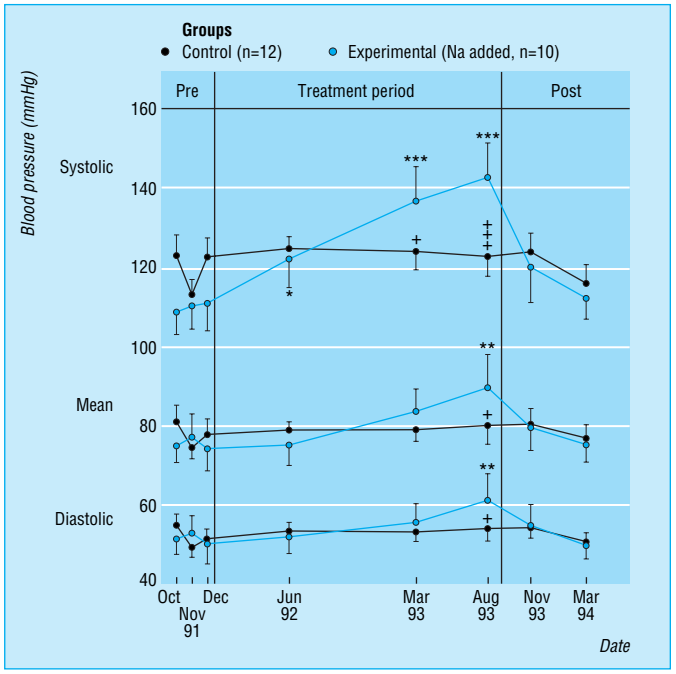
Chimpanzees, living on a natural (low sodium) vegetarian and fruit diet in Gabon (West Africa) were given a liquid infant formula with or without salt up to 15 g/day for 1.5 years.[10] Blood pressures rose progressively in eight of the 10 animals given typical human salt intakes and in none of the controls.

Salt has been used since Neolithic times by most cultures as an important food preservative. Most of mankind has become used to the taste of more salt than we need now that canning, freezing, refrigeration, etc, are widely used to preserve our food.

For the general adult population, mainly as a measure to help prevent hypertension, Australia (from 1982), the USA (from 1989), WHO (1990) and the UK Department of Health (1994)[11] all recommend a target intake of 100 mmol sodium per day equivalent to 6.0 g NaCl or 2.3 g Na *or less.*

Sodium accumulation and arterioles

The mechanism of action of sodium is undoubtedly complex and involves kidney tubules and several hormones. One aspect is that if sodium tends to accumulate in cells it interferes with calcium transport, and elevated free calcium in the cytosol of arteriolar smooth muscle cells increases their tone and consequently the arterial blood pressure.

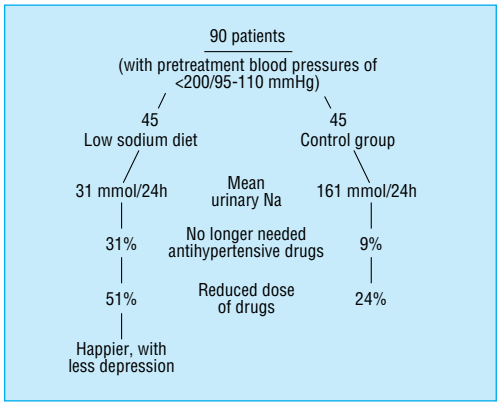


Mean systolic and diastolic blood pressures of 10 salt-added (experimental) and 12 control chimpanzees over 2.5 years. Blood pressure rose in most of the experimental chimpanzees. It returned to normal when the salt was discontinued (post treatment). (Adapted from Denton *et al.*[10])

In people with hypertension, how much reduction of blood pressure can be achieved with a low salt diet and how difficult is this to organise (and persist with)?

Elevated blood pressure can usually be lowered by salt restriction.[12] Diuretic drugs work by increasing urinary sodium excretion. Alternatively a sufficient reduction of dietary sodium can achieve the same degree of negative sodium balance. In mild to moderate hypertension, a reduction of sodium intake (which can be monitored with 24-hour urinary sodium) by 50 mmol/day will usually give a useful reduction of blood pressure, so that the patient may be able to come off the hypotensive drugs (or not start them) or reduce the dose (and with this the probability of side effects). Salt restriction increases sensitivity to all hypertensive drugs except slow channel calcium blockers, like nifedipine. Some people are more responsive than others. Older people may be more responsive to salt reduction and they are particularly susceptible to the side effects of drugs.

When people change to a lower salt diet their taste adjusts after a few weeks. Other flavours are perceived and appreciated more. The major obstacle to eating low salt is that most of the salt in food is put in during processing and is outside the individual's control.



Trial of low-salt diet in people with mild to moderate hypertension[13]

Sodium in foods

Most of the salt that we eat is not that added at the table or in cooking water (much of which goes down the sink). It is salt added in food processing, particularly of staple foods. Wheat flour contains 3 or 4 mg sodium/100 g but average breads have 520-550 mg/100 g. Oils like sunflower or olive oil contain only traces of sodium but butter averages 750 mg/100 g and margarines 800 mg/100 g. Many cereal products—biscuits, cakes and breakfast cereals (though not all)—are very high in sodium, which consumers cannot taste (being masked by the sugar content). Salted peanuts contain less sodium than breads; consumers can taste the salt because it is all on the surface. Anyone wanting to reduce salt intake must find low-salt breads and breakfast cereals and cheeses as well as cutting out the more obvious bacon and olives in brine which people eat less

Average percentages of sodium from different sources[14]

• **Discretionary**	
Added at table	9.0
Used in cooking	6.0
• **Food**	
Naturally occurring	18.5
Added salt in processing	58.7
Non-salt additives	7.2
• **Salt in water supply** (average)	0.6
	100.0

often. Other sodium compounds in food, bicarbonate and glutamate, have less effect on blood pressure than sodium chloride.

Body weight

Obese people are likely to have a higher blood pressure than lean people. In a cohort of over 5000 people born in Britain in the same week, blood pressures at the age of 36 were progressively higher in those with a body mass index (weight (kg)/height (m^2)) above 26. Typically a 3 mmHg higher diastolic pressure may be expected for every 10 kg increase in body weight.[3] In a large Swedish study of 60-year-old men, a quarter of the fattest fifth were taking antihypertensive drugs compared with only 4% of the thinnest fifth.[16] Raised blood pressure and hyperlipidaemia are both major risk factors for cardiovascular disease, and effective weight reduction will improve both.

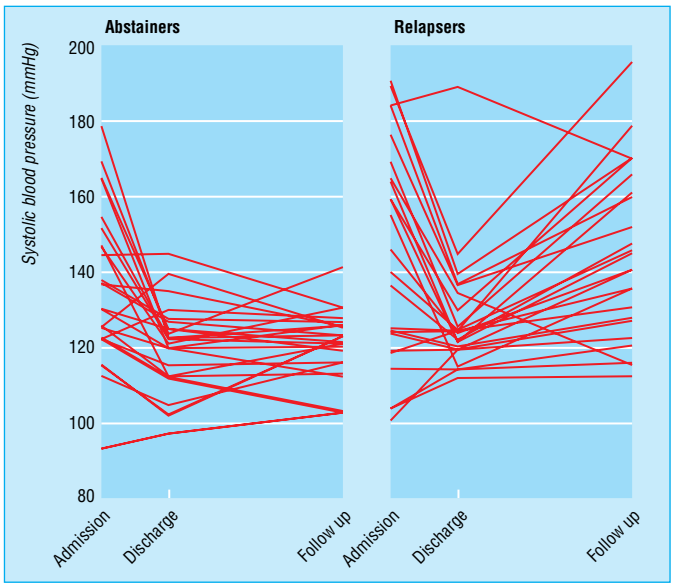
Reduced food energy and falling blood pressure

- People who do not eat enough food energy and lose weight usually have a fall of (normal) blood pressure.
- If hypertensive obese patients reduce their weight they show falls of blood pressure like 10 mmHg systolic/5 mmHg diastolic for a 5-kg weight loss.
- Less food means less sodium eaten. Some weight loss occurs even if sodium intake is maintained, but the combination of weight loss and a low sodium intake is more effective.
- In a randomised placebo-controlled trial of first-line treatment of mild hypertension in overweight patients, the weight reduction group (mean loss 7.4 kg) had a 13 mmHg fall of systolic blood pressure while those treated with metoprolol *(200 mg/day) had a 10 mmHg fall. Plasma lipids improved in the weight reduction group, but changed adversely in those on drug therapy.*[15]

Alcohol

Alcohol intake is emerging as one of the important environmental factors associated with raised blood pressure. Heavy drinkers have higher blood pressure than light drinkers and abstainers. The effect starts above about three (stated) drinks a day. Systolic pressure is more affected than diastolic.

The pressor effect of alcohol can be demonstrated directly. It was seen, for example, in men with essential hypertension who were moderate to heavy drinkers. They continued their habitual intake of beer and antihypertensive drugs; when low alcohol beer (0.9% alcohol) was substituted for the same intake of regular beer (5% alcohol), their blood pressure fell 5/3 mmHg.[18] The mechanism(s) have not yet been established. Acute ingestion of alcohol causes peripheral vasodilatation, but there are features of a hyperadrenergic state in the withdrawal syndrome. Plasma cortisol concentrations are sometimes raised in alcoholics. Increased red cell volume, and hence increased blood viscosity, is a possible mechanism.



When alcoholic patients were admitted to hospital for detoxification their systolic blood pressure fell by about 20 mmHg; it stayed down if they continued to abstain, but rose again if drinking was resumed. The same pattern was seen with diastolic pressure[17]

Components in the diet that may lower blood pressure

Potassium

In a placebo-controlled, crossover trial in mild to moderate hypertension, blood pressure fell by (average) 7/4 mmHg with a supplement of eight Slow-K tablets (64 mmol potassium) a day. But the same (London) clinic found little or no effect in similar hypertensive patients who had managed to reduce their sodium intake (and urinary sodium) to around 70 mmol a day—potassium acts as a sodium antagonist and has little effect when sodium intake has been halved.[19]

Calcium

Analyses of a diet and health study in the USA suggested that people with low calcium intakes had more hypertension, and in Britain less cardiovascular disease is reported in areas with hard water (which contains more calcium). Over 30 controlled trials with calcium supplements have been summarised in three meta-analyses,[21] which showed that 1000 mg per day or more of calcium has only a trivial effect on systolic (not diastolic) blood pressure. Increased calcium, by diet or supplements, might be useful in a very small number of hypertensive patients who have a low calcium intake or increased plasma parathyroid levels.

Potassium in foods

- **Moderate to high (mmol per usual serving):**
  Potatoes (12-26), pulses (19), dried fruits (5-12), fresh meat and fish (8-10), All Bran (8), fresh fruit (2-10), vegetables (2-10), orange juice (6), oatmeal (5), cows' milk (5), nuts (2-6), wine (3-4), beer (3), coffee (2).
- **Low (1-3 mmol per usual serving):**
  Rice, chocolate, egg, biscuits, bread, cheese, flour, cornflakes.
- **Very low or absent:**
  Sugar, jam, honey, butter, margarine, cream, oils, spirits.

Potassium is the major intracellular cation; the more concentrated the cells in a food, the higher the potassium is likely to be.

In Britain potassium intakes are around 70 mmol/day in men and 60 mmol/day in women: 17% from potatoes (10% from fried potatoes), 14% from cereals, 14% from milk products, 13% from meat and products, 11% from other vegetables, 5% from fruit and 16% from beverages (coffee 6%, tea 4%, beer 3%, fruit 2%).[20]

Magnesium

Magnesium can sometimes lower blood pressure. In patients who had received long term diuretics (mostly for hypertension) and potassium supplements, half were also given magnesium aspartate hydrochloride for six months. Their blood pressure fell significantly. The diuretics had presumably led to subclinical magnesium depletion.

Vegetarianism

Healthy (normotensive) hospital staff in Perth, Western Australia, were provided with all their meals as one of two diets—mixed omnivore or (lacto-ovo) vegetarian. Sodium intakes were kept the same. After six weeks the subjects were changed to the other diet. Blood pressures were significantly lower by about 6/3.5 mmHg while on the vegetarian diet.[23] The responsible ingredient(s) have not been clearly demonstrated.

DASH 1 and 2

*D*ietary *A*pproaches to *S*top *H*ypertension was a multicentre randomised controlled dietary trial in over 400 middle aged US adults with BP in the normal or mildly elevated range. In DASH 1 three diets were compared for eight weeks.[24] Blood pressures were lower with extra fruits and vegetables than on control diet and lower still with a combination of low fat dairy food and low saturated fat with the extra fruits and vegetables (cf control diet): −7.2/−2.8 on extra fruits plus vegetables and −11.4/−5.5 mmHg on the combination diet. Sodium and alcohol intakes and body mass index were held the same between groups.

In DASH 2 BPs were compared on control diet or DASH combination (extra fruits and vegetables, low fat dairy) at three different levels of salt intake (for one month in each subject in random order).[25] As before BPs were lower on the DASH combination diet. Reduction from usual Na (143 mmol/day) to intermediate (105 mmol/day) they averaged 2.1 and 1.3 mm systolic on control and DASH diets. Between intermediate and low sodium (65 mmol/day) systolic BPs were −4.6 and −1.7 mmHg respectively. Black people with mild hypertension showed the largest falls of BP.

Magnesium distribution in foods

***Magnesium* is distributed in foods somewhat similarly to potassium. Bran, wholegrain cereals, and legumes are the richest sources. Most vegetables contain similar moderate amounts to meat.**

From management guidelines of the British Hypertension Society[22]

Non-pharmacological measures … should be offered to all hypertensive patients whether taking drugs or not. This advice should also be offered to people with a strong family history of hypertension. In mild hypertension non-pharmacological measures may obviate the need for drugs.

- Reduce energy intake to achieve ideal weight.
- Alcohol <21 units/week in men and <14 units per week in women. One or two days/week no alcohol.
- Reduce salt intake.
- Regular physical exercise and improve level of fitness.

And to reduce the risk of cardiovascular disease stop smoking and reduce saturated fat intake.

References

1 Page LB, Damon A, Moellering RC Jr. Antecedents of cardiovascular disease in six Solomon Islands societies. *Circulation* 1974; **49**: 1132-45.
2 Truswell AS, Kennelly BM, Hansen JDL, Lee RB. Blood pressure of Kung Bushmen in northern Botswana. *Am Heart J* 1972; **84**: 5-12.
3 Intersalt Cooperative Research Group. Intersalt: an international study of electrolyte excretion and blood pressure. Results for 24-hour urinary sodium and potassium excretion. *BMJ* 1988; **297**: 319-28.
4 Frost CD, Law MR, Wald NJ. By how much does dietary salt reduction lower blood pressure? II. Analysis of observational data within populations. *BMJ* 1991; **302**: 815-18.
5 Morimoto A, Uzu T, Fujii T *et al*. Sodium sensitivity and cardiovascular events in patients with essential hypertension. *Lancet* 1997; **350**: 1734-7.
6 Beard TC, Blizzard L, O'Brien DJ, Dwyer T. Association between blood pressure and dietary factors in the dietary and nutritional survey of British adults. *Arch Intern Med* 1997; **157**: 234-8.
7 Tuomilehto J, Jousilahti P, Rastenyte D *et al*. Urinary sodium excretion and cardiovascular mortality in Finland: a prospective study. *Lancet* 2001; **357**: 848-51.
8 Elliott P, Stamler J, Nichols R *et al*. Intersalt revisited: further analyses of 24-hour sodium excretion and blood pressure within and across populations. *BMJ* 1996; **312**: 1249-53.
9 Department of Health. *Dietary reference values for food energy and nutrients for the United Kingdom. Report of the Panel of the Committee on Medical Aspects of Food Policy*. London: HMSO, 1991, pp 152-5.
10 Denton D, Weisinger R, Mundy NI *et al*. The effect of increased salt intake on blood pressure of chimpanzees. *Nature Med* 1995; **1**: 1009-16.
11 Department of Health. *Nutritional aspects of cardiovascular disease. Report of the Cardiovascular Review Group Committee on Medical Aspects of Food Policy*. London: HMSO, 1994.
12 Law MR, Frost CD, Wald NJ. By how much does dietary salt reduction lower blood pressure? III. Analysis of data from trials of salt reduction. *BMJ* 1991; **302**: 819-24.
13 Beard TC, Cooke HM, Gray WR, Barge R. Randomised controlled trial of a no-added-sodium diet for mild hypertension. *Lancet* 1982; **ii**: 455-8.
14 Edwards DG, Kaye AE, Druce E. Sources and intakes of sodium in the United Kingdom diet. *Eur J Clin Nutr* 1989; **43**: 855-61.
15 McMahon SW, Macdonald GJ, Bernstein L, Andrews G, Blacket RB. Comparison of weight reduction with metaprolol in treatment of hypertension in young overweight patients. *Lancet* 1985; **i**: 1233-5.
16 Larsson B, Björntorp P, Tibblin G. The health consequences of moderate obesity. *Int J Obesity* 1981; **5**: 97-116.
17 Saunders JB, Beevers DG, Paton A. Alcohol-induced hypertension. *Lancet* 1981; **ii**: 653-6.
18 Puddey IB, Beilin LJ, Vandongen R. Regular alcohol use raises blood pressure in treated hypertensive subjects. *Lancet* 1987; **i**: 647-50.
19 Smith SJ, Markandu MD, Sagnella GA, MacGregor GA. Moderate potassium chloride supplementation in essential hypertension: is it additive to moderate sodium restriction? *BMJ* 1985; **290**: 110-13.
20 Ministry of Agriculture, Fisheries and Food. *The Dietary and Nutritional Survey of British Adults—Further Analysis*. London: HMSO, 1994.
21 Bucher HC, Cook RJ, Guyatt GH *et al*. Effects of dietary calcium supplementation on blood pressure. A meta-analysis of randomised controlled trials. *JAMA* 1996; **275**: 1016-22.

22 Sever P, Beevers G, Bulpitt C *et al.* Management guidelines in essential hypertension: report of the second working party of the British Hypertension Society. *BMJ* 1993; **306**: 983-7.
23 Rouse IL, Beilin LJ, Armstrong BK, Vandongen R. Blood pressure-lowering effect of a vegetarian diet: controlled trial in normotensive subjects. *Lancet* 1983; **i**: 5-10.
24 Appel LJ, Moore TJ, Obarzanek E. A clinical trial of the effects of dietary patterns on blood pressure. *N Engl J Med* 1997; **336**: 1117-24.
25 Sacks FM, Svetkey LP, Vollmer WM *et al.* Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. *N Engl J Med* 2001; **344**: 3-10.

Further reading

Scientific Advisory Committee on Nutrition. *Salt & Health.* London: Stationery Office, 2003.

3 Nutritional advice for some other chronic diseases

Dental caries

Dental caries affects people predominantly in the first 25 years of life. Dental enamel is the hardest material in the body. Its weakness is that, because it is basically calcium phosphate, it is dissolved by acid. Three factors together contribute to caries.

Infection

A specific species of viridans streptococci, *Streptococcus mutans*, metabolises sugars to lactic acid and also polymerises sugars to a layer of covering polysaccharide in which the bacteria are shielded from saliva and the tongue. Some people harbour more of these bacteria than others.

Substrate

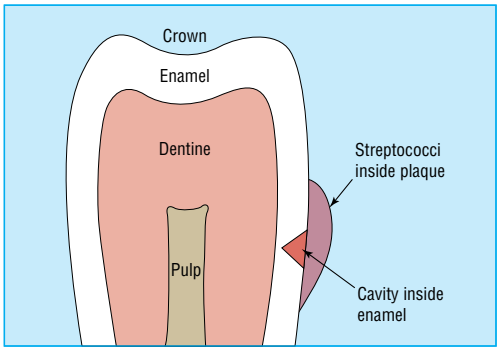
Most sugars serve as substrate—sucrose, glucose, fructose, and lactose (not sorbitol or xylitol). Starches too, if they stay in the mouth, are split to sugars by salivary amylase. Consumption of sugary foods between meals, especially if they are sticky and consumption is repeated, favours the development of caries. Brushing the teeth and flossing between them after meals reduces the likelihood of caries.

Resistance of the teeth

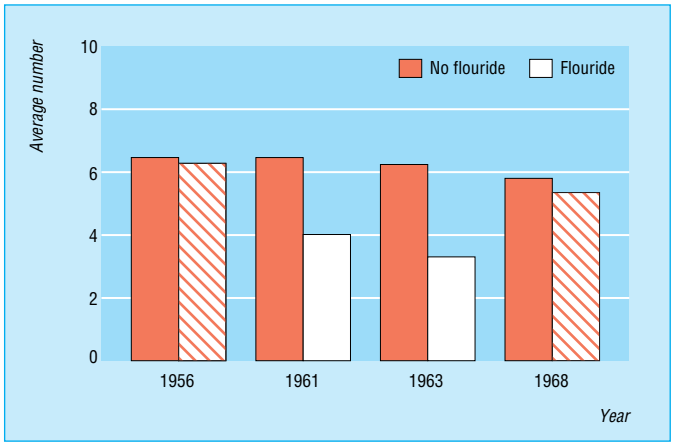
Caries is more likely in fissures. In older people the "mature" enamel is more resistant. An intake of 1-3 mg/day of fluoride—as occurs, for example, if drinking water is fluoridated at a concentration of 1 mg/l—increases the enamel's resistance, especially if taken while enamel is being laid down before the tooth erupts.

The cariostatic effect of fluoride in natural water was noticed in Maldon, Essex in 1933, and confirmed by comparing children's teeth and water fluoride across the United States in the early 1940s. Water fluoridation is widespread in the United States, Australia and New Zealand but still unusual in Scandinavia and The Netherlands. In Britain only about 10% of the population receive fluoridated water. Dental caries has nevertheless become less prevalent in most industrialised countries. Most toothpastes now contain fluoride and this, rather than any change in children's sugar consumption, seems the main reason for the decline where water is not fluoridated. A controlled study in the north of England found 44% less caries in 5 year olds in 1991-94 in towns with a fluoridated water supply.[2]

Mottling of the (anterior permanent) teeth occurs if the fluoride intake is too high in the first eight years of life. Young children should either be persuaded not to swallow their toothpaste or be provided with a "junior" product with half-strength fluoride.



Dental caries

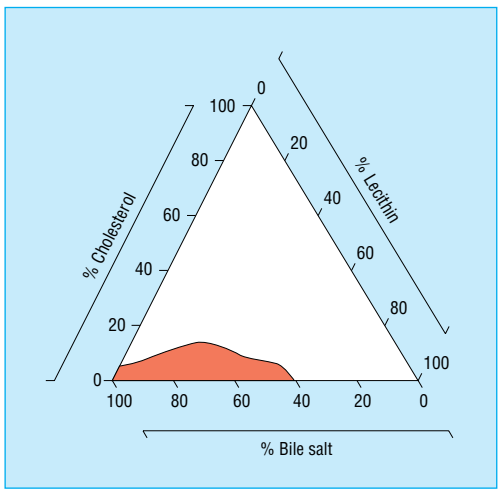


The shaded bars show what happened to the number of decayed temporary teeth in Kilmarnock after fluoridation of water, which started in 1956 and was discontinued in 1962. Unshaded bars are findings in Ayr, which never had fluoridated water.[1] Figures for children aged 5 years

Gallstones

Most gallstones are composed mainly (about 85%) of crystallised cholesterol with small proportions of calcium carbonate, palmitate, and phosphate. Cholesterol, which is excreted by the liver into the bile, would be completely insoluble in an aqueous fluid like bile if it were not kept in micelle microemulsion by the combined detergent action of the bile salts and phospholipids (chiefly lecithin) in bile.

Non-dietary risk factors include female sex, pregnancy, oral contraceptives, age, ileal disease, clofibrate therapy, and certain



Three major components of bile (bile salts, lecithin, and cholesterol) on triangular coordinates. Each component is expressed as percentage moles of total bile salt, lecithin, and cholesterol. The shaded area shows conditions required for cholesterol to be soluble in micellar form. If the concentration of cholesterol goes up or bile acids or lecithin go down then cholesterol is likely to precipitate out[3]

ethnic groups—for example, Pima Amerindians have a high incidence of gallstones.

In obesity and during dieting (with rapid weight loss) cholesterol secretion into bile tends to increase. During fasting and on total parenteral nutrition the gall bladder does not contract normally. In people on vegetarian and high cereal fibre diets the pattern of biliary bile acids change favourably, with less deoxycholate and more chenodeoxycholate.[4] Moderate alcohol intake appears to be protective; decreased cholesterol saturation of bile has been reported. Regular exercise also appears to protect against gallstones.[5] These associations do not apply to the less common pigment stones.

Gallstone formation

Gallstones are more likely to form if:
- biliary cholesterol is increased, *or*
- biliary bile acids are reduced, *or*
- the gall bladder is less motile, *or*
- factors in the bile favour nucleation of cholesterol crystals.

Urinary tract stones

Calcium stones

Dietary factors which tend to increase urinary calcium or have been associated with stones are high intakes of protein, sodium, refined carbohydrate, vitamin D, calcium (spread over the day), alcohol, curry, spicy foods, and Worcester sauce, and low intakes of cereal fibre and water. Since most patients with hypercalciuria have intestinal hyperabsorption of calcium it has been common to recommend a low calcium diet or phytic acid or a resin to reduce calcium absorption. Long term trials have been lacking. Now a diet providing usual calcium intake (1200 mg/day) but very low salt (50 mmol Na/day) and reduced animal protein (50 g/day) has reduced calcium stone recurrences significantly over five years compared with a low calcium diet (400 mg/day).[6] The normal calcium, low protein, low salt diet reduced urinary excretion of both calcium and oxalate.

Foods rich in oxalate

Spinach, rhubarb, beetroots, cocoa, chocolate, currants, dried figs, tea, swiss chard, blackberries, oranges, turnip greens.

Oxalate stones

Associated dietary factors are high intakes of oxalate or vitamin C and low water intake.

Uric acid stones

Uric acid stones are associated with an acid urine, a high purine diet, and low water consumption.

The one common dietary association with all the common types of stone—and with the rare ones also—is a low water intake. Drinking plenty of water is an important habit for anyone liable to stones, especially if the weather is hot. Last thing at night is the important time to take water.

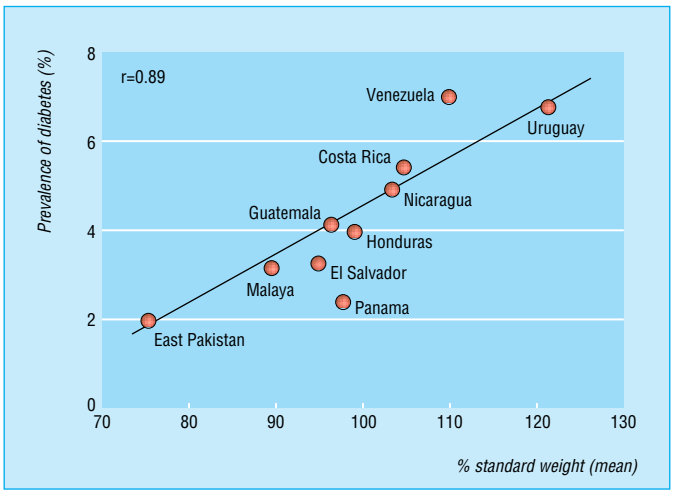
Uric acid stones

- One dietary cause of acid urine is a high protein intake. The amino acids methionine and cystine are metabolised to urinary sulphuric acid.
- Foods traditionally rich in purines include liver, kidneys, sweetbreads, sardines, anchovies, fish roes, and yeast extracts, but there are no modern tables and dietary RNA may raise plasma urate more than DNA.

Diabetes mellitus

Insulin-dependent diabetes (Type 1) is usually caused by autoimmune damage to the β-cells in the pancreatic islets, which lose their ability to secrete enough insulin. This type of diabetes typically starts in adolescents or younger adults. Several epidemiological studies have reported that patients with type 1 diabetes were less often exclusively breast fed for the first 3-4 months of life than unaffected controls.

The prevalence of non-insulin dependent diabetes (Type 2) increases with age; overall it is about six times more common than Type 1. This type 2 diabetes is closely associated with overweight or obesity and with lack of exercise. Beyond Europe and Anglo-Celtic north Americans there is almost a pandemic of type 2 diabetes occurring in some communities that may have earlier experienced undernutrition but are now sedentary and eating refined, high energy "Western" foods. The thrifty genotype hypothesis attempts to explain this phenomenon, which is especially affecting people of south Asian descent in



Relation of body weight to prevalence of diabetes (standardised criteria) between countries[7]

Britain and elsewhere, Pacific islanders, and north American and Australian aboriginals.[8]

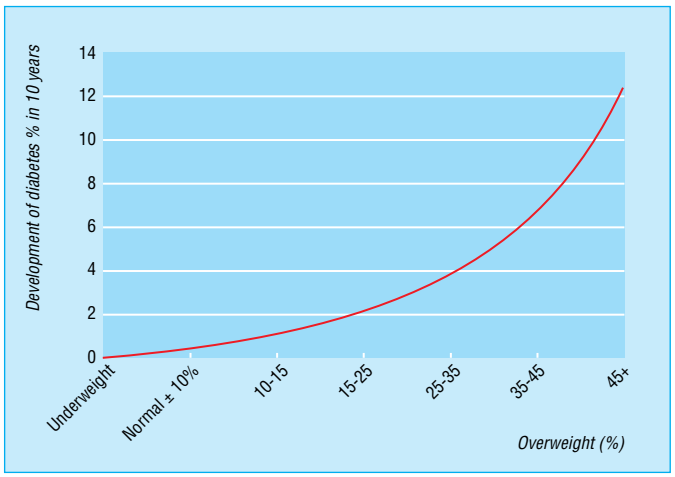
Looked at another way, diabetes is the complication of **obesity** whose incidence goes up at the steepest gradient with degree of overweight. The risk of developing diabetes is greater in people whose obesity is mainly intra-abdominal rather than on the hips or buttocks (subcutaneous)—people with a high ratio of waist:hip circumferences.[9]

Diabetes is a multifactorial disease. There is a strong family influence, though this may be partly because eating habits and body weight are influenced by family behaviour. But a **genetic** factor is clear in some groups: the Pima Amerindians in North America and Micronesians in Nauru. When these people are obese (which most of them are these days) the incidence of diabetes (in older life) is over 50%.

The popular belief that eating a lot of sugar predisposes to diabetes is not confirmed by several epidemiological and prospective studies. High fat intake is more likely to lead to diabetes, a hypothesis first put forward in Britain in 1935 by Sir Harold Himsworth. High total carbohydrate (mostly starch) and high fibre intakes are characteristic of peasant communities, in which type 2 diabetes is uncommon.

In a **prospective study** of 7735 middle-aged men, drawn from group practices in 24 towns in England, Wales, and Scotland and followed for 12 years, the incidence was 2 per 1000 person years.[11] The risk of developing diabetes increased exponentially with increasing body mass index (BMI). It was 11 times higher in the upper fifth of BMI (>28 kg/m^2). Men with moderate physical activity had less than half the risk. Moderate drinkers also developed less diabetes. On average those who developed diabetes had higher plasma triglycerides, higher blood pressures and higher casual blood glucose. Another finding in people who will later develop type 2 diabetes has been an increased fasting insulin and/or insulin response to standard glycaemic stimulus, due to insulin resistance.

Diets for managing established diabetes are discussed in chapter 13.



Relation of body weight to subsequent development of diabetes in 10 years[10]

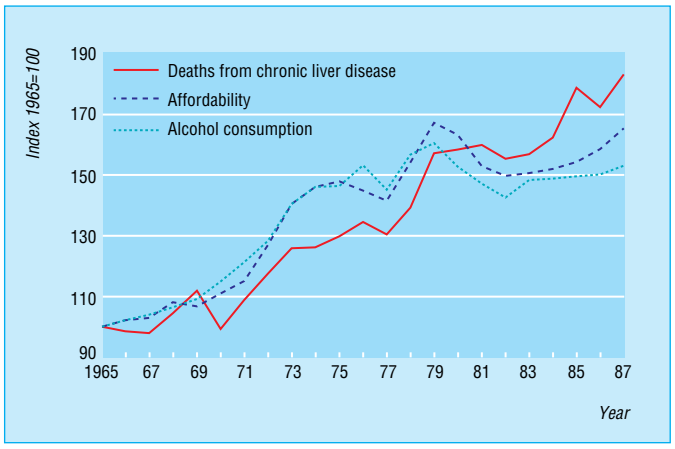
Two large *randomised controlled trials*, one in 27 centres in the United States,[12] the other in five cities in Finland[13] have both shown that lifestyle intervention can halve the incidence of type 2 diabetes in middle-aged, overweight, sedentary people with impaired glucose tolerance. In the US trial lifestyle intervention was weight reduction averaging 6 kg and increased physical activity. This was more effective in preventing diabetes than metformin. In Finland the subjects were also asked to reduce saturated fat and increase whole grain foods.

Alcoholic liver disease

Countries with high alcohol consumption per head have high mortalities from cirrhosis. These have fallen when there has been a reduction in the supply of alcohol—for example, during prohibition in the United States and during the two world wars in Europe. Correlation of alcohol consumption and deaths from cirrhosis between countries is close, but there are deviations. Britain has a lower incidence of cirrhosis than might be expected from the rate of alcohol consumption but mortality from cirrhosis has doubled here since 1970. Where alcohol consumption is high most cases of cirrhosis are due to alcohol. Other causes—for example, viral hepatitis B or C, account for important proportions of cases.

In heavy drinkers pre-cirrhotic liver disease—fatty liver or alcoholic hepatitis—is more common than cirrhosis. A fourth condition, primary liver cell cancer, is a complication of alcoholic cirrhosis.

Within countries the risk of developing cirrhosis is related to the dose and duration of alcohol intake. Daily heavy drinking for years is the typical pattern—80 g (eight drinks) a day in men, and usually well over this. In a large Italian study, cirrhosis appeared to be less likely in those who drank only with meals.[15]



Alcohol affordability and consumption and deaths from cirrhosis in the United Kingdom, 1965-87. As alcohol became more affordable, consumption of alcohol and harm related to alcohol increased[14]

The essential treatment of alcoholic liver disease is complete and permanent abstinence from alcohol. Although alcoholics may become deficient in nutrients, those who develop cirrhosis are often socially organised and well nourished. There is no evidence that a high protein diet or choline can prevent alcoholic cirrhosis in man. Even when cirrhosis is established, an improved clinical state and prognosis may be expected in those who manage to abstain completely.

No precise safe level of alcohol intake can be given—only a clinical impression—because people who drink heavily underestimate their consumption when asked about it, and no prospective epidemiological study has been done. Women are more susceptible to hepatic damage from alcohol because they have smaller livers (where most metabolism of alcohol occurs) and also lower rates of gastric (first pass) oxidation of alcohol than men.[16] Only a minority of heavy drinkers get cirrhosis; there is presumably a synergy between alcohol and hepatitis viruses.

Some types of cancer[17]

Differences in diets are thought to account for more variation in the incidence of all cancers than any other factor (with smoking in second order).[18] The big questions are which dietary components are active, and how do they work? Our bodies have three routes of entry for foreign compounds: the skin, lungs, and intestines. As a function of surface area the chances of absorption are skin 1, lungs 1000, and intestines 1 000 000. There are countless natural non-nutrient substances in foods and several are mutagens. The fact that they can induce mutations in a standard bacterial culture does not, however, establish that they are dangerous to man: there are many available protective mechanisms.

Poor diet may have a more decisive effect by weakening defence mechanisms than by supplying potent carcinogens. Epidemiologists estimate that synthetic chemical additives in food account for under 1% of all cancers.[18] The cancers most clearly related to habitual diet are oesophageal, gastric, and large intestinal cancers.

Oesophagus

In the Chinese focus of oesophageal cancer, nitrosamines have been found in mouldy food and there is a deficiency of molybdenum. Domestic fowl are affected too. In the Iranian focus there are some vitamin deficiencies and people may take opium by mouth. In the Transkei researchers think that fusarium mycotoxins, together with deficiencies of niacin, zinc, and other micronutrients, are responsible for the epidemic of oesophageal cancer. In Europe alcohol, especially that derived from apples, and tobacco are associated factors.

Oesophageal cancer

- 300× range in incidence
- Highest rates: Linxian, People's Republic of China; East Mazandaran, Iran; and Transkei, South Africa.
- In Europe there are moderately high rates in NW France and in Switzerland.

Stomach

From present epidemiological data protective factors are fruits and vegetables, refrigeration of foods and vitamin C intake. Apparent causative factors are intake of salt, pickled and salted foods, *Helicobacter pylori* infection, and smoking.

Gastric cancer

- Incidence in Britain has spontaneously fallen to half in the past 25 years.
- There have been similar reductions in many developed countries.[19]
- Highest rates, in Japan, are three times those in England and Wales, 10 times those in the United States of America, 20 times those in countries with the lowest rate, for example Gambia and Kuwait.
- Chronic atrophic gastritis is a precancerous state.

Large intestine

Cancer of the large intestine usually arises in a polyp. Different dietary factors may be involved in the successive stages: formation of polyps; malignant transformation; growth and spread of a cancer. Having a halfway stage of polyps should make study of causative factors easier. In some epidemiological studies animal fat and meat have emerged as risk factors. But in the majority of epidemiological studies meat has not been significantly associated.[20] One possible mechanism is the formation of heterocyclic amines (1Q, MelQ, PhlP, etc.), which are potent mutagens, on the surface of well-cooked meat. Some types of beer have been associated with rectal cancer. Wheat fibre appears the best established protective factor. It dilutes and moves on potential carcinogens in the lumen and promotes fermentation. Brassicas and other vegetables also appear protective; they contain several anticancer substances and also folate, which may prevent hypomethylation of DNA, a characteristic change in this cancer. In a trial wheat bran plus low fat prevented polyp development. β-Carotene or vitamin E (α-tocopherol) have been ineffective; other prevention trials are underway.

Cancer of the large bowel

- Fourth largest cause of death from cancer in Britain (after lung cancer, breast cancer in women, and prostate cancer in men).
- Ten times more common in developed Western countries such as Britain and USA than in the Third World.
- Rates in Scotland have been among the highest in world
- Epidemiology of rectal cancer shows some minor differences from the larger group of colon cancer.
- Left side of the large bowel is usually affected.

Breast

Between-country comparisons and animal experiments suggest that high fat intake increases the risk of breast cancer but prospective and case-control epidemiological studies have not confirmed a role for fat, unless it operates in childhood or adolescence. Weight gain in adult life increases the risk of postmenopausal breast cancer. Adipose tissue is a major source of oestrogen after the menopause. Alcohol consumption also shows some association but this is not dose related. Plant foods appear protective. The two most promising of these are wheat fibre (which can bind oestrogens in the bowel, reducing reabsorption) and soya (which contains phytoestrogens, isoflavones).

Breast cancer

In countries with a high incidence the majority of cases are postmenopausal. Incidences are four times higher in western Europe and North America than in East Asia. Early menarche and/or late menopause increase the risk; bilateral oophorectomy protects, and endogenous plasma oestrogens are higher in patients with postmenopausal breast cancer.

References

1 Department of Health & Social Security. *The fluoridation studies in the United Kingdom and the results achieved after eleven years.* London: HMSO, 1969 (Reports on Public Health & Medical Subjects no 122).
2 Jones CM, Taylor GO, Whittle JG, Evans D, Trotter DP. Water fluoridation, tooth decay in 5 year olds, and social deprivation measured by the Jarman score: analysis of data from British dental surveys. *BMJ* 1997; **315**: 514-17.
3 Small DM. Gallstones. *N Engl J Med* 1968; **279**: 588-93.
4 Low-Beer TS. How the colon begets gallstones. *Lancet* 1998; **351**: 612-13.
5 Vega KJ, Johnston DE. Exercise and the gallbladder. *N Engl J Med* 1999; **341**: 836-7.
6 Borghi L, Schianchi T, Meschi T *et al.* Comparison of two diets for the prevention of recurrent stones in idiopathic hypercalciuria. *N Engl J Med* 2002; **346**: 77-84.
7 West KM, Kalbfleisch JM. Influence of nutritional factors on prevalence of diabetes. *Diabetes* 1971; **20**: 99-108.
8 Report of a WHO Study Group. *Prevention of diabetes mellitus.* WHO Tech Rep Ser 844. Geneva: WHO, 1994.
9 Ohlson LO, Larsson B, Svarsudd K *et al.* The influence of body fat distribution on the incidence of diabetes mellitus. 13.5 years of follow up of the participants in the study of men born in 1913. *Diabetes* 1985; **34**: 1055-8.
10 Westlund K, Nicolayson R. Ten year mortality and morbidity related to serum cholesterol. *Scand J Clin Lab Invest* 1972; **30**: 3-24.
11 Perry IJ, Wannamethee SG, Walker MK, Thomson AG, Whincup PH, Shaper AG. Prospective study of risk factors for development of non-insulin dependent diabetes in middle aged British men. *BMJ* 1995; **310**: 560-4.
12 Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002; **346**: 393-403.
13 Tuomilehto J, Lindström J, Eriksson JG *et al.* Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* 2001; **344**: 1343-50.
14 Department of Health. *On the state of the public health for the year 1988.* London: HMSO, 1989, fig 2.1.
15 Bellantani S, Saccoccio G, Costa G *et al.* Drinking habits as cofactors of risk for alcohol induced liver damage. *Gut* 1997; **41**: 845-50.
16 Frezza M, di Padova C, Pozzato G. High blood alcohol levels in women. The role of decreased alcohol dehydrogenase activity and first pass metabolism. *N Engl J Med* 1990; **322**: 95-9.
17 Department of Health. Nutritional aspects of the development of cancer. *Report of the Working Group on Diet and Cancer of the Committee on Medical Aspects of Food and Nutrition Policy.* The Stationery Office, 1998.
18 Doll R, Peto R. The causes of cancer: qualitative estimates of avoidable risks of cancer in the United States today. *J Natl Cancer Inst* 1981; **66**: 1191-308.
19 Howson CD, Jiyama T, Wynder EL. The decline in gastric cancer: epidemiology of an unplanned triumph. *Epidemiol Rev* 1986; **8**: 1-27.
20 Truswell AS. Meat consumption and cancer of the large bowel. *Eur J Clin Nutr* 2002; **56** suppl 1: 19-24S.

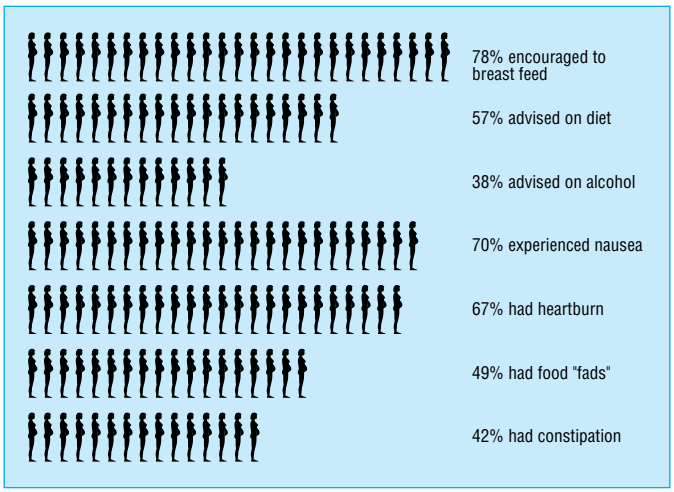
4 Nutrition for pregnancy

Pregnancy is a time when appetite is altered and nutritional needs change. What the expectant mother eats or drinks can affect her baby's health and her own comfort. In pregnancy women develop a new interest in the consequences for health of what they eat. They are entitled to advice from their doctors.

The first advice should ideally be communicated before pregnancy, when a woman decides to try to have a baby. Pregnancies in women who are overweight, have anorexia nervosa, or whose growth is not completed are more difficult, and these women need extra nutritional care.

A good intake of folate is important in preventing neural tube defects and some other malformations in the fetus of a minority of women. The stage when this vitamin is most needed is the first 28 days after conception so supplementation or high folate diet has to be periconceptional. The supplement dose is 400 or 500 μg/day. Likewise, it is the early weeks when excess alcohol intake may lead to malformations.

During pregnancy extra nutrients are required, especially from 20 weeks, for the growing fetus and for the placenta. Tissue is also laid down in the uterus and breasts, blood volume is increased, and, in healthy women with adequate food, adipose tissue increases by around 2.7 kg. This fat is deposited more on the hips and thighs.



BBC television survey of 6000 women, 1982[1]

Folate and neural tube defects

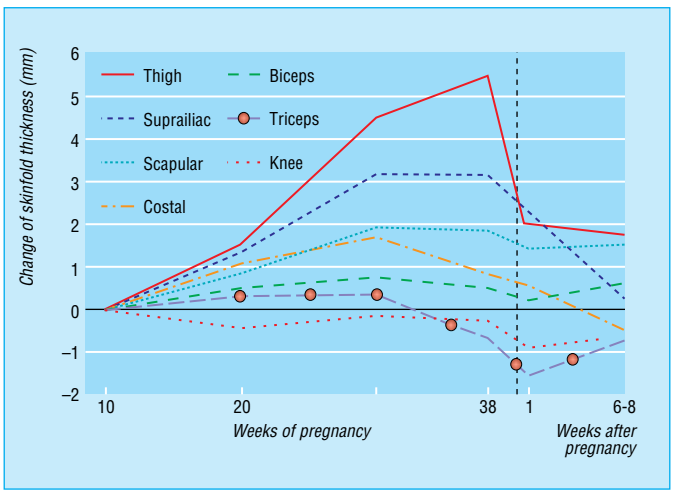
- Folate is the most important nutrient for replication of DNA in cell division. Evidence for the role of folic acid in preventing neural tube defects (NTDs) has been accumulating for 50 years. The folate antagonist aminopterin, taken in pregnancy, led to NTDs.
- Lower biochemical folate levels in women who gave birth to babies with NTD were reported in 1975 and 1976. The first secondary prevention trials (reported in 1980) were encouraging but not randomised. So the MRC conducted a large randomised double-blind trial in seven countries and found that folic acid could prevent three-quarters of recurrences.[2] Other epidemiological studies are supportive and so is a primary prevention trial in Hungary.[3] Evidently at the time of closure of the neural tube there is extra demand for folate for cell division and in some pregnancies on ordinary diets the level of folate at the site is inadequate.

Alcohol in pregnancy[4]

- Heavy drinkers have a greatly increased risk of inducing the fetal alcohol syndrome—characteristic underdevelopment of the mid face, small size, and mental retardation.
- Women who intend to become pregnant should not sit drinking whatever the occasion: they could be two or three weeks pregnant.
- Once pregnancy is established the rule should be no more than one alcoholic drink a day to be sure of preventing minor effects, chiefly growth retardation.

Nutrition for pregnancy

The extra energy need for a pregnancy can be calculated as about 250 MJ (60 000 kcal).[6] This includes energy stored in fetal fat and protein, and in maternal reproductive tissues and adipose tissue. It takes account of the mother's increased basal metabolic rate and the energy needed to move a heavier body. This corresponds to 1 MJ (240 kcal) a day (excluding the first month, for 250 days), and in Britain the recommended daily intake of energy during pregnancy (10 MJ, 2400 kcal) until 1991 was 1 MJ (240 kcal) above the non-pregnant amount (9 MJ or 2150 kcal). When actual food intakes are carefully measured, however, little indication exists of extra energy intake in Western women. This was found in careful intake measurements in London, Cambridge, Aberdeen, Glasgow,[7] Wageningen, and Sydney. In all these centres women ate an average of almost 9 MJ (2150 kcal) per day. The extra energy need is probably balanced by decreased exercise and increased



Changes in skinfold thickness at different sites during pregnancy[5]

efficiency of metabolism. Pregnant women seem to reduce their exercise if they can. Postprandial cholecystokinin concentrations increase, which enhances nutrient absorption and the anabolic actions of insulin.[8] So it is not true that a pregnant woman has to eat calories for two, but a few nutrients should be substantially increased. In 1991 the Department of Health revised the estimated average extra requirement of energy in pregnancy to 0.8 MJ (200 kcal) a day and this is only for the third trimester.[9] However, in a developing country like rural Thailand,[6] where pre-pregnant food intake may be marginally adequate and women are involved in agricultural labour, food intake may—and should—increase in pregnancy.

The amounts of different nutrients which the mother has to put into her fetus by the time of delivery have been worked out by chemical analysis of stillbirths. These can be estimated more accurately for stable inorganic elements than for the vitamins. From these figures for nutrients accumulated and from information on whether there is any change in their absorption and turnover, the extra requirements for pregnancy can be estimated.

The metabolism of protein is more efficient and so is the absorption of iron in pregnancy. For most nutrients like **protein** the small extra amounts required are covered adequately by a normal diet. But intakes are more critical for the other five nutrients in the table showing recommended daily intakes.

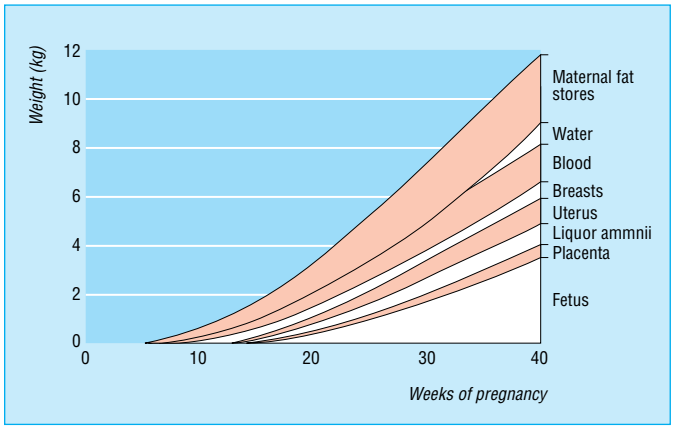
Recommended daily intakes* for six critical nutrients in pregnancy[10]

	Addition for pregnancy	Non-pregnant women	Total
Protein (g)	+10	50	60
Folate (μg total folate)	+220	180	400
Calcium (mg)	+400	800	1200
Iron (mg)	+15	15	30
Zinc (mg)	+3	12	15
Iodine (μg)	+25	150	175

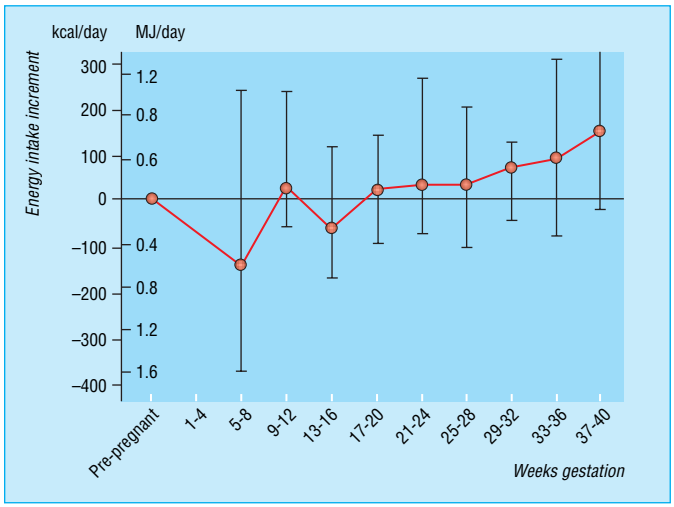
*United States recommended dietary allowances, 1989

Folate is the only vitamin, and iron the only nutrient element whose requirements double in pregnancy. Extra folate is needed for the first month and again for the last trimester. Serum and red cell folate concentrations decline in pregnancy and, if looked for, some degree of megaloblastic change can be found in substantial minorities of women in late pregnancy. Such changes have been reported in 6-25% of women not taking supplements in Britain. The word folate comes from the Latin *folia* (leaf) because it was first found in spinach, but food sources are not the same as for vitamin C. Whole grain cereals, nuts, and legumes are good sources of folate. The folate content of vegetables varies from about 10 μg/100 g in potatoes and carrots up to 155 μg in asparagus, averaging round 50 μg per 100 g. Fruits average round 5 μg per 100, citrus and blackberries higher. The vitamin is largely destroyed by prolonged boiling.

The **iron** content of the fetus (about 300 mg), placenta (50 mg), and average postpartum blood loss (200 mg) add up to some 550 mg. The red cell mass also increases after 12 weeks by an amount which corresponds to about another 500 mg of iron, but this is a temporary internal borrowing from stores and causes no extra demand provided the stores are sufficient. Against these extra needs there is the saving from no menstruation (some 200 mg) and improved intestinal absorption. Maternal haemoglobin concentration declines by about 10% because of physiological haemodilution; and serum



Contributions to weight gain in average pregnancy



Energy intake increments (and confidence limits) for 71 Glasgow women throughout pregnancy[7]

Contribution of food groups to total folate content per head, Great Britain 1998[3]

- Vegetables 32%, of which potatoes 10%
- Cereals 32%, of which bread 11%*, breakfast cereals 14%*
- Milk and products 10%
- Fruit 7%
- Meat 5%
- Tea 4%
- Eggs 2%

*A significant proportion of the folate from cereals comes from fortification. In general the folic acid used for food fortification is more biologically available than naturally occurring folate.

Iron in pregnancy

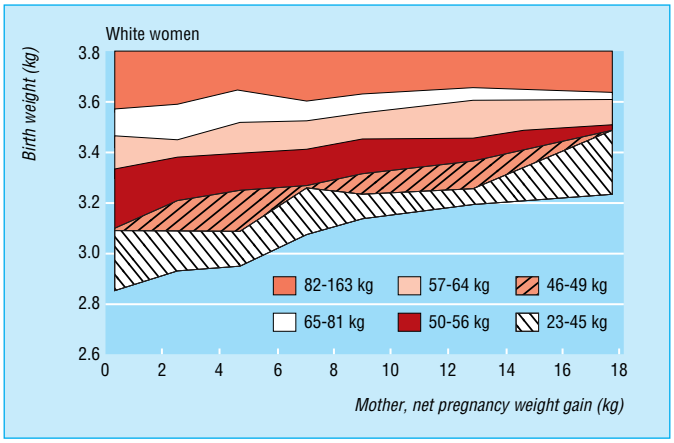
There is no universal policy. Some doctors are more interventionist than others. Iron tablets can cause indigestion or constipation. The following is generally agreed.

- Women should be advised to eat meat regularly (unless vegetarian). This is the best absorbed source of iron in the diet.
- A woman with a history of anaemia, menorrhagia, poor diet, or repeated pregnancies should be given iron supplements or an iron-folate preparation.
- Haemoglobin should be checked and iron given if it is below 110 g/l (with a low mean cell volume).
- For prophylactic purposes one iron tablet a day is adequate.
- With the smaller dose of iron, side effects are fewer and compliance should be better.

iron concentration, transferrin saturation, and ferritin concentration all go down. These changes can be partly—but only partly—prevented by iron supplementation.

With **calcium**, absorption becomes more efficient.[11] Without any change of vitamin D intake or exposure to the sun, plasma concentrations of calcitriol (the active form of the vitamin converted in the kidney) are increased. Some of this extra conversion takes place in the placenta. The easiest way of obtaining the extra calcium needed for pregnancy and lactation is from milk; 0.5 litre supplies about 600 mg calcium.

The increased need for **iodine** may be taken for granted in Britain, but in areas where goitre is endemic (see chapter 8) there is a risk of cretinism. In such areas expectant mothers should be given an injection of iodised oil, preferably before conception.



Birth weight is related positively to amount of maternal weight gain and to pre-pregnancy weight of the mother[13]

Weight gain

The amount of weight gained from before conception to shortly before delivery ranges considerably in normal women—from about 6 to 24 kg. A good average to try to achieve is 12 kg (26 lb). This might be made up of about 115 g (1/4 lb)/week for the first 10 weeks and 300 g (2/3 lb)/week for the remaining 30 weeks. A mother's height, her weight for height at the start of pregnancy, and her weight gain can all influence the size of the fetus. Birth weights are lower in babies of mothers who choose (against medical advice) to continue to smoke during pregnancy. In affluent countries the body fat gained during pregnancy can persist after childbirth. Pregnancy is one of the factors that can predispose to obesity.[12]

In Third World countries, where mothers often start small and thin and gain little weight because of restricted and bulky food, and heavy physical work, birth weights are lower than in affluent communities. They have been increased, in controlled trials, by providing food (energy) supplements during pregnancy. Average gains of birth weight in eight different trials have been from 40 to 300 g.

Obesity in pregnancy increases the chances of a heavier and fatter baby and also of hypertension and gestational diabetes. Since 3 to 4 kg of the usual 12 kg weight gain is fat, overweight women should try to put on only 7 to 8 kg overall during their pregnancy.

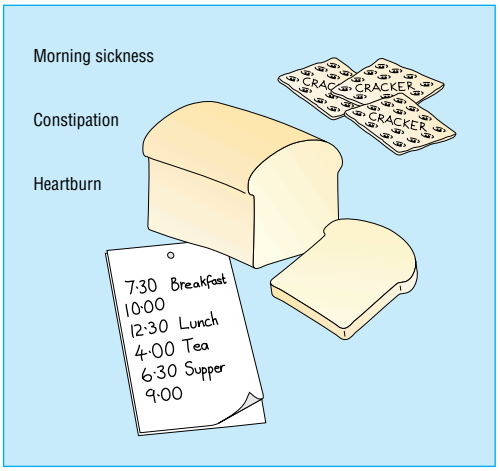
The Barker hypothesis

- Professor David Barker (Southampton) found very good birth weight records in Hertfordshire for 1911-30 and was able to trace health records of most of this cohort in later life. Coronary heart disease (CHD) mortality was higher in those who had had low birth weights (at term).[15]
- The large US Nurses prospective study provides good supportive evidence (birth weights here were by recall).[16] Low birth weight, reflecting subnormal intrauterine growth, can only influence CHD incidence if there are risk factors for CHD in adult life (chapter 1). Low birth weights have also been reported to be followed in middle age by hypertension and type 2 diabetes.
- All these discoveries emphasise the importance of good nutrition in young women before and during pregnancy. But this does **not** mean that women who enter pregnancy with a body mass index over 26 need to "eat for two"![17]
- In Britain little or no relation has been found between nutrient intakes in pregnancy and birth weight.[18]

Hypertension and "toxaemia"

In pregnancy-induced hypertension (toxaemia) no excess of sodium is retained. It is proportional to the fluid retained. No evidence exists that either a high or a low salt diet predisposes to pregnancy-induced hypertension or that any other dietary component—energy, protein, or any micronutrient—is directly responsible, except perhaps calcium deficiency.[14]

Diet and discomforts of pregnancy



Diet and discomforts of pregnancy

Nausea and vomiting of pregnancy (NVP) is not confined to the mornings (so "morning sickness" is a misleading name). It is probably due to rising levels of pregnancy-associated hormones and often accompanied by increased olfactory sensitivity and aversion to strongly flavoured food and drink. In developed countries normal NVP appears to be beneficial, not detrimental to the fetus.[19] There has been no controlled trial of simple management. One opinion is that it is related to a low blood glucose concentration and that a dry biscuit or similar light snack before getting up may help. It now seems

possible that the increased cholecystokinin concentration could explain the symptoms. Unlike other conditions that cause nausea, women tend to put on weight during the phase of morning sickness.[8] Severe, persistent NVP, hyperemesis gravidarum is uncommon. When it occurs, note that thiamin is the most critical micronutrient.[20]

Constipation and its complication haemorrhoids are very common in pregnancy. All pregnant women should be advised to eat more wholemeal bread, bran, or bran cereals to loosen and increase the bulk of their faeces.

Heartburn should improve if the woman eats smaller meals and avoids foods which she finds indigestible. The common meal pattern of tiny breakfast, small lunch, and large dinner becomes unsuitable in late pregnancy. It is a good plan for her to have four, five, or even six small meals throughout the day. This also helps NVP.

Cravings and aversions—at some stage in pregnancy most women experience a distortion of their usual range of likes and dislikes of foods. Women may develop a nine-month aversion to foods they usually like—for example, fried foods, coffee, tea. Contrariwise and at the same time they may experience a craving for certain foods. These are often sweet foods, such as fruits and chocolate ice cream, and sometimes salty, but some remarkable non-foods—coal, soap, soil—have been recorded.

Vegetarians who are pregnant may need extra dietary advice. There are several types of vegetarian (chapter 7). Those most at risk are vegans. It is essential for them to take a supplement of vitamin B-12 for normal cerebral development of the fetus. Other lacto-ovo vegetarians, especially if they are prosperous and belong to a traditional vegetarian group, usually manage well enough but may want or need advice to optimise their protein and iron intakes. Legumes and nuts are an important part of a balanced vegetarian diet.

Food safety in pregnancy

- Avoid unpasteurised milk, soft cheeses and paté, raw eggs—danger of listeria and salmonella infection is more serious in pregnancy.
- Pre-cooked foods (for example, pies) should be thoroughly re-heated before eating.
- Avoid extra vitamin A, in the form of supplements or multivitamins containing vitamin A, or liver more than occasionally in early pregnancy. Retinoic acid is involved in normal morphogenesis and excess can be teratogenic.

What are pregnant women thinking about their food?

- "Eating the right foods", for example plenty of meat, fish, eggs, milk and fresh vegetables
- "Watching weight", taking care how much weight is gained
- "Eating for two", a largely outmoded idea

Based on Baric and MacArthur, cited by Anderson[21]

References

1 Boyd C, Sellars L. *The British way of birth*. London: Pan, 1982.
2 MRC Vitamin Study Research Group. Prevention of neural tube defects: results of Medical Research Council vitamin study. *Lancet* 1991; **338**: 131-7.
3 Department of Health. *Folic Acid and the prevention of disease*. Report of the Committee on Medical Aspects of Food and Nutrition Policy. London: Stationery Office, 2000.
4 Forrest F, Florey du VC. The relation between maternal alcohol consumption and child development: the epidemiological evidence. *J Publ Health Med* 1991; **13**: 247-55. [Review by members of the Dundee prospective team.]
5 Taggart NR, Holliday RM, Billewicz WZ, Hytten FE, Thomson AM. Changes in skinfolds during pregnancy. *Br J Nutr* 1967; **21**: 439-51.
6 Durnin JVGA. Energy requirements of pregnancy: an integration of the longitudinal data from the five country study. *Lancet* 1987; **ii**: 1131-3.
7 Durnin JVGA, McKillop FM, Grant S, Fitzgerald G. Energy requirements of pregnancy in Scotland. *Lancet* 1987; **ii**: 897-900.
8 Unväs-Moberg K. The gastrointestinal tract in growth and reproduction. *Sci Am* 1989; **July**: 60-5.
9 Department of Health. *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*. Report of the Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy. London: HMSO, 1991.
10 Subcommittee on the 10th edition of the Recommended Dietary Allowances, Food and Nutrition Board, National Research Council. *Recommended dietary allowances*. 10th edn. Washington, DC: National Academy Press, 1989.
11 Prentice A. Calcium in pregnancy and lactation. *Annu Rev Nutr* 2000; **20**: 249-72.
12 Harris HE, Ellison GTH. Do the changes in energy balance that occur during pregnancy predispose parous women to obesity? *Nutr Res Rev* 1997; **10**: 57-81.
13 Naeye RL. In: Dobbing J, ed. *Maternal nutrition: eating for two?* London: Academic, 1981.
14 Bucher HC, Guyatt GH, Cook RJ *et al*. Effect of calcium supplementation on pregnancy induced hypertension and pre-eclampsia: a meta-analysis of randomised controlled trials. *JAMA* 1996; **275**: 7113-17.
15 Barker DJP. *Mothers, Babies and Diseases in Later Life*. London: BMJ Books, 1994.
16 Rich-Edwards JW, Stampfer MJ, Manson JAE *et al*. Birth weight and risk of cardiovascular disease in a cohort of women followed up since 1976. *BMJ* 1997; **315**: 396-400.
17 Fraser R, Cresswell J. What should obstetricians be doing about the Barker hypothesis? *Br J Obstet Gynecol* 1997; **104**: 645-7.
18 Matthews F, Yudkin P, Neil A. Influence of maternal nutrition on outcome of pregnancy: prospective cohort study. *BMJ* 1999; **319**: 339-43, and subsequent correspondence 2000, **320**: 941-2.
19 Pirisi A. Meaning of morning sickness still unsettled. *Lancet* 2001; **357**: 1272.
20 Tesfaye S, Achari V, Yang YC, Harding S, Bowden A, Vora JP. Pregnant, vomiting and going blind. *Lancet* 1998; **352**: 1594.
21 Anderson AS. Pregnancy as a time for dietary change? *Proc Nutrition Soc* 2001; **60**: 497-504.

5 Infant feeding

Infant feeding is the dominant nutritional interest in less affluent countries and it gets much attention in Western countries because infants depend on others to feed them. For their first few months babies are fed only one food, so its composition is much more critical than the compositions of the many different foods in a mixed diet. Babies cannot eat ordinary adult food or say how they feel after the feed. Though there are still many questions, scientific knowledge is perhaps fuller about nutrition for this age of man than any other.

Breast or bottle?

For the first 4-6 months of life the infant should be fed either by breast feeding or on a formula based on cows' milk modified to make its composition suitable for infants—that is, more like breast milk. The decision on which method to use should be made well before delivery, and it should be made by the mother. The doctor's role is to give advice to help her make up her mind and then, whichever method she wants to use, to provide support and arrange instruction.

Advantages of breast feeding

- Breast feeding is natural and may confer advantages that science has not yet discovered.
- Breast milk is microbiologically clean.
- Breast milk's nutrient composition is the standard against which infant formulas for bottle feeding must be judged. Many of the differences between cows' and human milk have been minimised in modern infant formulas, but by no means all and some nutrients such as iron and zinc are known to be better absorbed from human milk.
- Only breast milk provides a complex range of anti-infective components: macrophages, lymphocytes, immunoglobulins (especially IgA), lactoferrin, lysozyme, complement, interferon, oligosaccharides (for example, bifidus factor), sialic acid, xanthine oxidase, gangliosides, glycoconjugates, growth factors, and enzymes.
- Breast feeding reduces the risk of gastrointestinal, respiratory and other infections (otitis media, meningitis, urinary tract infections), SIDS, childhood lymphomas, early allergic diseases, and type 1 diabetes.
- For most women breast feeding is a satisfying, convenient and enjoyable experience that is beneficial to the mother-child relationship.
- Mothers' milk is always at the right temperature.
- A mother can always change from breast to bottle feeding but not the other way round.

Breast feeding is recommended by the UK Department of Health, the WHO, the American Academy of Pediatrics and all authorities. The organisation of maternity wards (encouraged by UNICEF's Baby Friendly Hospital Initiative), control of advertising watched by WHO's Code of Marketing Breast Milk Substitutes, and change of social attitudes all make it easier than it used to be. On the other hand, modern technology makes bottle feeding easy and safe in developed countries and the newer infant formulas are closer to breast milk in nutrient composition.

> Breastfeeding from a woman who is in good health and nutritional status provides a complete food, which is unique to the species. There is no better nutrition for healthy infants at term and during the early months of life. … Breast feeding is preferable to feeding with infant formulas and should be encouraged.
>
> DHSS[5]

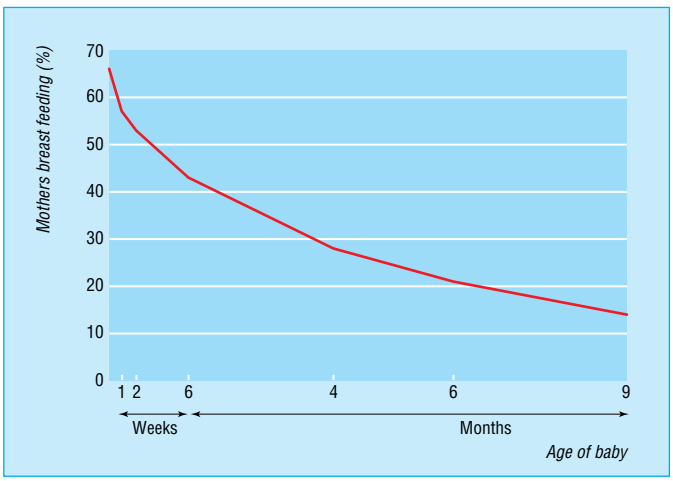
Composition of cows' milk compared with human milk and a modified infant formula (breast milk substitute) (All per 100 ml)

	Human* (mature)[1,2]	Cows' (full cream) (unfortified)	A modified milk formula† (powder diluted as directed)
Energy (kcal)	70	67	67
Protein (total) (g)	1.1	3.5	1.5
Casein (% protein)	40%	80%	40%
Carbohydrate (g)	7.4	5.0	7.2
Fat (total) (g)	4.2	3.7	3.6
Saturated fat (% fat)	46%	66%	44%
Linoleic (% fat)	7-11%	3%	17%
Sodium (mmol)	0.6	2.2	0.71
Calcium (mg)	35	120	49
Phosphorus (mg)	15	95	28
Iron (mg)	0.075	0.050	0.8
Vitamin C (mg)	3.8	1.5	6.9
Vitamin D (μg)	0.8	0.15	1.1

* The composition of breast milk varies considerably with stage of lactation, between individuals, and with maternal nutrition
† Mean of Cow and Gate Premium and SMA Gold Cap

In Third World countries breast feeding unquestionably reduces infant mortality.

In affluent countries, however, epidemiologists have difficulty in showing an appreciable reduction in mortality when confounding factors are taken into account. Mothers who breast feed tend to have higher educational and income levels. A well designed study in Dundee seems to have corrected for all such confounding variables. It showed that breast feeding for the first three months of life confers a protection against gastrointestinal illness, which persists beyond the period of breast feeding itself.[3,4]



Prevalence of breast feeding in Great Britain, 1995[6]

Two examples of ongoing research about human milk

Docosahexaenoic acid (DHA)

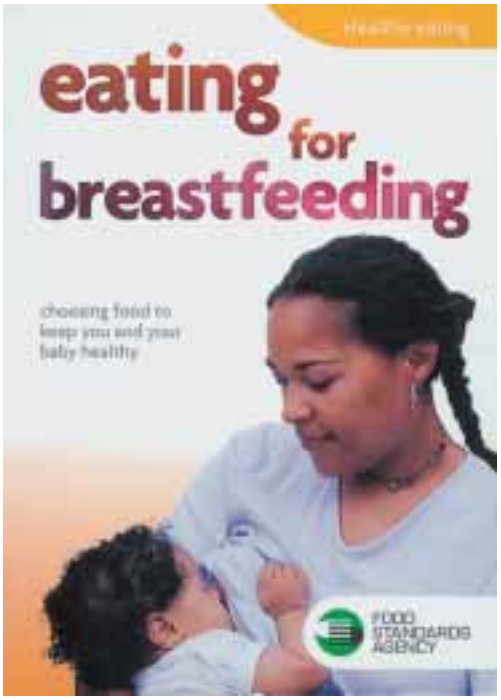
The brain grows rapidly in infancy, from 350 g at birth to 1000 g at 12 months. Sixty per cent of its solids are lipids and two very long chain polyunsaturated fatty acids are more abundant here and in the photoreceptors of the retina than elsewhere—DHA (22:6, ω-3) and AA (arachidonic, 20:4, ω-6). DHA is present in human, not cows' milk. It is synthesised in the body from α-linolenic (18:3, ω-3) but probably not fast enough for the brain's requirements especially in premature babies. Young infants fed on standard formulas had lower DHA concentration in red cells and brain (SIDS post-mortem) than breast fed infants.[7]

Randomised controlled trials have found better eye function in pre-term infants who were breast fed or given formula enriched with DHA and AA than in babies fed on standard formulas. In full term infants a few (not all) trials have had similar results.[8] It is not known whether there will be benefits beyond infancy. Socio-economic and psychological confounding factors will make interpretation of brain function tests difficult. Some manufacturers now add DHA and AA to premium infant formulas.

Oligosaccharides

Lactose is not the only sugar in human milk. The concentration of oligosaccharides is higher than the protein! Over 100 of these oligosaccharides have been chemically defined—all made up of five monosaccharides: fucose, galactose, glucose, *N*-acetylglucosamine, and sialic acid (NANA) and ranging from three to ten residues in length. Cows' milk, and infant formulas, contain only trace amounts. These human milk oligosaccharides (HMOs) are not digested in the small intestine. Small amounts are absorbed and found in the urinary tract. Most passes to the large intestine where it acts like dietary fibre.[9] Oligosaccharides containing N-acetylglucosamine promote the growth of bifidobacteria, which are the dominant colonic bacteria in breast fed infants.

Micro-organisms and their toxins gain entry to cells by attaching to specific sugars on the cell surface. Oligosaccharides in mucus and in human milk include particular sugars that can act as decoys for many specific micro-organisms and so prevent their access to the body. HMOs have been shown to include receptors for *E. coli, E. coli* toxins, Campylobacter, Candida, Rotavirus and *Strep pneumoniae*.[10]



Eating for breast feeding. Reproduced with kind permission from the Food Standards Agency

How to manage breast feeding

Knowing how to establish breast feeding is no longer instinctive in the women of our complex industrial societies. Some take to it naturally but others will not do well without guidance and a sympathetic environment.

- The mother should be adequately nourished during pregnancy.
- She should have watched others breast feeding and talked about it.
- She needs to involve and consider her partner. There can be sexual implications in breast feeding. His support (or opposition) is important.
- The National Childbirth Trust or La Lèche League or the Association of Breastfeeding Mothers can help to provide information and support (for addresses see end of this chapter).
- The baby should be put to the breast as soon as possible after delivery.
- The midwife can play an important part, giving advice on positioning the baby, and encouragement in the first few days.[11]
- Frequent suckling stimulates prolactin secretion. Suckling more than six times a day maintains high basal prolactin as well as initiating prolactin surges with feeding.
- Feeding should be on demand or baby-led.
- After delivery the baby should be in a crib next to its mother all or most of the time and suckled whenever it seems to be hungry. Colostrum is a concentrated anti-infective fluid.
- Relaxation and privacy are needed.

- The baby should not be given other complementary milk or juice—only water if necessary.
- The baby should feed from both breasts each time and start the feed with the breast used last.
- Advice may be needed about sore nipples or breasts, oversupply, or undersupply.
- In developed countries rates of breast feeding are lower in younger, less educated mothers in less skilled occupations, and in single mothers.

Obstacles to initiation and continuation of breast feeding are listed in the box alongside.

Contraindications to breast feeding are rare: galactosaemia in the infant; mother uses illegal drugs; mother has active untreated tuberculosis; mother has HIV infection (controversial); mother has to take therapeutic drugs that adversely affect her milk—radioisotopes, cancer chemotherapy, etc.

Obstacles to breast feeding[12]

- Doctor's apathy and misinformation
- Insufficient prenatal education in breast feeding
- Disruptive hospital policies
- Inappropriate interruption of breast feeding
- Early hospital discharge
- Lack of regular home health visits, post-partum
- Maternal employment (especially if no workplace facilities or support for breast feeding)
- Lack of broad societal support
- Portrayal by media of bottle feeding as normative
- Commercial promotion of infant formula, for example distribution of hospital discharge packs

Nutrition for the lactating mother

Except in malnourished communities, there is little evidence that dietary calories, protein, fat, water, or anything else have a consistent effect on milk volume. Regular and fairly frequent suckling is the well established stimulus. Human lactation works more by pull than by push.

Some constituents in the milk are affected by the mother's intake.

(1) Fatty acid pattern, vitamin A, thiamin, riboflavin, biotin, folate, vitamin B-12, and vitamin C are affected, especially downwards if the mother's diet is deficient.
(2) Zinc, iron, fluoride, and vitamin D may be responsive in some circumstances, but more research is needed.
(3) Protein, lactose, total fat content, calcium—that is, the major proximate constituents of milk—do not appear to be affected.
(4) Specific proteins in the mother's diet might be excreted intact in small amounts and an allergic (IgE) reaction occasionally occurs in the baby.
(5) The amount of caffeine in the milk after a cup of coffee is only about 2% of the maternal dose. Likewise, the alcohol concentration of breast milk is about the same as that of plasma so single drinks of coffee or alcohol, well spaced out, are harmless, but the babies of alcoholics can be affected. Beer stimulates prolactin secretion (at least in non-lactating women) and so might increase lactation. Milk production is reduced in heavy smokers.
(6) The fat-soluble environmental contaminants, polychlorinated biphenyls, dry cleaning solvents, and organochlorine insecticides (DDT, etc), are stored in adipose tissue and excreted in the cream of breast milk (though the DDT group is fairly innocuous in man).

Drugs and lactation

- For most drugs the concentration in human milk is of the same order of magnitude as the plasma concentration or in some cases less. The infant would thus receive around 1% of the maternal dose. But the milk/plasma ratio is 12 for propylthiouracil and 25 for iodine-131.
- Other drugs are contraindicated if they are radioactive, can cause allergy, agranulocytosis or bleeding disorders, or are poorly metabolised in the newborn, or can suppress lactation. These include chloramphenicol, indomethacin, diazepam, reserpine, anti-cancer drugs, lithium, and some others.
- Tetrahydrocannabinol is concentrated in the milk of cannabis smokers, as are opiate narcotics in the milk of those taking them.
- The *British National Formulary* has an appendix on prescribing during breast feeding.

Constituents of milk affected by mother's intake

- Fatty acid pattern, vitamin A, thiamin, riboflavin, biotin, folate, vitamin B-12 and vitamin C
- Possibly zinc, iron, fluoride and vitamin D
- Protein, lactose, total fat content, calcium
- Some proteins in mother's diet
- Caffeine in milk after coffee, alcohol after alcohol consumption (only in large doses)
- Environmental contaminants

The mother's need for extra nutrients

A good average production of breast milk is 800 ml/day, and the mother's extra nutritional requirements are calculated from this and the average composition of milk, taking into account the available information about efficiency of absorption. The gross energy value of average human milk is 280 kJ/100 g and efficiency of conversion from maternal dietary energy to milk energy is assumed to be 80%. Hence the energy lost in exclusive breast feeding in the first three months is:

$$800\,\text{ml} \times 280\,\text{kJ} \times 100/80 = 2.8\,\text{MJ}\ (675\,\text{kcal}).^{13}$$

If, as is usual, the mother does not eat the full amount of this extra energy she will lose some of the body fat put on during

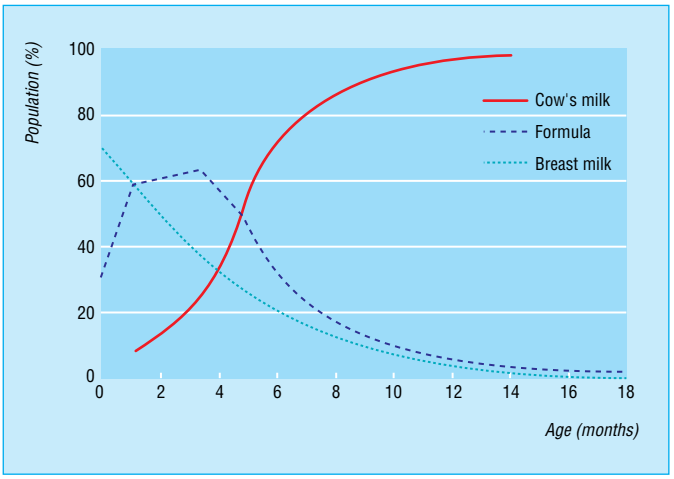
pregnancy. When the infant is getting other foods the energy expenditure on breast milk usually declines.

Most of the nutrients come along with the extra calories; lactating women usually have a good appetite and if this is satisfied by a mixed diet the nutrients that need watching (because there is little excess in the diets of non-lactating women) are calcium, iron, folate, and vitamin D. The extra **calcium** can come from a pint of milk or two cartons of yoghurt. Calcium metabolism changes during lactation. There is some loss of bone density, which is apparently not prevented by calcium supplements. These changes are reversed when lactation ceases.[14] There is no evidence that women who have breast fed have increased incidence of osteoporosis. **Iron** supplements may be advisable, and **vitamin D** supplements are recommended for any mother whose vitamin D status is in doubt (such as Asian mothers eating a wholly vegetarian diet). **Folate** deficiency incurred during pregnancy may first show as anaemia in the puerperium. **Zinc** is secreted in the milk but staple isotope studies show increased zinc absorption during lactation.

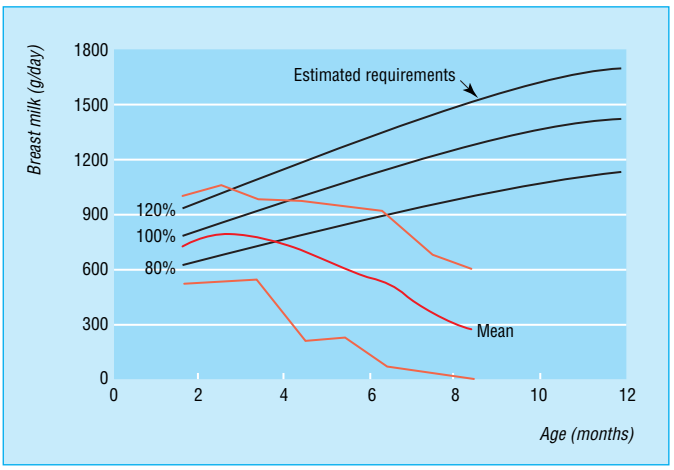
Mothers return to pre-pregnant weight?
Mothers are more likely to lose the fat stores put on during pregnancy if they choose to breast feed. The energy lost in lactation is usually more than the mother's increased food intake over her non-pregnant, non-lactating level. DHSS estimated an average energy deficit of 0.5 MJ (120 kcal) per day, which corresponds to a fat loss of 0.5 kg per month (14.5 MJ).[13] But in fact appetite and weight loss during lactation is highly variable.[15] Obviously lactation cannot contribute to worthwhile weight loss if it is only brief. The question arises whether milk production will suffer if the mother deliberately restricts her food intake. Lovelady *et al.* tested this out in a randomised controlled trial in overweight (not obese) women. (BMI 25-30).[16] They lost approximately 0.5 kg per week between 4 and 14 weeks post partum from moderate food restriction and exercise: their infants gained the same weight and length as the controls, but some of the control mothers put on weight.

Ending lactation
In an industrial population the prevalence of breast feeding goes down with infant's age in a curve reminiscent of first order elimination kinetics. A few mothers continue breast feeding towards or beyond 12 months. In a British national sample the major reasons for stopping in the first six weeks were insufficient milk (54%) and painful breasts or painful or inverted nipples (18%); the commonest reason for stopping between 6 and 16 weeks was also insufficient milk (66%). Those with insufficient milk early on never got lactation well established. Those with insufficient milk later may have had normal volume production but the baby's energy needs started to outgrow this.

Complementary and supplementary bottles of milk
Complementary bottle feeds are used to finish off a breast feed and supplementary bottle feeds replace a breast feed. The occasional bottle feed once a day or less is convenient if the mother has to leave the baby with a friend, but regular topping up of the baby's intake with bottle milk is likely to reduce sucking and breast milk production. Some mothers produce less milk than others, however, and if the baby is not gaining, and hungry on pure breast feeding with good technique, extra bottle feeding may be necessary.



Mean consumption of different types of milk in normal Canadian infants[17]



Measured breast milk intakes of Cambridge infants. Mean and ranges against estimated requirements[18]

Bottle feeding

Some mothers choose to bottle feed from the start and others will change over from breast to bottle feeding after weeks or months, so they need practical advice.

- A cows' milk formula specially modified for infants should be used in which the protein has been reduced, the casein partly replaced by whey protein, the fat made more unsaturated, the lactose increased, sodium and calcium reduced, and enough of all the essential micronutrients added.
- Bottles and teats should be washed in water and detergent (the bottle brush used only for this), rinsed and sterilised by boiling in water or by standing covered in sterilising solution (usually hypochlorite) in a plastic container. It saves time to prepare several bottles at once. Empty the water out of each bottle, without touching the inside, then fill to the mark with recently boiled water that has cooled some minutes, not too hot or it will destroy some vitamins and may produce clumping.
- Exactly the amount of power in the manufacturer's instructions should be put into the (wide mouthed) bottle, using the scoop provided (levelled with a clean knife, not pressed down). "One for the pot" can lead to obesity. Mothers and even nurses are often found to prepare feeds inaccurately. Screw on the cap and shake the bottle well. Bottles may be kept in the refrigerator for up to 24 hours.
- If the hole in the teat is too small it can lead to aerophagia or underfeeding. Milk should drip from the inverted teat at about one drop per second. Teats need replacing every few weeks.
- Babies do not mind cold milk but usually prefer it warm. The bottle should be not warmed for too long and the milk's temperature should be checked by dropping some on the parent's skin. Infant feed should be not warmed in a microwave oven once it is in the feeding bottle. Very hot fluid at the centre of the bottle may be missed and may scald the baby.[5] For about the first eight weeks of life babies need to be fed every three to four hours, including the small hours of the morning. (Fathers can bottle feed as well as mothers.) By the end of the first week most babies are taking 120-200 ml/kg per day (160 ml/kg corresponds to the old $2\frac{1}{2}$ fluid ounces per lb bodyweight).
- Cereals or rusks should not be added to milk in the bottle and babies should not be left to sleep with a bottle in their mouth.
- Vitamin drops, fruit juices, are not required as supplements to modern infant formulas.
- Uncles, grannies, and baby sitters can give a bottle feed but parents should feed their infant themselves as much as possible with the same sort of closeness, cuddling, and communication as in breast feeding.

Fifteen month old child bottle feeding. (Courtesy of Mr PM Whitfield, reproduced with permission)

Weaning[19]

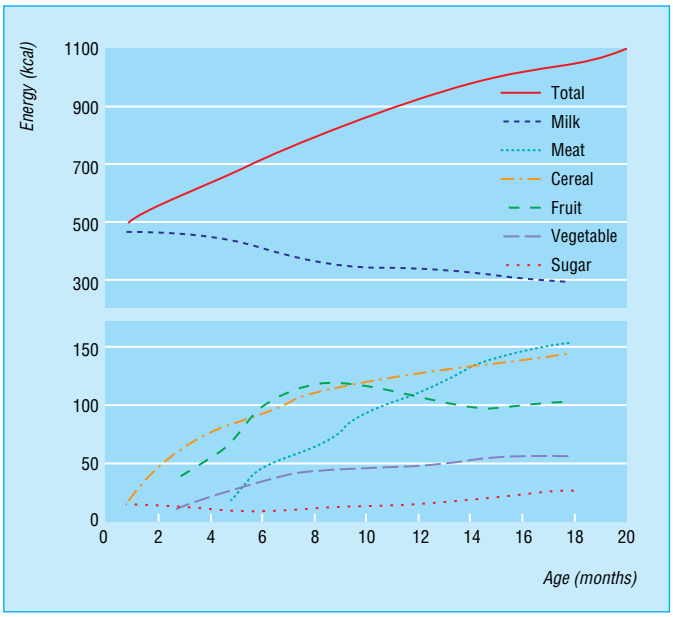
In the first six months

Young infants cannot deal properly with solid foods (in reality semisolid foods at first) for the first four months. The natural time for starting solids (beikost) is when the energy provided by well established breast feeding starts to become insufficient. The Department of Health and other authorities advise that the introduction of any food to the baby, other than milk, should be unnecessary before the age of 4 months, but mothers may be tempted to jump the gun. Most babies should start a mixed diet not later than the age of 6 months.

Weight in the lower half of the standard percentiles without other symptoms is not an indication to augment breast feeding. Breast fed babies tend to put on weight (and length) a little more slowly than bottle fed infants. Indeed, the standard percentiles, derived mostly from bottle fed babies, may not be ideal. The time to start thinking about adding solids is when the infant still seems hungry after a good milk feed. But by

Babies cannot cope with solid food in the first few months because:

- the extrusion reflex prevents spoon feeding
- they cannot swallow solids
- pancreatic amylase is not produced for the first three months
- pancreatic lipase is absent for the first month (fat digestion in breast milk is facilitated by the bile salt-activated lipase it contains)
- there is an increased likelihood of absorption of intact foreign (food) proteins.



Mean consumption of energy from different foods in normal Canadian infants[17]

six months body stores of several nutrients, such as iron, zinc, and vitamin C, are often falling in exclusively milk fed infants, whether from breast or bottle.

When solids are introduced, single ingredient foods should be used and started one at a time at half weekly intervals so that there is time to recognise allergy or other intolerance to each food. A little of the food on the tip of a teaspoon is enough at first, given after a milk feed when the baby is wide awake.

Infant cereals (usually enriched with iron) are traditional foods to start with; rice is better before wheat. They can be thinned with baby's usual milk (mother's or formula) or water. Thereafter different soft foods can be added: mashed potato; soft porridges; puréed fruit and vegetables, meat, or chicken. Foods should be semisolid—sieved or blended or commercial baby food. It is nutritionally sensible to give a balance of foods from the four major food groups: cereals, vegetables/fruit, dairy products, and meats/fish. Combination foods should not be given until tolerance to their individual components is established. Egg should not be started before six months because of the chance of allergic reactions, and then it is best to begin with a small amount of cooked yolk. Spinach, turnip, and beets can contain enough nitrate to cause methaemoglobinaemia in young infants. Coffee and tea should not be given. Babies should not be left alone while they are eating.

In the second six months

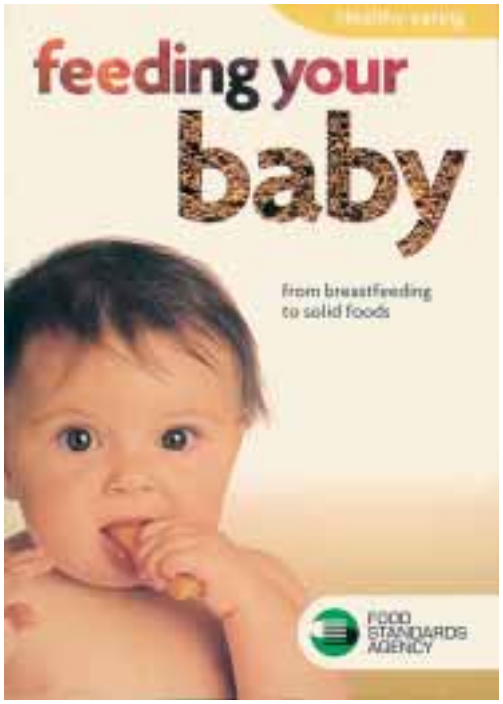
In the second six months other liquids can be given from a cup, especially citrus fruit juices. Untreated cows' milk can sometimes cause gastrointestinal bleeding from irritation by the bovine serum albumin. This does not happen with boiled milk or infant formulas (which have been heat treated). Iron-fortified infant formula contributes to iron intake, which is critical in the second six months of life. It is wrong to add any salt to the foods given to infants. A fully breast fed infant receives only about one-twentieth of the sodium in a typical British adult diet. There has been a quiet revolution in commercial baby foods; most contain no added salt or colours and only up to 4% sugar (needed with sour fruits). Infants' sodium intakes have been found to shoot up after six months but more from home prepared rather than commercial baby foods.

An increasing range of foods is given in the second six months. Variety is likely to cover the needs for most nutrients and provide a basis for healthy food habits. Some fruits or vegetables should be given each day, but the most critical nutrients at this stage are protein and iron: finely minced beef and legumes should be given regularly and the protein in cereal foods should not usually be diluted by refining or by added fat or sugar. Foods should become progressively more chewy and fibrous and include rusks and other finger foods like bread or cheese. Babies do not usually like strongly flavoured foods like pickled onions. Nuts, popcorn, raw peas, and similar small hard foods should be avoided; they can be breathed in accidentally. Commercial baby food manufacturers offer a succession of "strained", "junior", and "toddler" foods for maturing babies, and similar meals are usually made at home. Some cookbooks for babies are more sensible than others.

Milk continues to be the main source of calories but a diminishing one. Sweetened fruit juices should be given by cup not bottle because the latter can promote dental caries. Infantile obesity is probably becoming less common in the United Kingdom now that people are aware of it. It is not usually caused by bottle feeding or early introduction of solids in themselves, but by more concentrated feeds, by pushing food at mealtimes, or by snacks in between. Between feeds, water for thirst and a minimum of snacks or sweets are good general rules.

A suggested timetable for the introduction of solid foods

- 1-4 months — Breast milk only
- 4-6 months — Cereal(s) added
- 6-7 months — Vegetables (puréed) added
- 8-9 months — Start finger foods (rusk, banana) and chopped (junior) foods
- 9 months — Meat, citrus juice (from a cup)
- 10 months — Egg yolk (cooked), bite-sized cooked foods
- 12 months — Whole egg, most table foods
  No peanuts or hard particles of similar size



Feeding your baby (from breast feeding to solid foods). Reproduced with kind permission from the Food Standards Agency

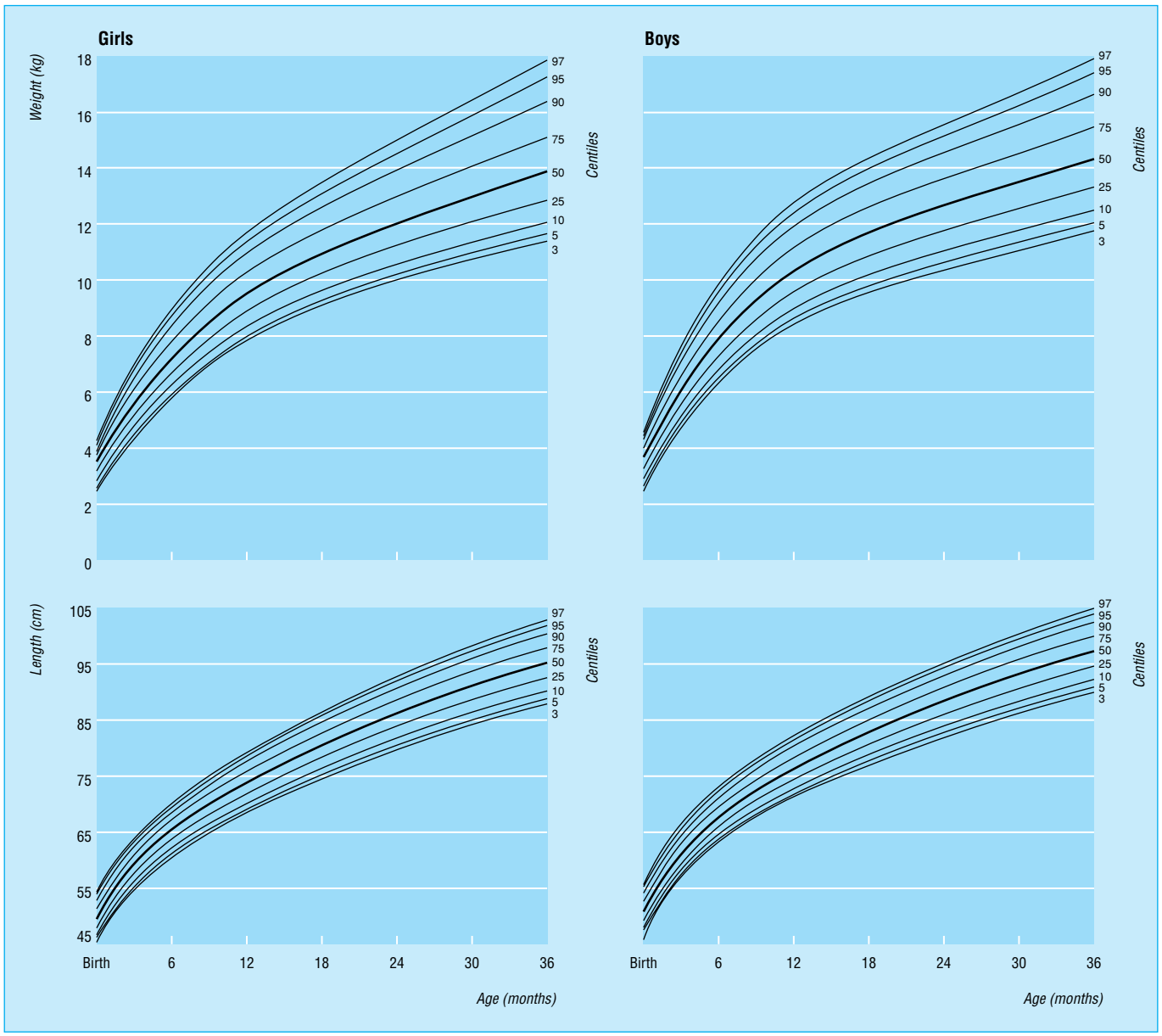
In a survey for MAFF of food and nutrient intakes of British infants aged 6-12 months,[20] the percentage contribution of food types to energy intake were:

- *In infants 6-9 months old (median energy 792 kcal)*, family foods 30%, infant formula 23%, infant foods 23%, cows' milk 18%, breast milk 6%
- *In infants 9-12 months old (median energy 894 kcal)*, family foods 53%, cows' milk 28%, infant foods 11%, infant formula 7%, breast milk 1%.

Two other nutrients are not adequately supplied in all mixed diets. In communities where rickets occurs—for example, among Asian babies in northern cities—a supplement of vitamin D 10 μg (400 IU) a day is good insurance. Bottle fed infants who are consuming 500 ml infant formula as follow on formula a day do not need vitamin supplementation because these manufactured products are fortified with vitamin D. In areas where the drinking water is not fluoridated, sodium fluoride prophylactic tablets or drops (0.25 mg/day) should be considered.

Useful addresses for help with breast feeding

- Association of Breast Feeding Mothers, PO Box 207, Bridgewater, Somerset, TA6 7YF http://home.clara.net/abm,(0)20 7813 1481.
- La Lèche League, PO Box BM 3424, London WC1N 3XX. http://www.laleche.org.uk, (0)20 7242 1278.
- National Childbirth Trust, Alexandra House, Oldham Terrace, Acton, London W3 6NH. http://www.nctpregnancyand babycare.com, 0870 770 3236.



Center for Health Statistics graph of percentile weights and heights for girls and boys from birth to 36 months, adopted by the World Health Organization[21]

References

1 Department of Health & Social Security. *The Composition of Mature Human Milk*. (Report on Health & Social Subjects no 12.) London: HMSO, 1977.

2 Reeve LE, Chesney RW, de Luca HF. Vitamin D of human milk: identification of biologically active forms. *Am J Clin Nutr* 1982; **36**: 122-6.

3 Howie PW, Forsyth JS, Ogston SA, Clark A, Florey C du V. Protective effect of breast feeding against infection. *BMJ* 1990; **300**: 11-16.

4 Wilson AC, Forsyth JS, Greene SA, Irvine L, Hau C, Howie PW. Relation of infant diet to childhood health: seven year follow up of cohort of children in Dundee infant feeding study. *BMJ* 1998; **316**: 21-5.

5 Department of Health & Social Security. *Present Day Practice in Infant Feeding: Third Report*. (Report on Health and Social Studies no 32.) London: HMSO, 1988.

6 Foster K, Lader D, Cheesbrough S. *Infant feeding 1995*. London: HMSO, 1997.

7 Farquharson J, Jamieson EC, Abbasi JA, Patrick WJA, Logan RW, Cockburn F. Effect of diet on the fatty acid composition of the major phospholipids of infant cerebral cortex. *Arch Dis Childhood* 1995; **72**: 198-203.

8 Gibson RA. Long chain polyunsaturated fatty acids and infant development. *Lancet* 1999; **354**: 1919-20.

9 Brand-Miller JC, McVeigh P, McNeil Y, Messer M. Digestion of human milk oligosaccharides by healthy infants evaluated by the lactulose hydrogen breath test. *J Pediatr* 1998; **133**: 95-8.

10 Newburg DS, Street JM. Bioactive materials in human milk. Milk sugars sweeten the argument for breast-feeding. *Nutr Today* 1997; **32**: 191-201.

11 Anonymous. *Successful breastfeeding. A practical guide for mothers and midwives and others supporting breastfeeding mothers*. London: Royal College of Midwives, 1989.

12 American Academy of Pediatrics: Work Group on Breast feeding. Breast feeding and the use of human milk. *Pediatrics* 1997; **100**: 1035-9.

13 Department of Health. *Dietary reference values for food energy and nutrients for the United Kingdom. Report of the Panel of the Committee on Medical Aspects of Food Policy*. London: HMSO, 1991.

14 Prentice A. Calcium in pregnancy and lactation. *Annu Rev Nutr* 2000; **20**: 249-72.

15 Butte NF, Hopkinson JM. Body composition changes during lactation are highly variable among women. *J Nutr* 1998; **128 Suppl**: 381S-385S.

16 Lovelady CA, Garner KE, Moreno KL, Williams JP. The effect of weight loss in overweight, lactating women on the growth of their infants. *N Engl J Med* 2000; **342**: 449-53.

17 Yeung DL. *Infant nutrition. A study of feeding practices and growth from birth to 18 months*. Ottawa: Canadian Public Health Association, 1983.

18 Whitehead RG, Paul A, Rowland MGM. Lactation in Cambridge and in the Gambia. In: Wharton BA (ed) *Topics in paediatrics 2*. Tunbridge Wells: Pitman (for the Royal College of Physicians), 1980.

19 Department of Health. *Weaning and the weaning diet. Report of the Working Group on the Weaning Diet of the Committee on Medical Aspects of Food Policy*. (Report on Health & Social Subjects no 45.) London: HMSO, 1994.

20 Mills A, Tyler H (Ministry of Agriculture, Fisheries & Food). *Food and nutrient intakes of British infants aged 6-12 months*. London: HMSO, 1992.

21 National Center for Health Statistics/National Center for Chronic Disease Prevention and Health Promotion. New childhood growth charts, 2000. http://www.cdc.goc/growthcharts

6 Children and adolescents

Pre-school children (1-5 years of age)

What children eat between being weaned and starting school is important. Food habits are established in this phase of life. They are less under parental control than they used to be. Mothers often go out to work, the toddler eats at a child care centre or pre-school. Food companies make some products specially for this age group and young children are targeted by advertisements on television. Young children have different taste perception from adults. They prefer sweeter foods, dislike bitter tastes, and often object to vegetables of the cabbage family.

Young children's eating is still mostly controlled by their mothers. Four simple principles should help them.

(1) Eat from each of the **four basic groups** every day (the fifth will accompany the basic four). These four basic food groups contain all the essential nutrients between them.
(2) Aim for **variety** within each food group—for example not always the same vegetable. The modern child should be prepared to eat food from other cultures. Food neophobia is a social handicap in later life. Young children should be able to eat most of the same food as the rest of the family.
(3) To minimise **dental caries**, sweets and other sugary foods should be rationed and not eaten between meals as a rule.
(4) The most likely **nutrient deficiency** in Britain is a low iron status with mild anaemia.[2] The easiest way to prevent this is for the child to be given meat, chicken or fish regularly, which provide haem iron.

Rickets in the northern winter, though now uncommon, is prevented by children getting out of doors regularly, eating plenty of the foods that contain some vitamin D and, in vulnerable groups, taking BNF vitamin capsules or children's vitamins drops (contains 7.5 μg vitamin D) during the winter. In a sample of Asian children, $1\frac{1}{2}$-$2\frac{1}{2}$ years old in England, 20-34% had 25-OH vitamin D levels under 25 nmol/l.[3] Low vitamin D status tended to be associated with low haemoglobin and serum iron.

Overweight increasing

Health visitors in the Wirral, NW England weigh and measure pre-school children regularly. Records from over 28 000 children show that the percentage overweight (>85% of the British growth reference BMI) rose from 14.7% in 1989 to 23.6% in 1998.[4] What has changed since the 1980s? The *Lancet* asked:

> Children's leisure activities have shifted from active outdoors play—often because of parents' perception that neighbourhoods are unsafe—to television watching and computer games. Cars are used even for short distance journeys. And children's diets including school meals, commonly consist of fast food with high fat and sugar contents. Many parents, especially when overweight themselves, do not recognise obesity in their children.[5]

Faddy toddlers

After having a reliable appetite for their first twelve months, some children aged 1-3 go through a phase of poor eating, which can make parents anxious or even exasperated. Children are less enthusiastic about eating and refuse foods that the rest of the family eat, especially vegetables.

Food groups for nutrition[1]

(1) Bread, other cereals and potatoes
(2) Fruits and vegetables
(3) Milk and dairy foods
(4) Meat, fish, and alternatives
(5) Foods containing fat
Foods containing sugar

1-4 are the basic four food groups. The fifth group includes foods low in nutrient density that should be used sparingly or enjoyed as treats.

Foods that contain some vitamin D

- Milk, eggs, cheese, butter, liver, all contain some.
- Fatty fish (herring, mackerel, sardines, pilchards, tuna, and salmon) contain more.
- Infant milks, infant cereals, margarines, some breakfast cereals, some yoghurts, and branded food drinks have vitamin D added during manufacture.

Favourite foods

- In the 1992-93 survey of British children aged $1\frac{1}{2}$-$4\frac{1}{2}$ years[2] the foods or drinks most frequently recorded over 4 days were (in descending order): biscuits (88%); white bread (86%); non-diet soft drinks (86%); whole milk (83%); savoury snacks (potato crisps or cereal-based) (78%); boiled, mashed or jacket potatoes (77%); chocolate confectionery (75%); potato chips (71%); chicken (70%); breakfast cereals (66%); cheese (59%); sugar confectionery (58%); buns, cakes and pastries (55%); sausages (53%); beef dishes (47%); pasta (51%); carrots (54%); peas (53%); baked beans (49%); apples & pears (50%); bananas (46%); fruit juice (36%); eggs (46%); yoghurt (40%); coated and fried white fish (36%); leafy green vegetables (39%).
- Many of the top half of items in this frequency list are high in fat, added sugar or salt. It would be preferable if the lower half of the list had been at the top.

One reason for this is that growth slows at around twelve months. This can be seen in the changes in gradient of normal weight for age, and height for age curves. It is even more obvious in weight and height velocity curves, which descend to about a quarter and a half, respectively, of the values in early infancy. Energy intake can be very variable from time to time in toddlers.

Other things are also happening. Children are discovering their independence and testing their choice in food selection. Once they have some control over what is offered, foods that they find unattractive are displaced by those they think delicious: cakes, biscuits, chocolate, crisps, ice cream, etc.

Schoolchildren

In developed countries undernutrition is exceptional nowadays—a big advance from Dickens' time. Children are growing taller than ever.[8] Deficiency is limited to a few micronutrients and usually subclinical. Iron nutrition is low in a minority of older girls (their requirements increased by menstruation) and plasma 25-OH vitamin D is low in some susceptible children.

The main nutritional concerns in schoolchildren are:

- the increasing percentage who are overweight, even obese
- whether the most frequently consumed foods are healthiest in the long term? Vegetable, fruit and whole grain consumptions are often below dietary guideline recommendations (see chapter 7)
- the continuing opportunity to minimise dental caries by reducing the time that teeth are exposed to acid-producing carbohydrates in the mouth
- diets of socially deprived children can be of poorer quality.[9]

Overweight

It is now agreed that overweight can be diagnosed with body mass index (kg/m^2) using reference curves worked out for children's different ages. There are BMI reference curves for the United Kingdom (1995) and for the United States (2000). The cut offs shown on page 34[11] were calculated by Cole *et al.* for international use, based on data from Brazil, Great Britain, Hong Kong, the Netherlands, Singapore and the USA.

The role of high fat foods in the epidemic of overweight is generally known, and realised among schoolchildren themselves, though cheese or biscuits may not be recognised as high fat. Another contributor to overweight, more recently revealed, is the widespread consumption of sugar-sweetened soft drinks.[12] It is sometimes hard to visualise how much sugar and calories are in these liquids. In the NDNS survey[8] 76% of 4-18 year olds consumed "carbonated soft drinks, not low calorie" over seven days while only 61% drank tap water!

Energy expenditure has fallen in young people. Many do not participate in sport, are driven to school and spend hours everyday looking at television or a personal computer. Children who watch more TV have higher BMIs[13] and most of the food adverts they see are for fast foods, high in fat, sugar or salt.

How could schools help?

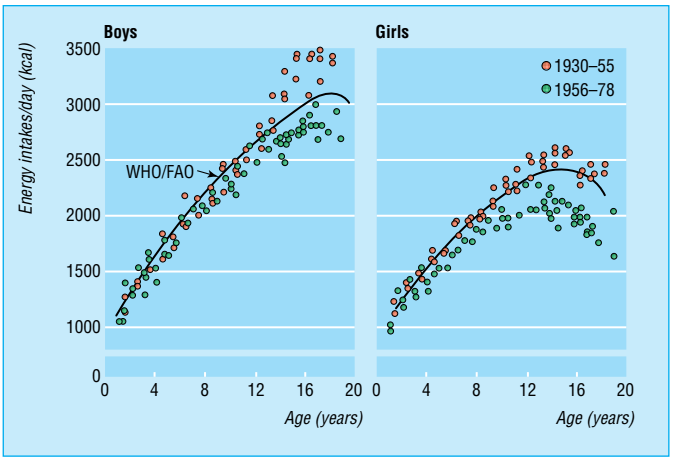
(Suppose a local doctor is on the school board.)

Much could be done (though nutrition is low priority at present for management, teachers, and pupils), including:

- there needs to be modernised food education across the years
- ALL pupils should get regular physical exercise; there is too much emphasis on competitive sport
- foods served in and around the school should help children eat a varied and nutritious diet

Advice for parents of faddy toddlers[6,7]

- Most toddlers will eat some form of bread, cereals, meat, milk, and some fruits and fruit juices. The fibre foregone in vegetables can be replaced by breads and the vitamins by fruits and juices.
- Do not have battles over foods. Do not use bribes or force. Try to avoid tension at meal times.
- Keep meal preparation for toddlers easy and quick, so you can accept it without anger if they will not eat what you have made for them.
- Make eating fun—for example, cut sandwiches, pieces of fruit, cheese into patterns. Let toddlers eat at a small table, etc.
- When it is clear that they will not eat any more, let them leave the table. Do not insist on clean plates; serve a little less next time. With proper education, parents may be led to understand that a "good" eater is not a big eater but a moderate eater.
- Rejection of new foods is usually a transitory phenomenon which can be reduced by repeated exposure to small quantities without undue coercion (followed by praise if the child eats some). Parents should not interpret initial rejection to mean the child has an immutable distaste for the food.
- Children are more likely to try new foods offered in small amounts at the start of meals when they are hungry. If they help preparing (or even growing) vegetable foods (for example, shelling peas or making a salad) they may like to taste them.
- Do not give delicious (for them) high fat and sugar foods (cake or chocolate) before a meal **or** if they will not eat their main meals.
- Many toddlers cannot adjust to their parents' three meals but will eat nutritious foods as snacks.



Mean energy intakes of boys and girls studied in 1930-55 compared with more recent studies in relation to WHO/FAO (1973) recommendations[10]

National Diet & Nutrition Survey, Great Britain[8]

- Young people, aged 4-18 years from 1700 weighed dietary records.
- Most frequently consumed (% consuming over 7 days): White bread (95), milk (93), potato chips (89), biscuits (84), chocolate confectionery (82), other potato dishes (82), carbonated drinks (76), cakes (74), chicken (74).
- But consumption frequency was lower for: sausages (64), baked beans (60), apples, pears (54), beef, veal (51), peas (50), fruit juice (49), fried coated fish (46), pizza (45), green vegetables (39), bananas (38), citrus fruits (26).
- Average fat intake total 35.5% of energy, saturated 14.25% of energy, polyunsaturated 5% of energy.

- resist sponsorship by food companies and vending machines
- all children's weights and heights should be measured at school once a year.

James and McColl have written a paper for government departments setting out the justification for **schools that promote a healthy lifestyle** and how changes can be made.[14]

Adolescents

> Teenagers are not fed; they eat. For the first time in their lives they assume responsibility for their own food intakes. At the same time they are intensely involved in day-to-day life with their peers, and preparation for their future lives as adults. Social pressures thrust choices at them: to drink or not to drink, to smoke or not to smoke, to develop their bodies to meet sometimes extreme ideals of slimness or athletic prowess. Few become interested in foods and nutrition except as part of a cult or fad such as vegetarianism or crash dieting.[15]

Several facets of eating behaviour are different or more pronounced in adolescents than in other people and each may cause concern in the older generation.

- **Missing meals**, especially breakfast. This does not usually affect classroom performance, partly because of—
- **Eating snacks** and confectionery. The major snack is usually in the afternoon, after school. Snacks tend to be high in "empty calories"—fat, sugar, and alcohol—but some provide calcium (for example milk) or vitamin C (fruit).
- **"Fast", take-away, or carry-out foods**. These provide some nutritious portions, but adolescents may not choose balanced meals from what is offered. There is not enough accessible information about the nutrient composition of fast foods.
- **Unconventional meals** may be eaten in combinations and permutations that other members of the family do not approve of, but they often add up to an adequate nutritional mix.
- **Start of alcohol consumption.** This is the most dangerous of the new food habits. Alcohol-related accidents are the leading cause of death in the 15-24 year age group.
- **Soft drinks and other fun drinks.** If they are an alternative to alcoholic drinks soft drinks should not be discouraged, but (unless sweetened with aspartame, etc) they provide only empty calories and by replacing milk can reduce the intake of calcium. Bottled pure water is a healthy trend.
- **High energy intakes.** Many adolescents go through a phase of eating much more than adults, sometimes up to 16.7 MJ per day (4000 kcal). This seems to occur near the age of peak height velocity in girls (around 12 years), but in boys may come later than the age of peak height velocity (usually 14 years). Presumably the larger, more muscular male adolescent is expending more energy at this stage.
- **Low levels of some nutrients.** Iron deficiency is quite common in adolescent girls who are menstruating, still growing, and often restricting their food intake. It may sometimes occur in boys too. Calcium accretion in the skeleton can be as much as 100 g/year at peak height velocity. Around 20% is absorbed so that about 500 grams per year are needed in the diet—that is, 1370 mg/day.
- **Adolescent dieters.** There are two aspects to this: overweight/obesity and social dieting. Obese adolescents are usually inactive and tend to have low socio-economic status. Dietary management should aim to hold the weight constant while the young person continues to grow and so thins out. Increased exercise should be emphasised and anorectic drugs should not be used.

Body Mass Index international cut off points for overweight (≡25 kg/m² in adults) and obesity (≡30 kg/m² in adults)[11]

Age	Overweight		Obese	
(years)	Boys	Girls	Boys	Girls
2	18.4	18.0	20.1	20.1
3	17.9	17.6	19.6	19.4
4	17.6	17.3	19.3	19.1
5	17.4	17.1	19.3	19.2
6	17.6	17.3	19.8	19.7
7	17.9	17.8	20.6	20.5
8	18.4	18.3	21.6	21.6
9	19.1	19.1	22.8	22.8
10	19.8	19.9	24.0	24.1
11	20.6	20.7	25.1	25.4
12	21.2	21.7	26.0	26.7
13	21.9	22.6	26.8	27.8
14	22.6	23.3	27.6	28.6
15	23.3	23.9	28.3	29.1
16	23.9	24.4	28.9	29.4
17	24.5	24.7	29.4	29.7
18	25	25	30	30

Reproduced courtesy of Steve Parry/ Impact

Percentage of daily nutrients provided by different meals among 290 school students (boys and girls) aged 16-17 in Sydney, Australia (based on 4-day records and interviews)[16]

	Breakfast	Lunch	Dinner	Snacks
Energy	16.7	23.8	33.0	26.1
Protein	16	25	41	18
Fat	15	25	37	23
Carbohydrate	20	23	27	31
Alcohol	0	14	36	50
Calcium	26	19	26	28
Iron	23	23	37	9
Vitamin A	18	26	36	21
Thiamin	28	22	28	20
Riboflavin	27	17	26	18
Vitamin C	14	19	27	27

As well as overweight youngsters there are adolescent girls of normal weight modifying their diet because they are not as thin as they or their peers think they should be. Some may fast and binge alternately. With a smaller energy intake they are more likely not to reach their requirements for iron and other essential nutrients. In a small minority this social dieting goes on to anorexia nervosa (incidence in some places as high as 1% of middle class girls aged 15-25) or bulimia. Treatment of anorexia nervosa is best handled by a specialised team of psychiatrist and dietitian. The general practitioner's main role is to recognise the early case. The longer the duration the worse the prognosis. A young woman whose weight goes below a body mass index (weight (kg)/height(m)^2) of 17 should be warned, with her parents, that her thinness is unhealthy and referred for treatment if she cannot put on weight. (See page 56 for more on eating disorders.) By contrast, adolescent boys are more likely to worry that they are not growing tall enough or not developing enough muscles.

Does diet affect acne?

The popular belief is that chocolate, fatty foods, soft drinks, and beer can all aggravate acne vulgaris. This is not surprising since 85% of people have acne at some time during adolescence and most adolescents eat and enjoy these foods.

Controlled trials—for example, of chocolate—have proved negative but their design can be criticised. It is very difficult to produce double blind conditions. Individual cases appear to respond to cutting down confectionery, fatty foods, or alcoholic drinks and there are other reasons to recommend such a dietary change. Zinc, polyunsaturated fats, and vitamin A are reported to improve acne, and adolescents can be advised to eat foods that are good sources of each: meat and wholemeal bread (zinc), polyunsaturated margarine or cooking oil, and (for vitamin A) carrots or liver.

Perspective

Patience, and a sense of humour help in watching and advising on adolescent food habits. The serious concerns are drinking with driving, usually in boys, and excessive slimming, usually in girls.

No young person wants to lose their teeth and spoil their good looks. As in children, sticky sugary foods should not be eaten between meals, or if they are, the teeth should be thoroughly brushed afterwards.

Parents have more influence than they may think. They can choose which foods and drinks they buy and prepare and keep in the refrigerator. The adolescent's food habits are laid down in the family and the family remains one influence. The other three are the peer group, the need to develop an independent personality, and society in general.

Adolescence is a transitional stage when the structure of food habits is loosened. In a few years the young person will usually get married, work out a compromise set of food habits with the spouse or partner,[17] and settle down to re-establish the eating behaviour of a new family. It is here, in preparing for and starting marriage that nutrition education should probably focus more.

Between the ages of 10 and 20 years:

- **lean body mass** goes from (average) 25 to 63 kg in boys, 22 to 42 kg in girls
- **body fat** goes from (average) 7 to 9 kg in boys, 5 to 14 kg in girls
- **triceps skinfolds (subcutaneous fat)** in boys stay at about 9 mm with a dip at year 15; in girls they climb steadily from 11 to 16 mm.

Too thin: entering anorexia nervosa range, BMI=17

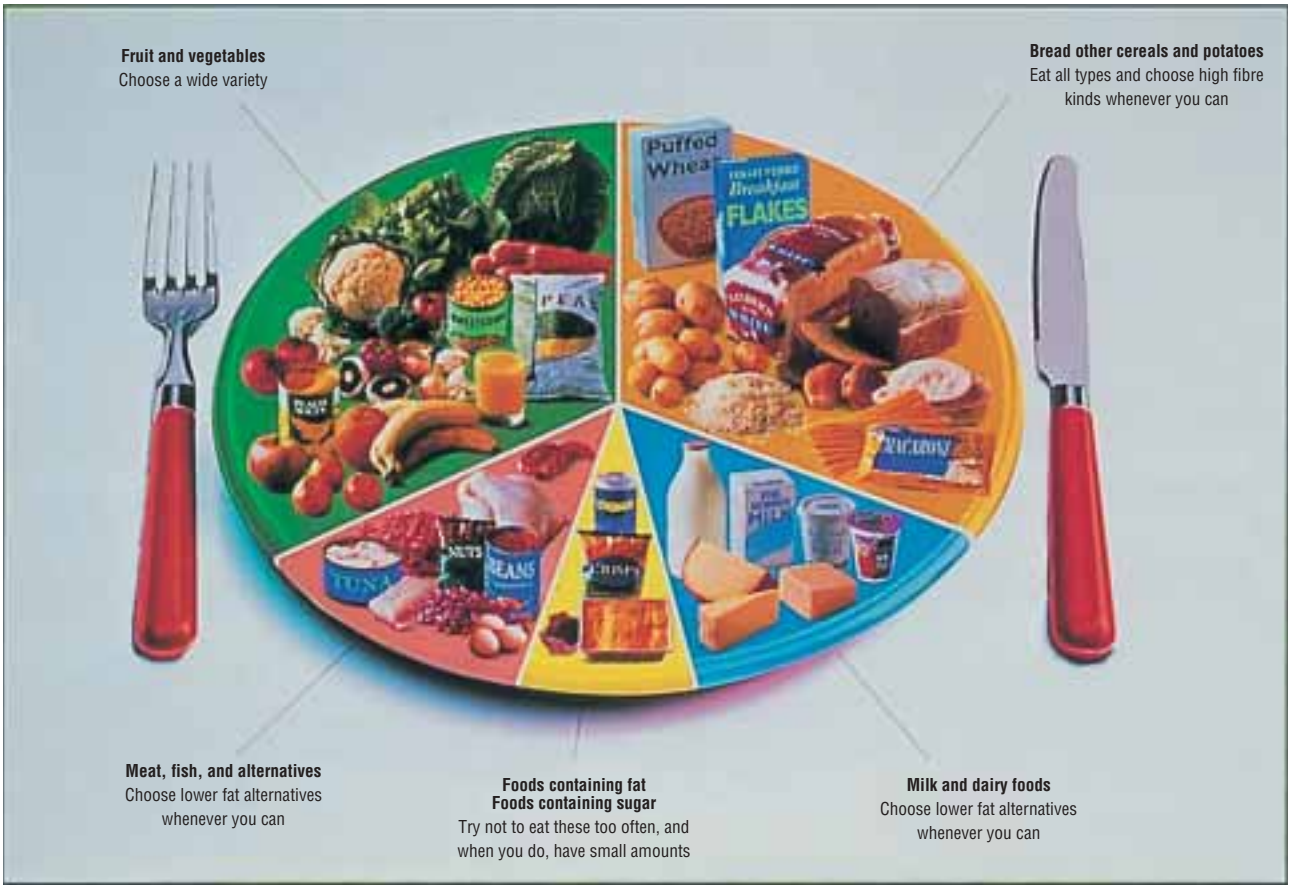
Height (no shoes)		Weight (min clothes)
Metres	Ft/inches	kg
1.45	4, 9	36
1.48	4, 10	37
1.50	4, 11	38.5
1.52	5, 0	39.5
1.54	5, 1	40.5
1.56	5, 1	41.5
1.58	5, 2	42.5
1.60	5, 3	43.5
1.62	5, 4	44.5
1.64	5, 5	45.7
1.66	5, 5	47
1.68	5, 6	48
1.70	5, 7	49
1.72	5, 8	50.3

Reproduced courtesy of Bruce Stephens/Impact

References

1 Health Education Authority. *Enjoy healthy eating*. London: Health Education Authority, 1995.
2 Gregory JR, Collins DL, Davies PSW, Hughes JM, Clarke PC. *National Diet and Nutrition Survey: children aged $1\frac{1}{2}$ to $4\frac{1}{2}$ years*, vol 1. *Report of the diet and nutrition survey*. London: HMSO, 1995.
3 Lawson M, Thomas M. Vitamin D concentrations in Asian children aged 2 years living in England: population survey. *BMJ* 1999; **318**: 28.
4 Bundred P, Kitchiner D, Buchan I. Prevalence of overweight and obese children between 1989 and 1998: population based series of cross sectional studies. *BMJ* 2001; **322**: 326-8.
5 Editorial. Childhood obesity: an emerging public-health problem. *Lancet* 2001; **357**: 1989.
6 Skuse D. Identification and management of problem eaters. *Arch Dis Childh* 1993; **69**: 604-8.
7 Green C. *Toddler training. A parents' guide to surviving the first four years*. Sydney: Doubleday, 1986.
8 Gregory J, Lowe S, Bates CH *et al. National Diet and Nutrition Survey. Young people aged 4 to 18 years, Vol 1. Report of the Diet and Nutrition Survey*. London: Stationery Office, 2000.
9 Nelson M. Childhood nutrition and poverty. *Proc Nutrition Soc* 2000; **59**: 307-15.
10 Whitehead RG, Paul AA, Cole TJ. Trends in food energy intakes throughout childhood from one to 18 years. *Hum Nutr Appl Nutr* 1982; **36A**: 57-62.
11 Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ* 2000; **320**: 1240-3.
12 Ludwig DS, Peterson KE, Gortmaker SL. Relation between consumption of sugar-sweetened drinks and childhood obesity: a prospective, observational analysis. *Lancet* 2001; **357**: 505-8.
13 Anderson RE, Crespo CJ, Bartlett SJ, Cheskin LJ, Pratt M. Relationship of physical activity and television watching with body weight and level of fatness among children. *JAMA* 1998; **279**: 938-42.
14 James WPT, McColl KA. *Healthy English schoolchildren: a new approach to physical activity and food*. Aberdeen: Rowett Research Institute, 1997.
15 Hamilton EMN, Whitney EN. *Nutrition: concepts and controversies*. St Paul, MN: West Publishing, 1979.
16 Truswell AS, Darnton Hill I. Food habits of adolescents. *Nutr Rev* 1981; **39**: 73-88.
17 Craig P, Truswell AS. Dynamics of food habits in newly married couples: who makes changes in the food consumed? *J Hum Nutr Diet* 1994; **7**: 347-61.

7 Adults young and old



UK food guide plate[1]

Most adults

Adults should eat enough of the essential nutrients by eating a healthy varied diet. This is more difficult to achieve if people have a low energy intake and if much of their diet consists of fats, alcohol, and sugar, which provide empty calories but little or no protein, or micronutrients.

Food guides

The food guide plate, with pictures of real foods, shows the four groups of foods that should between them provide enough of all the essential nutrients and the fifth group that provides mostly energy. (The plate picture makes a good poster for the waiting room.) The areas indicate relative amounts, so that people can see they should eat plenty of vegetables and fruit, plenty of cereal foods, and potatoes; moderate amounts of meat, fish and alternatives, and of dairy foods. The fifth group, "fatty and sugary foods", are foods and drinks on which to go easy. Other countries have food guides rather like this but use different shapes. The United States and Australia use (different) pyramids. Canada has a quarter rainbow.

It becomes complicated to specify amounts (in grammes or servings) because there is a big range of energy intake from small old ladies to large male athletes. UK health authorities, however, do recommend that adults aim to eat five (or more) servings of vegetables or fruit each day.[2]

Fats and Sugars leaflets. Reproduced with kind permission from the Food Standards Agency

HEA/MAFF/DoH eight guidelines for a healthy diet[3]

(1) *Enjoy your food*
No food needs to be excluded from the diet—except occasionally for special medical reasons. There is enough stress in life already without adding to it by worrying unnecessarily about food. It is good to be a little adventurous too with foods that have not previously been tried.

(2) *Eat a variety of different foods*
The food plate guide shows the types and proportions. Under this heading is a note on **salt:** eating too much can contribute to high blood pressure. Most people eat more salt than they need.

(3) *Eat the right amount to be a healthy weight*
Fat has more than twice as many calories as the same weight of starch or protein. Exercise does not have to be strenuous. Lots of small changes like walking … and using the stairs … can add up.

(4) *Eat plenty of foods rich in starch and fibre*
Foods like bread, other cereals (rice and pasta) and potatoes and legumes. Cooking potatoes in oil or fat greatly increases their calorie content.

(5) *Eat plenty of fruit and vegetables*
There are many biologically active substances in fruit and vegetables, as well as essential nutrients, that may help to reduce the risk of chronic diseases.

(6) *Don't eat too many foods that contain a lot of fat*
It is best to cut down on saturates as much as possible.

(7) *Don't have sugary foods and drinks too often*
Eating sugary foods frequently is the main cause of tooth decay. Children's teeth are the most susceptible.

(8) *If you drink alcohol, drink sensibly*
Up to 3-4 units* a day for men; 2-3 units* for women (except in pregnancy) will not significantly increase the risk to health over time.

*British units are 8 g of alcohol; US units are approximately 12 g of alcohol

US National Research Council 1989[4]

Reduce total fat to 30% or less of calories and saturated fat to less than 10% of calories and dietary cholesterol to less than 300 mg/day. Polyunsaturated fatty acid optimal intake 7% to 8% of calories (not over 10%). ω-3-polyunsaturates from regular fish consumption. (Concentrated fish oil supplements not recommended for general public).

Eat five or more servings of vegetables or fruits daily, especially green and yellow vegetables and citrus fruits, and six or more daily servings of bread, cereals, and legumes. Do not increase intake of added sugars.

Maintain a moderate protein intake, not more than twice the RDA.

Balance food intake and physical activity to maintain appropriate body weight.

If you drink alcohol limit it to no more than two standard drinks a day. Women who are pregnant or attempting to conceive should avoid alcoholic beverages.

Limit total salt intake to 6 g a day sodium chloride.

Maintain adequate calcium intake.

Avoid eating nutrient supplements with dose above the RDA (that is, avoid megavitamin supplements).

Maintain an optimal intake of fluoride, particularly during the years of primary and secondary tooth formation and growth.

RDA = recommended dietary allowance

Dietary guidelines for Americans[5]

- *Aim for a healthy weight.* If you are at a healthy weight aim to avoid weight gain. If you are already overweight (BMI $>25\,kg/m^2$) first aim to prevent further weight gain, and then lose weight to improve your health.
- *Be physically active each day.*
Choose activities that you enjoy and that you can do regularly. Be physically active for at least 30 minutes most days of the week. You can do it all at once or spread it over two or three times during the day.
- *Let the pyramid guide your food choices.*
That is eat most bread, cereal, rice and pasta; next level (not so much) vegetable group (including potatoes) and fruit group; next level (less again) milk, yoghurt, cheese group and meat, poultry, fish, dry beans, eggs, nuts group. The smallest group (use sparingly) is fats, oils and sweets.
- *Choose a variety of grains daily, especially whole grains.*
- *Choose a variety of fruits and vegetables daily.*
- *Keep food safe to eat.*
Safe means that the food poses little risk of foodborne illness from harmful bacteria, toxins, parasites, viruses or chemical contaminants.
- *Choose a diet that is low in saturated fat and moderate in total fat.*
Use moderate amounts of food high in unsaturated fat, taking care to avoid excess calories.
- *Choose beverages and foods to moderate your intake of sugars.*
In the United States the number one source of added sugar is non-diet soft drinks (soda or pop)… Drink water often.
- *Choose and prepare foods with less salt.*
Many people can reduce their chance of developing high blood pressure by consuming less salt … Most of the salt you eat comes from foods that have salt added during food processing.
- *If you drink alcoholic beverages, do so in moderation.*
Moderation here is defined as no more than one drink per day in women (that is, 12 g alcohol) and no more than two drinks per day in men (that is, 24 g alcohol).

Dietary guidelines

The box on page 38 shows three sets of dietary guidelines. These mostly give advice about how the part of the diet that provides energy—fats, carbohydrate, protein, etc,—should be made up. They are expressed in terms of averages. The average person, eating the average amount of (say) fat, should reduce this by or to a new national average. The advice has to be modified for individuals: Jack Spratt should not reduce fat intake as much as his wife. Recommendations from the US National Research Council use some numbers and technical terms. They need to be translated by health professionals for most consumers. The UK and US guidelines are written in non-technical language. The box shows the headings. There are more details in the booklets.[3-5]

Other aspects

Other aspects of a healthy diet, sometimes taken for granted, are that food should be wholesome and not contaminated with pathogenic micro-organsims or their toxins or with other toxins (chapter 14). Some of our foods are routinely enriched with micronutrients—for example, B vitamins and iron in white flour, vitamins A and D in margarine.

The municipal drinking water should likewise be nearly chemically pure. Calcium, magnesium, sodium, and fluoride concentrations in it vary from place to place. There is suggestive—but not conclusive—evidence that hard water (more calcium or magnesium, or both), is associated with lower cardiovascular mortality. Fluoridation at 1 ppm is recommended by all orthodox medical and dental authorities.

Vegetarianism

Are vegetarians more healthy or less? The answer depends first on the degree of vegetarianism.

Vegans, who eat no animal products, are at risk of vitamin B-12 deficiency. Supplements are essential during pregnancy and for infants of vegans. Vegans lack the best dietary sources of calcium—milk, yoghurt, and cheese.

Lacto-ovo-vegetarians have no absolute nutritional risk. They miss the best absorbed form of iron in the diet, haem iron, but may largely compensate because ascorbic acid enhances the absorption of non-haem iron.

The other determinant is the reason for the vegetarianism. People belonging to long traditions of vegetarianism have the necessary recipes to prepare vegetarian centres for their dishes, using legumes (including soya) and nuts, and so have a good protein intake. It is new vegetarians, some of whom simply remove meat from the centre of the plate, who may eat inadequately.

On the whole vegetarians appear to have lower risk of obesity, coronary heart disease,[7] hypertension, and possibly some cancers. However, many of the figures come from well documented groups such as Seventh Day Adventists, who have a more healthy lifestyle than average in other ways—for example, they do not smoke or drink alcohol.

Nutrition and poverty in industrialised countries

In different countries of the EU between 4% and 25% of people live below the poverty line (50% of the national average household expenditure, adjusted for family size).[8] In these subcultures of otherwise rich societies people are deprived of jobs, education, security, family support, adequate housing, transport, and/or language skills. The food and nutrients they eat and resulting nutritional status contribute to inequalities of health.[9]

Five grades of vegetarianism

- Do not eat meat of some animals (for example, horse) or some organs (for example, brain). *This is the norm for omnivorous humans except in a disaster.*
- Do not eat meat but eat fish (and dairy produce). *Very little nutritional risk.*
- Do not eat meat or fish but eat milk and eggs = lacto-ovo-vegetarian. *This is the most common degree of vegetarianism. The only nutritional risk is of iron deficiency.*
- Do not eat any animal products = vegan. *Vitamin B-12 deficiency is likely and can be very serious in infants.*[6] *Adequate protein can be ensured with regular legume products and nuts. Calcium, iron, and zinc nutrition should be watched.*
- Do not eat anything but fruit = fruitarian. *Unlike some primates, people cannot usually manage on such a diet for long and seldom try to. It is inadequate in protein (unless nuts are included) and even in sodium as well as the nutrients above.*

Supplements

- Most people do not need nutritional supplements, but women with high menstrual losses may benefit from iron supplements.
- Women who are, or who are planning to become, pregnant are advised to take 0.4 or 0.5 mg of folic acid each day and to eat folate-rich foods (chapter 4).
- Older house-bound people, or people who wear enveloping clothes out of doors all year round, do not make enough vitamin D for their needs by the action of sunlight on their skin and should take 7-10 μg of vitamin D daily.
- Vitamin drops are available for young children below the age of 5.
- Before starting to take a dietary supplement it is advisable for people to consult a medical practitioner.
- Someone drinking alcohol too heavily can at least prevent Wernicke's encephalopathy by taking thiamin (or vitamin B Co) tablets.

Sports nutrition[10]

- Dietary guidelines are the foundation of nutritional health, and competitive athletes are motivated to take their health seriously. There are no magic foods, no supplements which will improve sports performance on their own. Body weight and build differ for the various sports and events. Marathon runners are lightly built with little body fat. Champion weight lifters have a BMI in the "obese" range but this is due to an unusual amount of muscle, not fat. Athletes in regular training use and need more energy than sedentary folk. Compared with an average energy intake of around 10 MJ/day (2500 kcal), competitors during the Tour de France use a very high average of 33.7 MJ/day (8000 kcal).
- Athletes take various regimes of high carbohydrate intakes before major events to maximise muscle glycogen content and hence extend the duration of top performance. During long events, drinks similar to oral rehydration fluid (chapter 8) (with glucose, NaCl, and potassium) may sustain performance and prevent dehydration. Different sports drinks and schedules have their devotees.
- Two conditions are important in female athletes, including ballet dancers: mild iron deficiency anaemia, which limits performance, and osteopenia associated with amenorrhoea. For the latter, reduced training and/or increased body fat to restore menstruation, and calcium supplements are advised.

Low birth weight, lower rates of breast feeding, shorter children and overweight adults are more common in these people.

People in this group have to pay more for their food because they lack transport and funds to shop in large supermarkets. They consume less lean meat, fresh fruit and vegetables, and wholemeal bread than average. They suffer food insecurity at times. Their intake of micronutrients is lower, they are liable to anaemia, and their plasma lipids are higher than the average of the country. The Rome Declaration of World Food Security (1996) stated: "Governments will implement cost-effective public works, programmes for unemployed (and) develop social welfare and nutrition safety nets to meet the needs of the food insecure". In Britain the Low Income Project Team of the Nutrition Taskforce has published a report with far-ranging recommendations.[11] This problem of poor nutrition in deprived sections of prosperous countries is a failure of present capitalism and all who are trying to alleviate it deserve our support.

Older people

In developing countries children suffer most of the malnutrition. In developed countries it is the elderly who are most at risk of nutritional deficiency, though this is usually mild or subclinical and often associated with other disease(s). But it is very misleading to lump everyone over 65 together and expect them all to show the same problems and diseases. Healthy older people who are socially integrated are no more likely to get into nutritional trouble than anyone else.

For the majority people most of their life after 65 should be healthy and enjoyable. This "third age" is a time when people want to look after their health. They can now give more attention, time and money to getting and keeping healthy. They can take plenty of gentle exercise, have none of the stress of the workplace, few deadlines, and plenty of rest, and have time to choose food carefully and prepare it nicely. The dietary guidelines for younger adults all apply after retirement.

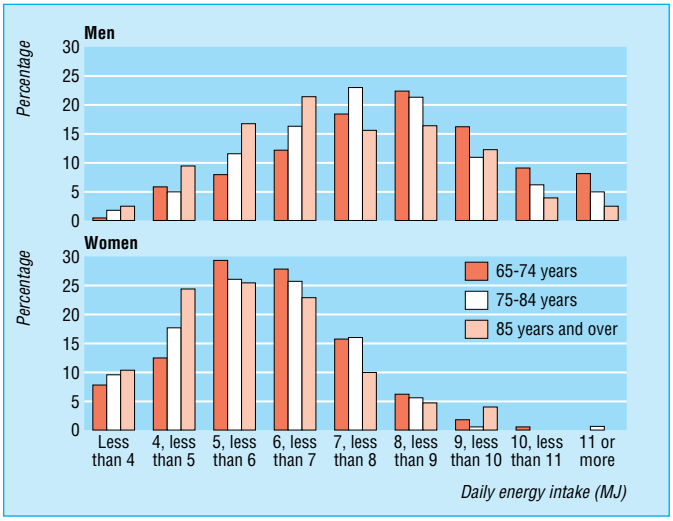
Dietary guidelines after retirement age

- A nutritious diet from a variety of foods is more important than when people are younger because the total energy intake is usually smaller than in young adults. The number of calories needed is less but not the requirements for most essential nutrients.
- To be light in weight eases the load on osteoarthritic joints and ageing heart and lungs and reduces the risk of accidents. Judicious regular exercise is much better than food restriction.
- Cut down on fat, especially saturated fat. Fat supplies more empty calories than any other dietary component. It predisposes to thrombosis, raised plasma cholesterol, and atherogenesis.
- Eat plenty of bread and cereals (preferably wholegrain) and vegetables and fruits; older people are liable to constipation, and a good intake of fibre will help to control this.
- Limit alcohol consumption. The smaller liver cannot metabolise as much alcohol as in young adults and the consequences of falls or accidents are more serious. No more than one or two drinks a day.
- Cut down on salt and salty foods. They tend to raise blood pressure; salt sensitivity increases with age and hypertension predisposes to strokes.
- Avoid too much sugar because of the empty calories, but dental caries is less troublesome in surviving mature teeth.

It is wrong, however, to use these guidelines for people who are declining in health (the "fourth age"); and some doubt even exists about whether low serum cholesterol concentration or body weight give any survival advantage to very old people.

Eating for the environment

People may rightly choose foods because the environment was considered in their production, for example more humane housing of hens and chickens (free-range eggs), fish nets that do not catch dolphins, or organic farming. The basis for preferring such produce has to be environmental rather than nutritional. Analysis of all the nutrients in foods adds up to be very expensive and no one can guarantee an overall nutritional advantage.[12]



Average daily intakes of 1275 free living older people in Britain (age 65-97)[13]

With ageing (from 20-30 to over 70 years):

- average body weight goes down after middle age (partly because of selective mortality of obese people)
- lean body mass declines from average 60 to 50 kg in men and 40 to 35 kg in women
- there is a loss of height and of mass of the skeleton
- muscle mass declines from about 450 g/kg to 300 g/kg
- body density goes down from 1.072 to 1.041 in men and from 1.040 to 1.016 in women
- body fat (as % of body weight) increases from about 20% to 30% in men and from 27% to 40% in women. It becomes more central and internal
- liver weight falls from about 25 g to 20 g/kg body weight
- basal metabolic rate goes down proportionally with lean body mass.

Diet and longevity 1[14]

- McCay in 1931, and others since, have found that rats and mice live longer if their diet is restricted to 60% of ad libitum intake long term. This is not because one type of food, fat or protein, is restricted and the effect still occurs after growth is completed. A favoured explanation is reduced exposure to reactive oxygen species from lower metabolic activity. Rats usually live about two years. Dietary restriction experiments have not yet been completed in larger animals. Whenever humans have been subjected to undernutrition they have usually been in an unhealthy environment with heavy exposure to infections but the diet-restricted rodents were kept away from pathogens. It is unlikely that a sufficiently long and controlled human experiment could ever be carried out.
- Humans are in any case an unusually long-lived species. We live twice as long as great apes. This may be because several antioxidant systems are unusually well developed in our species.[15]

Risk factors for impaired nutrition in older people

Social risk factors

- Loneliness
- Isolation
- Immobility (no transport)
- Poverty
- Ignorance (the widower who cannot cook)
- Bereavement
- Alcoholism
- Dependence
- Regression

Medical risk factors

- Cancer and radiotherapy
- Depression
- Chronic bronchitis and emphysema
- Anorexia
- Cardiac failure
- Angina
- Insomnia
- Blindness
- Deafness
- Paralysis
- Arthritis
- Dementia
- Gastrectomy
- Sjörgen's syndrome
- Diverticulitis
- No teeth
- Frailty

Some common drugs that can lead to malnutrition

- Aspirin and NSAIDs → blood loss, so iron deficiency
- Digoxin → lowers appetite
- Purgatives → potassium loss
- Cancer chemotherapy → anorexia
- Many diuretics → potassium loss
- Metformin → vitamin B-12 malabsorption
- Co-trimoxazole → can antagonise folate

NSAIDs = non-steroidal anti-inflammatory drugs

Some suggested extra dietary guidelines for older people

- Women especially should keep up a good intake of calcium from (low fat) milk or cheese, or both. This may help to delay osteoporosis.
- Those who are housebound or do not get out regularly should take small prophylactic doses of vitamin D (5-10 μg/day).
- Elderly people should avoid big meals, as Hippocrates, Galen, and Avicenna all advised. On the other hand, they should not miss any of the three main daily meals.

> Old men have little warmth and they need little food which produces warmth; too much only extinguishes the warmth they have.
>
> Hippocrates, *Aphorisms* 1, 14

- Coffee or tea in the evening may contribute to insomnia.
- There is a place for fatty fish or small amounts of fish oils containing fatty acids like eicosapentaenoic acid (20:5, ω-3), which can reduce the risk of thrombosis.

Assessment of nutritional status

There have been several studies of the nutrition of elderly people in Britain and similar countries. Findings have differed, partly because different sectors of the elderly population have been sampled, partly because different parts of the range of possible biochemical tests have been done. Nutritional deficiency is nearly always secondary to a social problem or to disease. The first step in diagnosing malnutrition is to recognise one or more **risk factors**. The chances of risk factors and of nutritional deficiencies in elderly people increase progressively from those at normal risk (the healthy and socially integrated), through those who have one or more chronic illnesses but are nevertheless socially organised, and the housebound, to those at high risk—the institutionalised.

Assessment of nutritional status can be difficult in old people. The history of food intake may be unreliable because of poor memory. We are not yet sure whether the recommended intakes of some nutrients should be adjusted downward or upward, or by how much. Height (stature) may be impossible to measure exactly because of deformities.

Nutrients most likely to be deficient in old people are (roughly in order of importance):

- total energy—thinness, wasting, undernutrition
- potassium—deficiency can present with confusion, constipation, cardiac arrhythmias, muscle weakness, etc
- folate—deficiency can present with anaemia or with confusion
- vitamin B-12—because of gastric atrophy. Serum methylmalonate may be elevated before vitamin B-12 is low
- vitamin D—deficiency can present with fractures or bone pains of osteomalacia
- water—frail old people may not drink enough, which can lead to urinary tract infection or dehydration
- dietary fibre—deficiency leads to constipation
- vitamin C—low plasma concentration, haemorrhages
- iron—anaemia, koilonychia
- protein—low plasma albumin, oedema
- calcium—low intake; decreased bone density
- zinc—low plasma concentration
- thiamin—biochemical features of deficiency (red cell transketolase)
- magnesium—low plasma concentration
- pyridoxine—biochemical features of deficiency.

Knee height (from heel to top of patella with the knee flexed) can be used to estimate stature in centimetres

- **for men** stature = 2.02 KH − (0.04 × age, y) + 64.2
- **for women** stature = 1.83 KH − (0.24 × age, y) + 84.9.

Weight-for-height standards (for younger adults) are not strictly applicable because of the decline in lean body mass. Some recent anthropometric data on 200 people over 75 living independently in two general practices are provided by Bannerman *et al*.[16]

Clinical examination is complicated by the presence of other diseases. Oedema is usually due to cardiovascular disease, loss of ankle jerks to ageing nerves rather than nutritional deficiency. For several biochemical tests we are not quite sure what the normal range is in old people.

Preventive measures

Only rarely are low intakes of nutrients and abnormal laboratory findings associated with a disturbance of function that would support the diagnosis of clinical malnutrition. The usual finding can best be called subclinical deficiency, and we are often not sure of its clinical importance. It is prudent to attempt to raise the level of nutrients to make people with subclinical deficiency more resistant to the effects of stress caused by non-nutritional diseases, which become increasingly common with advancing years.

General practitioners, with the younger family members or a friend, district nurse, or social worker can improve an old person's nutrition in several ways:

- suggest cooking lessons for retired men
- arrange help for partly disabled people to adapt cooking techniques
- organise delivery of heavy shopping
- suggest (where one is absent) buying a refrigerator or freezer
- ensure that every elderly person or couple has an emergency food store
- suggest that a younger relative helps with shopping and invites the elderly person for a regular good meal
- arrange for him or her to attend a lunch club
- arrange for meals on wheels
- possibly prescribe micronutrient supplements, but some multivitamin tablets do not contain them all (some miss folic acid) and there may be more need for potassium
- build on established eating patterns when advising about changing food consumption; drastic changes are likely to confuse
- warn that reduced sense of smell and sight make it hard to detect food that is no longer wholesome.

Nutrition in institutions

In nursing homes, practices that contribute to poor nutrition include:

- lack of communication between nursing and kitchen staff
- disregarding residents' suggestions about menus
- ignoring residents' requirements for special diets
- offering no choice of portion size or second helping
- monotonous menus
- not noting food left or residents not eating
- not weighing residents regularly
- little or no homestyle cooking
- inadequate help with feeding frail residents
- cooks that lack basic knowledge of nutrition
- meals rushed or served too early
- no facilities for residents to make hot drinks for themselves
- low fibre diets.

More details for those inspecting or advising on nutrition in nursing homes can be found in *Eating well for older people*.[18]

Diet and longevity: 2

- Claims that people live longer than usual in parts of Georgia (Europe), or in Vilcabamba in Ecuador have proved dubious on investigation. There were no birth certificates.
- The food that our centenarians ate when young they say was less processed and simpler than contemporary food patterns. Of course it was. That was what ordinary people ate in 1903.
- No special food emerged when a sample of centenarians recalled their lifestyles to the US House of Representatives Select Committee on Aging.

"Food, like sex, is one of the pleasures that stays with us all our lives" (as Alex Comfort put it).[17] When life is limited because of disabilities and loss of social and environmental stimulation, it is very important for the morale and well being of elderly people in institutions that their food is what they like, and is served with courtesy, and that they have at least some control over it. We younger adults are always making compromises between ideal nutrition and a little of what we fancy, and we must allow and help elderly people in nursing homes to do the same—by offering them a glass of sherry or beer, a favourite dessert, or a piece of confectionery.

References

1 Health Education Authority. *The balance of good health. Introducing the national food guide*. London: Health Education Authority, 1994.
2 National Heart Forum. *At least five a day. Strategies to increase vegetable and fruit consumption*. London: Stationery Office, 1997.
3 HEA/MAFF/DoH. *Eight guidelines for a healthy diet. A guide for nutrition educators*. Abingdon, Oxon: HEA Customer Services, 1997.
4 Committee on Diet and Health, Food and Nutrition Board, National Research Council. *Diet and health. Implications for reducing chronic disease risk*. Washington, DC: National Academy Press, 1989.
5 USDA, USDHHS. *Dietary guidelines for Americans*. Washington, DC: US Department of Agriculture, 2000.
6 Wighton MC, Manson JI, Speed I *et al*. Brain damage in infancy and dietary vitamin B-12 deficiency. *Med J Aust* 1979; **2**: 1-3.
7 Thorogood M, Mann JI, Appleby P, McPherson K. Risk of death from cancer and ischaemic heart disease in meat and non-meat eaters. *BMJ* 1994; **308**: 1667-71.
8 Dowler EA, Dobson BM. Nutrition and poverty in Europe: an overview. *Proc Nutr Soc* 1997; **56**: 51-62.
9 James WPT, Nelson M, Ralph A, Leather S. The contribution of nutrition to inequalities in health. *BMJ* 1997; **314**: 1545-9.
10 Burke L, Deakin V (eds). *Clinical sports nutrition*, 2nd edn. Sydney: McGraw Hill, 2000.
11 Department of Health. *Low income, food, nutrition and health: strategies for improvement*. Wetherby: Department of Health, 1996.
12 Williams CM. Nutritional quality of organic food: shades of grey or shades of green. *Proc Nutr Soc* 2002; **61**: 19-24.
13 Finch S, Doyle W, Lowe C *et al*. *National Diet and Nutrition Survey: People aged 65 years and over. Vol 1. Report of the diet and nutrition survey*. London: Stationery Office, 1998.
14 Masoro EJ. McCay's hypothesis: undernutrition and longevity. *Proc Nutr Soc* 1995; **54**: 657-64.
15 Cutler RG. Antioxidants and aging. *Am J Clin Nutr* 1991; **53**: 373-79S.
16 Bannerman E, Reilly JJ, MacLennan WJ, Kirk T, Pender F. Evaluation of validity of British anthropometric reference data for assessing nutritional state of elderly people in Edinburgh: cross sectional study. *BMJ* 1997; **315**: 338-41.
17 Comfort A. *A good age*. London: Mitchell Beazley, 1977.
18 Report of an Expert Working Group. *Eating well for older people. Practical and nutritional guidelines for food in residential and nursing homes and for community meals*. London: The Caroline Walker Trust, 1995.

8 Malnutrition in developing countries

Some former "developing countries" really have been developing, and as a byproduct malnutrition has largely disappeared. Other countries have been static or lost ground. A country can become poorer from reduced income, reduced gross national product (GNP) per head (caused by bad climate, economic mismanagement, or war) or from population growth faster than economic growth, or both. Many of the poorest countries have suffered civil wars.

About two-thirds of the world's six million people live in countries with low and lower middle incomes (on the current lists of the United Nations). In most of these countries people are very poor; the population is young and growing fast; there is no welfare state and little mechanisation. Food at an affordable price cannot be taken for granted; nor can clean drinking water. Tropical infections are an additional burden.

Public health indicators do not correlate closely with national income. Vietnam is an outstanding example of a poor country which nevertheless has enough food to go round, and health statistics better than that, for example, of South Africa, whose GNP/head is 10 times higher. Economic development is thus only one factor—an important one—that reduces malnutrition. But even if a country's income stays low there are things that doctors, nurses, agriculturists, administrators, and politicians can do to combat malnutrition.

Diagnosis and management of malnutrition in developing countries have to be mostly a public health operation. Many of the malnourished live in slums, shanty towns, or remote rural areas. They cannot be brought to a central teaching hospital. There are fewer doctors—in some countries only 1 for 50 000 people—so they have to work through teams of community health workers, who should be trained to recognise and cope with the common diseases, including malnutrition and the closely related infections.

1998 Figures	39 Poorest countries	Vietnam	United Kingdom
GNP/head (US$)	110 to 390	310	20 870
Population growth %	−0.7 to 4.7	1.9	0.2
Infant mortality Rate/1000	31 to 182	31	6
Under 5 mortality Rate/1000	42 to 316	42	6
Life expectancy at birth (years)	38 to 68	68	77

Data from *The State of the World's Children* 2000[1]

From the first three paragraphs of the World Declaration on Nutrition, 1992[2]

- We, the Ministers and Plenipotentiaries representing 159 states and the EEC … declare our determination to eliminate hunger and to reduce all forms of malnutrition. Hunger and malnutrition are unacceptable in a world that has both the knowledge and the resources to end this human catastrophe …
- … We all view with the deepest concern the unacceptable fact that about 780 million people in developing countries—20% of their combined population—still do not have access to enough food to meet their basic daily needs for nutritional well-being.
- We are especially distressed by the high prevalence and increasing numbers of malnourished children under five years of age in parts of Africa, Asia and Latin America and the Caribbean. Moreover, more than 2000 million people, mostly women and children, are deficient in one or more micronutrients: babies continue to be born mentally retarded as a result of iodine deficiency; children go blind and die of vitamin A deficiency; and enormous numbers of women and children are adversely affected by iron deficiency…

Protein-energy malnutrition[3,4]

What is described here is malnutrition in young children. They are dependent on adults for their food and are therefore especially vulnerable where there is food insecurity. They have high food energy needs for their size (kg body weight) and because of higher protein requirements per calorie (or kJ) they are more at risk of protein deficiency than adults.

The prevalence of protein-energy malnutrition (PEM) in its various forms is high in South and South-East Asia, in Africa and the Middle East, in some Caribbean islands, and in Central and South America. Severe forms affect around 2% and mild to moderate PEM affects around 20% of young children (in many places more) in developing countries. WHO has estimated that about 200 million children in the world at any time have moderate or severe PEM. Although it affects only some children in each community, it is a larger and more intractable problem than famines.

Protein-energy malnutrition	Bodyweight as % of standard	Oedema	Deficit in weight for height
Marasmus	60	0	++
Marasmic kwashiorkor	60	+	++
Kwashiorkor	80-60*	+	+
Nutritional dwarf	60	0	minimal
Underweight child	80-60	0	+

*Occasional cases are not underweight at the oedematous stage

Severe PEM

Nutritional marasmus is the commonest severe reform of protein-energy malnutrition, the childhood version of starvation. It usually occurs at a younger age than kwashiorkor. The cause is a diet very low in both calories and protein—caused, for example, by early weaning then feeding dilute food because of poverty or ignorance. Poor hygiene leads to

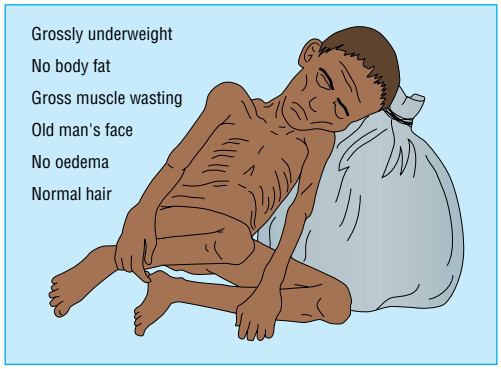
gastroenteritis and a vicious circle starts. Diarrhoea leads to poor appetite and more dilute foods. In turn further depletion leads to intestinal atrophy and more susceptibility to diarrhoea.

Kwashiorkor in its full blown form is less common than marasmus. It is most common in poor rural children, displaced from the breast by the next child and given a very low protein starchy porridge—for example, made with cassava or plantain. There have been several hypotheses about the antecedent diet because it is very difficult to reconstruct the exact dietary history of a malnourished child. But careful studies by Whitehead's group[5] of pre-kwashiorkor in Uganda compared with pre-marasmus in the Gambia, and other information support the classical hypothesis of protein deficiency with relatively adequate carbohydrate intake. Pure cases of kwashiorkor can develop in a few weeks and the patients sometimes have normal weight for age.

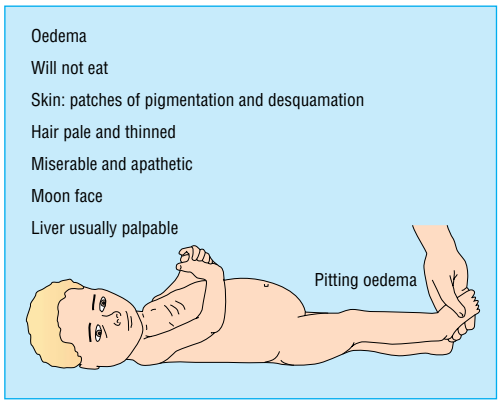
The pathogenesis of kwashiorkor appears to be: very low protein intake with more dietary carbohydrate leads to insulin secretion being maintained (unlike marasmus). Insulin spares muscle protein when there is shortage of amino acids, but there is loss of liver protein. So synthesis is reduced of two proteins made in the liver: (*a*) plasma albumin, hence oedema (potassium depletion makes it more likely), and (*b*) low density lipoproteins, hence lipids accumulate in the liver. Some of the features of kwashiorkor may be due to associated zinc deficiency.

Marasmic kwashiorkor has some features of both conditions. Severe protein-energy malnutrition can be thought of as a spectrum from marasmus to kwashiorkor. Most affected children have some skin lesions, hair changes, and fatty liver (as in kwashiorkor) together with the wasting of marasmus.

Malnourished children are likely to be depleted in other nutrients (see box opposite).



Marasmus



Kwashiorkor

Treatment

Management of severe protein-energy malnutrition is in three phases.[6]

(1) **Resuscitation**
Correction of dehydration, electrolyte disturbances, acidosis, hypoglycaemia, hypothermia, and treatment of infections.

(2) **Start of cure**
Refeeding, gradually working up the calories (from 100 to 150 kcal (420-630 kJ) per kg) and protein (to about 1.5 g per kg). There may be anorexia, and children often have to be hand fed, preferably in the lap of their mother or a nurse they know. Potassium, magnesium, zinc, and a multivitamin mixture are needed but iron should not be given for the first week.

(3) **Nutritional rehabilitation**
After about three weeks if all goes well the child has lost oedema and the skin is healed. The child is no longer ill and has a good appetite but is still underweight for age. It takes many weeks of good feeding for catch up growth to be complete. During this stage the child should be looked after in a convalescent home or by its mother, who should if possible have been educated about nutrition and provided with extra food. Locally available foods are best.

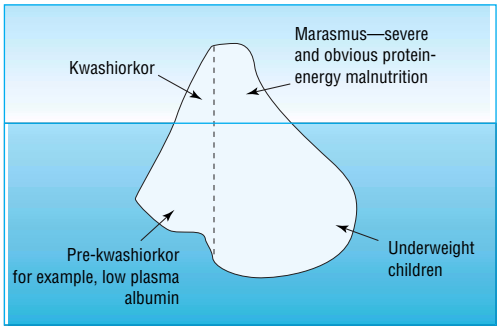
Other nutrients deficient in protein-energy malnutrition

Usually:
- Potassium
- Magnesium
- Zinc
- Vitamin A
- Iron
- Folate

In some areas:
- Thiamin—Thailand
- Riboflavin—Thailand
- Niacin—southern Africa
- Iodine—areas of endemic goitre
- Selenium

Mild and moderate protein-energy malnutrition

This is much more common than the obvious severe forms. Outside observers, even the mothers themselves, do not notice most of these cases because the children are similar in size and vitality to some of the other children of the same age. The condition is like an iceberg. For every severe case there are likely to be seven to 10 in the community with lesser degrees of malnutrition. These latter children do not grow normally and



The iceberg of malnutrition

are at increased risk of infection. They may also have more difficulty learning motor and cognitive skills.

National statistics used to indicate nutrition[1]

Underweight can be due to either wasting or stunting. **Wasting** is a fairly direct indication of undernutrition but the causes of **stunting** are complex. The NCHS anthropometric references used by WHO for international comparisons are based on large samples of US children measured in the 1960s and 1970s. A high percentage of stunted children means that they are short by US standards. The causes can be genetic, or intra-uterine growth retardation, or delayed growth from multiple infectious diseases, or insufficient nutrition. Small size is an adaptation. It cannot be equated with malnutrition.[7]

Most children with mild protein-energy malnutrition are thin and underweight. Because scales are difficult to carry children who are malnourished may be identified in the field by measuring the mid-upper arm circumference. From 12 to 60 months of age over 13.5 cm is a normal circumference; 12.5 to 13.5 cm suggests mild malnutrition and under 12.5 cm indicates definite malnutrition. The normal circumference stays the same for these four years.

Sometimes a child is seen who has adapted to chronic inadequate feeding by reduced linear growth but looks like a normal child a year or two younger—this is **nutritional stunting**.

Prevention of protein-energy malnutrition

Five measures to prevent protein-energy malnutrition are being actively promoted round the world.

Growth monitoring. The WHO has devised a simple growth chart—the Road to Health card. The mother (not the clinic) should keep the card in a cellophane envelope and bring the child (plus card) to the nearest clinic regularly for weighing and advice.

Oral rehydration. The UNICEF formula is saving many lives from gastroenteritis: NaCl 3.5 g, $NaHCO_3$ 2.5 g, KCl 1.5 g, glucose 20 g (or sucrose 40 g) and clean water to 1 litre.

Breast feeding is a matter of life and death in a poor community with no facilities for hygiene. Additional food, prepared from locally available products, is needed from four to six months of age.

Immunisation should be done against measles, tetanus, pertussis, diphtheria, polio, and tuberculosis.

Family planning advice and inexpensive or free contraception should be readily available.

Starvation and famine

When there is not enough food for an entire community children stop growing, and children and adults lose weight. The symptoms include craving for food, thirst, weakness, feeling cold, nocturia, amenorrhoea, and impotence.

The face at first looks younger but later becomes old and withered and expressionless; pupils react poorly to light. The skin is lax, pale, and dry and may show pigmented patches. Hair becomes thinned or lost except in adolescents. The extremities are cold and cyanosed. There may be pressure sores. Subcutaneous fat disappears, skin turgor is lost, and muscles waste. The arm circumference is subnormal. Oedema may be present; in adults this is famine oedema, which is not always associated with hypoalbuminaemia. Temperature is subnormal. The pulse is slow, blood pressure low, and the heart small with muffled sounds. The abdomen is distended. Diarrhoea is common, often associated with blood. Muscles are weak and tendon jerks diminished. Psychologically, starving

Underweight, wasting and stunting, defined by WHO

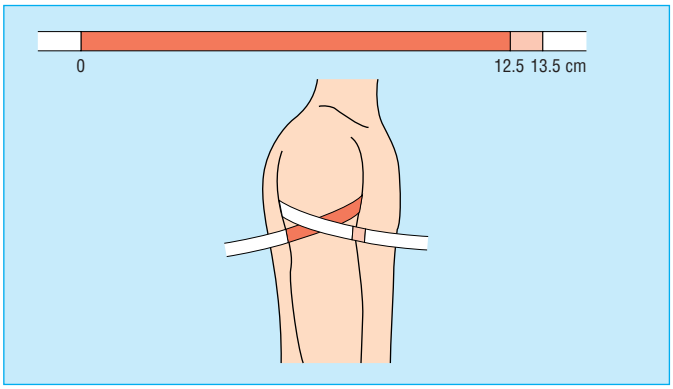
- Underweight is moderate if 2 to 3 standard deviations below the median NCHS **weight for age** (see page 30) and severe if more than 3 SDs below the reference.
- Wasting is moderate if 2 to 3 standard deviations (79-70%) below the median NCHS **weight for height** and severe if more than 3 SDs (<70%) below the reference.
- Stunting is moderate if 2 to 3 standard deviations (89-85%) below the NCHS median **height for age** and severe if more than 3 SDs (<85%) below the reference.

The standard deviation score is also called the "z score".

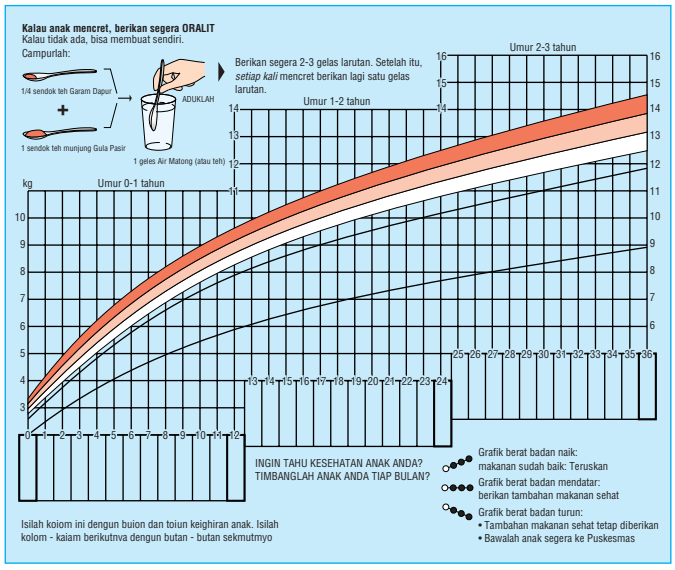
Percentage of children under 5 with moderate or severe underweight, wasting or stunting

	Under weight	Wasting	Stunting
Sub Saharan Africa	32	9	41
Middle East and North Africa	18	8	25
South Asia	51	18	52
Latin America and Caribbean	10	3	18
USA	1	1	2

From UNICEF *State of the World's Children* 2000[1]



Measuring circumference of mid-upper arm



Road to Health card for Indonesia with instructions for making oral rehydration fluid. Numbers along the base are months

people lose initiative; they are apathetic, depressed, and introverted but become aggressive if food is nearby.

Infections are to be expected, especially gastrointestinal infections, pneumonia, typhus, and tuberculosis. The usual signs of infection (pyrexia, leucocytosis) may not appear. Delayed skin sensitivity with recall antigens—for example, tuberculin—are falsely negative. But the erythrocyte sedimentation rate is normal unless there is infection. In advanced starvation patients become completely inactive and may assume a flexed, fetal position. Death comes quietly and often quite suddenly in the late stage of starvation. The very young and the very old are most vulnerable.

Inside the body plasma free fatty acids are increased; there is ketosis and may be a mild metabolic acidosis. Plasma glucose is low but albumin concentration is often normal. Insulin secretion is diminished, reverse triiodothyronine replaces normal T3, and glucagon and cortisol concentrations tend to increase. The resting metabolic rate goes down considerably; oxygen consumption per person goes down more than when expressed per kg body weight. The urine has a fixed specific gravity, and creatinine excretion becomes as low as 300 mg/day. There may be a mild anaemia, leucopenia, and thrombocytopenia. The electrocardiogram shows sinus brachycardia and low voltages. All the organs are atrophied and have subnormal weights at necropsy except the brain, which tends to maintain its weight.

Much the same clinical and metabolic features of starvation are seen in hunger strikers as in the much more common situation in a famine, but in the latter intercurrent infections usually compound the disorder. The problem in a famine is not so much loss of food availability as loss of food entitlement. People have to sell all their assets in the attempt to buy food. Practically all social and economic structures break down and there may as a last resort be mass migration of the sufferers. The worst famines of recent times have been in areas torn by civil war. This greatly hampers communication of early warning of food shortage and transport of relief food into the area.

Any doctor involved in relief operations should expect to have a mainly administrative and organisational role. It is impossible to give most time to treatment of a few very sick individuals. Therapeutic feeding is not an effective use of resources. Field workers have three options for distribution of food where supplies are insufficient to provide the minimum requirements of 1900 kcal (8MJ)/person/day:

(1) where community and family structure is still intact and community representatives can be identified, let the community decide how the limited food should be distributed
or
(2) where community structures have been disrupted field workers distribute food selectively to those at highest risk of mortality
or
(3) the third alternative is equitable distribution of the same basic ration to all members of the affected population with selection of particularly vulnerable members.[9]

Each option has its unsatisfactory aspects and work is easier when and if enough food is shipped in for all. The standard rations usually consist of cereals, legumes, and some oil. If the cereal is wholegrain, milling equipment is needed. Milk powder is used for malnourished children. Care must be taken that the population is getting the critical micronutrients, for example, vitamin C, potassium.[10]

To assess the degree of undernutrition in individuals two measures are used: mid-upper arm circumference (MUAC) and

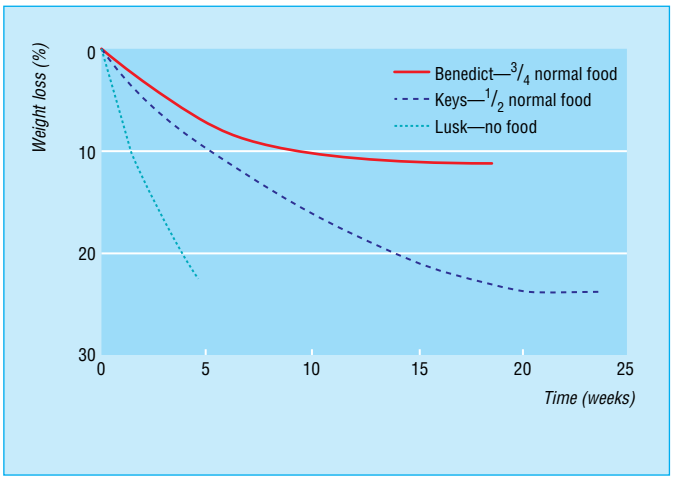
Famine is different from endemic undernutrition

Because present-day famines strike in developing countries where malnutrition is endemic, there is an unfortunate tendency to blur the differences between endemic malnutrition and famine-induced malnutrition, to treat the latter as if it were simply the former writ large …

Famines are distinct … They are different not only in severity, but in kind. This is because the famine year is neither characterised by poverty, nor even death, but by social disruption. Miserable though it is, chronic poverty in traditional societies is a situation to which considerable social, psychological and physiological adaptation has occurred. Only when these mechanisms of cultural homoeostasis are unable to cope does the situation shift into famine …

What distinguishes famine-induced malnutrition is not that it is *acute*, but that it is *extensive*.

Rivers JPW, in Harrison[8]



Body weight loss during fasting. (Adapted from Payne[7])

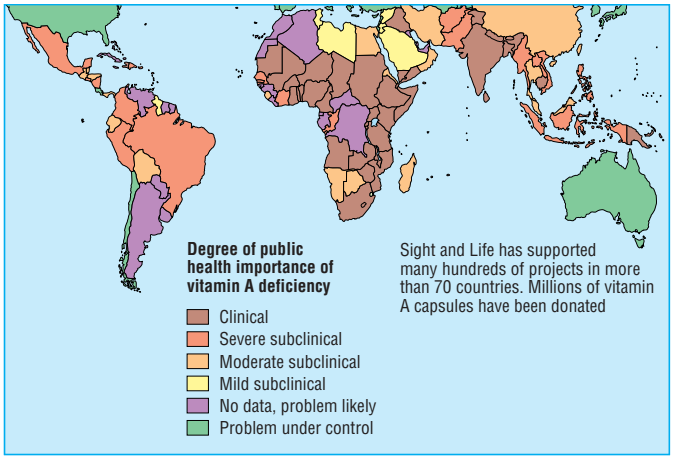
Degrees of underweight in adults at different heights

Height (m)	Weight (kg)	
	80% of standard*	70% of standard
1.45	38	33
1.48	39.5	34.5
1.50	40.5	35
1.52	41.5	36
1.54	42.5	37
1.56	44	38.5
1.58	45	39
1.60	46	40
1.62	47	41
1.64	48.5	42
1.66	49.5	43
1.68	50.5	44
1.70	52	45.5
1.72	53	47
1.74	54.5	48
1.76	56	49
1.78	57	50
1.80	58	51
1.82	60	52
1.84	61	53.5
1.86	62	54
1.88	63.5	55.5
1.90	65	57

*Standard weight = W/H^2 22.5. 80% standard = BMI 18
70% standard = BMI 16

weight for height in children or BMI (kg/m^2) in adults. MUAC is obviously quicker and a tape measure more portable, but in children low MUAC tends to select younger children as malnourished and miss older children with low weight for height.[11] In adults MUAC and BMI appear to correlate fairly well. A MUAC of 220 mm in men or 210 mm in women corresponds approximately to a critical BMI of 16 kg/m^2.[12] As a general rule **moderate starvation**=weight for height 80-71% of standard (BMI 18-16 kg/m^2) and **severe starvation**=weight for height ≤70% of standard (BMI ≤ 15.7).

Circumstances and resources are different in every famine. The problems are mainly non-medical: organising transport and repair of trucks and shelters, coordinating relief from different organisations, reconciling international workers with local politicians and administrators, arranging security of food stores, seeing that food is distributed on the basis of need, trying to procure the right food and the appropriate medical supplies. Civil disturbances do not occur during severe famine. They may happen at an early stage (food riots) or afterwards (revolution). Meanwhile, the future has to be planned for; agricultural workers are going to be needed with enough strength to plough and plant the next crop when the rains return.



Global map of VADD. Adapted with permission from *Sight and Life* (http:// www.sightandlife.org)

Vitamin A deficiency and xerophthalmia

In 1857 David Livingstone first suggested that eye lesions in some African natives were caused by nutritional deficiency: "The eyes became affected as in the case of animals fed pure gluten or starch." The antixerophthalmia factor was the first of the vitamins to be isolated, in 1915 by McCollum in the USA. Xerophthalmia is a late manifestation of vitamin A deficiency. Its global incidence has been estimated at some 500 000 new cases a year, half of which lead to blindness. Because of its social consequences vitamin A deficiency is given priority by the WHO for prevention programmes. The highest incidence is in South and South-East Asia—for example, India, Bangladesh, and the Philippines. It also occurs in some underdeveloped parts of Africa and Central and South America.

Vitamin A is not only the antixerophthalmia vitamin. It has also been called the "**anti-infective vitamin**" because rats with experimental deficiency had multiple infections. Children with full-blown xerophthalmia have high mortality. The importance of vitamin A for humans broadened when a longitudinal survey in Java showed that even mild forms of xerophthalmia were associated with a four-fold risk of death, often from respiratory or intestinal infections. Sommer's group[13] went on to a large randomised controlled trial in an area where cases of xerophthalmia occur. Children of 1-5 years given vitamin A capsules (one single capsule of 200 000 IU, repeated after six months) had a 34% lower mortality than untreated children in adjacent villages.[13] Subsequently similar prevention trials have been completed in several developing countries. Most reported significant benefits from vitamin A supplementation; the overall average reduction of death rate was 23%. By 1996, 40 developing countries had programmes for giving routine vitamin A supplements to young children.[1]

There is a strong synergistic association of measles and vitamin A deficiency. Measles can precipitate xerophthalmia and leads to low plasma vitamin A even in developed countries. In Cape Town (where clinical xerophthalmia is rare) a controlled trial in black children hospitalised with measles showed a strikingly better outcome in those given vitamin A (200 000 IU once on admission and repeated the next day).

Child receiving vitamin A capsule. Reproduced with permission from *Sight and Life*

Vitamin A should be given to any child with severe measles or from a deprived background.[14] With the new broader concepts of subclinical vitamin A deficiency, over 200 million young children have been estimated to be at risk and three million clinically affected.[15]

Stages of xerophthalmia

Severe xerophthalmia is virtually confined to infants and young children and usually associated with protein-energy malnutrition. The stages are classified by the WHO as follows.

- **Night blindness** (XIN) is the earliest symptom but not elicited in infants.
- In **conjunctival xerosis** (XIA) one or more patches of dry non-wettable conjunctiva emerge "like sand banks at receding tide" when the child ceases to cry. It is caused by keratinising squamous metaphasia of the conjunctiva.
- **Bitot's spots** (XIB) are glistening white plaques formed of desquamated thickened epithelium, usually triangular and firmly adherent to the underlying conjunctiva.
- **Corneal xerosis** (X2) is a haziness or a granular pebbly dryness of the cornea on routine light examination, beginning in the inferior cornea.
- **Corneal ulceration** (X3A) **or keratomalacia** (X3B). A punched out ulcer may occur or, in a severe case, colliquative necrosis of the cornea (keratomalacia). If promptly treated a small ulcer usually heals, leaving some vision. Large ulcers and keratomalacia usually result in an opaque cornea (X5) or perforation and phthisis bulbae.

Bitot's spot. Reproduced with permission from *Sight and Life*

Pathogenesis of xerophthalmia

In countries where xerophthalmia occurs, adults have much lower vitamin A stores in their livers than in well-fed people. Women start with low stores and throughout pregnancy have low intakes of vitamin A and carotene. Newborn babies have only one-fifth the liver vitamin A concentration of their mothers, even in well-fed communities, because vitamin A transport across the placenta is limited. Since the mother has low intakes and stores, her breast milk contains low concentrations of vitamin A and carotene. If the child has protein-energy malnutrition this impairs absorption and transport of vitamin A. Then a severe infection can precipitate clinical deficiency by increasing urinary loss of the vitamin and reducing hepatic synthesis of retinol binding protein.

The symptomatology of vitamin A deficiency can be explained by several functions of the vitamin. Retinaldehyde (retinal) is needed for the response of rods in the retina to light. Retinoic acid is needed to maintain differentiation of epithelia (for example, conjunctiva, respiratory), secretion of mucus, and tear production. In vitamin A deficiency cell-mediated immunity is impaired.

Colliquative necrosis of lower 2/3 of the cornea (Keratomalacia). Reproduced with permission from *Sight and Life*

Diagnosis and treatment

Xerophthalmia is rare in Britain. It is seen occasionally in patients with chronic jaundice or small bowel resection or very restricted diets. A British doctor going to work in a developing country should familiarise themself with the early features of xerophthalmia from colour photographs. (An excellent set is obtainable from American Foundation for Overseas Blind, 22 West 17th St, New York, NY 10011, USA. Or see a recent WHO publication on diagnosis and management of xerophthalmia.)

Treatment of xerophthalmia is urgent. The differential diagnosis includes smoke exposure, trauma, bacterial infections, measles, and trachoma. The child often has some other illness at the time like gastroenteritis, kwashiorkor, measles, or respiratory infection, which can distract attention from the eyes unless they are examined systematically. If in doubt a dose of vitamin A

In Britain conscientious objectors volunteered for a classic vitamin A depletion experiment in Sheffield during the Second World War. There were no features of deficiency until they were into the second year of a diet lacking vitamin A and carotene, and even then the symptoms were minor (night blindness and follicular hyperkeratosis). In well-nourished countries adults have enough vitamin A stored in their liver to last over a year of deprivation.

should be given. It can do no harm. The immediate treatment is 110 mg retinol palmitate or 66 mg retinol acetate (200 000 IU) orally or (if there is repeated vomiting or severe diarrhoea) 55 mg retinol palmitate (100 000 IU) **water soluble** preparation intramuscularly. For the next few days repeat the oral dose.

Prevention

There are four strategies for prevention. In some countries two or more are being used side by side.

(1) **Nutrition education**
This emphasises garden cultivation and regular consumption of locally grown plant sources of β-carotene (pro-vitamin A). The best sources include mango, papaya, pumpkin, yellow sweet potatoes, carrots and palm oil, as well as eggs and liver. Dark green leafy vegetables, formerly encouraged, contain useful amounts of β-carotene but this was found to be poorly absorbed (and converted to vitamin A) with traditional cooking methods. β-Carotene in plant leaves is mostly in the chloroplasts, which are not well digested. The carotene in fruits that contain it are more available[16] and absorption of β-carotene is improved if there is oil or fat in the meal.

(2) **Vitamin A for mothers**
The vitamin may be given to pregnant women, but it must not exceed 3300 IU (1 mg retinol) per day (or 23 300 IU once a week) because more vitamin A can be teratogenic. After delivery large single oral doses (200 000 IU) can be given to them in the first month. It should not be given later in case they become pregnant again.

(3) **Periodic dosing of young children**
This should be done in areas of high incidence with capsules of 110 mg retinol palmitate or 66 mg retinol acetate (200 000 IU) at six monthly intervals. Doses must be smaller in infancy.

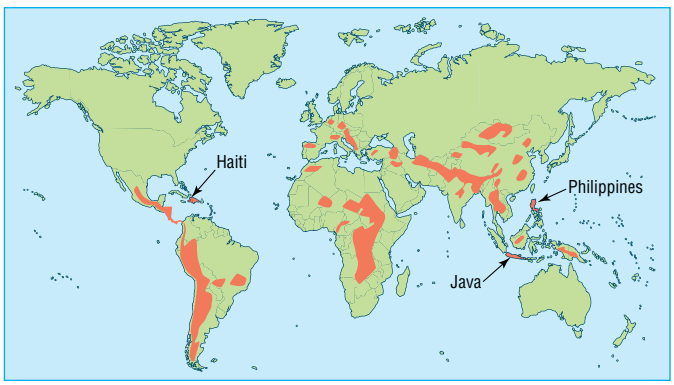
(4) **Fortification of staple foods with vitamin A**
In industrialised countries vitamin A is added to margarines to the level found in summer butter (2500 IU or 0.75 mg retinol per 100 g). In Central America sugar is fortified; the World Food Programme requires dried skim milk used in its aid schemes to be fortified with vitamin A.

Indian education poster (the plate contains green vegetables and orange-coloured fruits in additions to the staple, rice)

Iodine deficiency disorders (IDDs)[17]

Iodine deficiency disorders are also given priority by WHO for preventive efforts among nutritional diseases because of their extent—about 1500 million people live in iodine-deficient environments—and feasibility of prevention. Their social importance is greater than was formerly realised. In the major inland mountainous areas of the world, especially the Himalayas, the Alps, the Andes, inland mountainous areas of China and Africa, Indonesia and Papua New Guinea, the soil has had its original iodine leached out of it by heavy rainfall or glaciation, so that the human diet is lacking in iodine if people rely on locally grown foods. When the iodine intake is below the minimum (about 50-75 μg/day) required to replace the turnover of thyroid hormones, pituitary thyrotrophin secretion increases and the thyroid takes up more than its usual 50% of absorbed iodine. Hypertrophy of the gland develops—a goitre. The prevalence of goitre was estimated at about 200 million people worldwide in the early 1990s.

When just visible goitres occur in at least 5% of adolescents this is defined as **endemic goitre**. It usually shows first at puberty, and women are more affected than men. In some areas the iodine intake, indicated by the 24-hourly urinary iodine, is not very low and endemic goitre is attributed partly to thyroid antagonists such as glucosinolates or thiocyanate in certain brassicas or in cassava or soya beans.



Areas of endemic goitre

When endemic goitre occurs in almost all the women a small percentage of babies, 1% up to 5%, are born with **cretinism**. There are two types. In nervous cretinism there is mental deficiency, deaf mutism, spasticity, and ataxia but features of hypothyroidism are hard to find. In myxoedematous cretinism there are dwarfism, signs of myxoedema, and no goitre. The nervous type predominates in Papua New Guinea and parts of the Andes, while the myxoedematous type is seen in Zaire.

Endemic goitre has by now almost disappeared from the low iodine regions of developed, industrial areas like Derbyshire, the North American middle west, Switzerland, New Zealand, and Tasmania because much or all the salt that people eat is iodised; foods come in to the area that were grown or reared on soils with normal iodine; iodophors used as disinfectants in dairies get into the milk; dairy cows' winter rations have added iodine; and iodate may be used as a bread additive.

But in many remote, inaccessible parts of developing countries, endemic goitre and cretinism persist. Iodine status can be surveyed in such places by collecting single urine samples. Where goitre is common iodine excretions are all low and average less than 25 μg/1 g creatinine; the whole community is deficient. Endemic goitre was thought to be unaccompanied by functional effects (except for occasional local retrosternal pressure).

But in the 1980s it was recognised that "normal" people in goitrous districts (not diagnosed as cretins) have among them higher prevalences of deafness, slower reflexes, features of hypothyroidism, poorer learning ability, more stillbirths and malformed babies, and subnormal plasma thyroxines compared with control communities.[19] Any cretinism is thus the tip of the iceberg and the whole community on very low iodine intakes has a burden of miscellaneous impairments, iodine deficiency disorders (IDD), which reduce its capacity for productive work and development.

The collaboration of WHO, UNICEF, and ICCIDD (the International Council for Control of IDD) has achieved remarkable progress in reducing IDD during the 1990s. The major preventative measure is for governments of countries at risk to make iodisation of salt mandatory—and most of these countries have now done this, with particular success in South America. Where communities are isolated, away from the market economy the first line of prevention is to give all women of childbearing age 1 ml of iodised oil (rapeseed or poppy seed). In original trials this was injected but an oral capsule is nearly as effective and more convenient. UNICEF estimates that the number of children born with cretinism has been halved from the 1990 estimates of 120 000 worldwide. Doctors interested in activities towards IDD elimination can keep informed with the IDD Newsletter (obtainable free from the Editor, Dr J T Dunn, Box 800746, University of Virginia Medical Center, Charlottesville VA 22908, USA, email: jtd@virginia.edu).

Iodine deficiency disorders (IDDs)

Endemic goitre—just visible goitre in at least 5% of adolescents
When nearly all mothers have endemic goitre 1%-5% of babies are born with one of two types of **cretinism:**

• *Nervous cretinism*	mental deficiency deaf mutism spasticity ataxia, squint (features of hypothyroidism hard to find)
• *Myxoedematous cretinism*	dwarfism signs of myxoedema mental deficiency (no goitre)

"Normal" people (not cretins) in goitrous districts, when compared with control communities have:

higher incidence of deafness
slower reflexes
more pronounced features of hypothyroidism
poorer learning ability

Boys at an oasis in Egypt. Most have goitres[18]

Other types of malnutrition

Nutritional anaemia

The other WHO priority is nutritional anaemia. The commonest cause is iron deficiency, with folate deficiency second but well behind. Iron deficiency is probably the commonest of all nutritional deficiencies. WHO estimates more than 2 billion people—principally women and children—are iron deficient. It occurs in developed as well as developing countries and will be considered in the next chapter.

Distribution of other types of malnutrition

- Anaemia—common world wide
- Pellagra—seasonal in some developing countries in maize eaters —florid form rare
- Beriberi—occasional in infants in parts of South-East Asia

Pellagra

This is still seen in parts of Africa where people subsist on maize, in black people in rural areas of southern Africa, and in Egypt. Most of the niacin in maize is bound and not bioavailable. It is also poorer than other cereals in tryptophan, which can be partly converted to niacin in the liver. Clinical pellagra is seasonal and the florid form is no longer common anywhere in the world. Pellagra is also reported from Hyderabad, India, in people whose staple diet is sorghum. Sorghum eaters elsewhere in the world do not seem to be vulnerable. In Central America the staple food is maize (American corn), but pellagra is rare. This is because treating maize meal with lime ($Ca(OH)_2$) water, a traditional preliminary step in making tortillas, makes the bound niacin in cereals bioavailable. In developed countries, maize meal is fortified with niacin (as in the United States) or maize has been largely replaced by wheat in the diet.

Beriberi

In adults beriberi has almost disappeared but infantile beriberi is still occasionally seen in some underdeveloped rural areas of South-East Asia.

References

1 UNICEF. *The State of the World's Children 2000*. Oxford: Oxford University Press, 2000.
2 World Declaration on Nutrition. *Plan of action for nutrition. Adopted by the International Conference on Nutrition, jointly sponsored by FAO and WHO on 11 December 1992*. Geneva: WHO Division of Food & Nutrition.
3 Waterlow JC. *Protein energy malnutrition*. London: Edward Arnold, 1992.
4 Torun B, Chew F. Protein-energy malnutrition. In: Shils ME, Olson JA, Shike M, Ross AC (eds). *Modern Nutrition in Health and Disease*, 9th edn. Baltimore: Williams & Wilkins, 1999.
5 Whitehead RG, Coward WA, Lunn PG *et al*. A comparison of the pathogenesis of protein energy malnutrition in Uganda and the Gambia. *Trans R Soc Trop Med Hyg* 1977; **71**: 189-95.
6 World Health Organization. *Management of severe malnutrition: a manual for physicians and other senior health workers*. Geneva: WHO, 1999.
7 Payne P. Not enough food: malnutrition and famine. In: Harris-White B, Hoffenberg R (eds). *Food, multidisciplinary perspectives*. Oxford: Blackwell, 1994.
8 Harrison GE (ed). *Famine. (Biosocial Society Series no 1.)* Oxford: Oxford University Press, 1988.
9 Nutrition in times of disaster. *United Nations SCN News* 1989 no 3: 11-13.
10 Toole MJ. Micronutrient deficiencies in refugees. *Lancet* 1992; **339**: 1214-16.
11 Bern C, Lathanail L. Is mid-upper-arm circumference a useful tool for screening in emergency settings? *Lancet* 1995; **345**:631-3.
12 Ferro Luzzi A, James WPT. Adult malnutrition: simple assessment techniques for use in emergencies. *Br J Nutr* 1996; **75**: 3-10.
13 Sommer A, Tarwotjo I, Djunaedi E, West KP, Loeden AA, Tilden R. Impact of vitamin A supplementation on childhood mortality. A randomised controlled community trial. *Lancet* 1986; **i**: 1169-73.
14 American Academy of Pediatrics: Committee on Infectious Diseases. Vitamin A treatment of measles. *Pediatrics* 1993; **91**: 1014-15.
15 McLaren DS, Frigg N. *Sight and Life Manual on Vitamin A Deficiency Disorders (VADD)*, 2nd edn. Basel: Sight and Life, 2001.
16 de Pee S, West CE, Permaesih D, *et al*. Orange fruit is more effective than are dark-green leafy vegetables in increasing serum concentrations of retinol and β-carotene in schoolchildren in Indonesia. *Am J Clin Nutr* 1998; **68**: 1058-7.
17 Hetzel BS, Dunn JT, Stanburg JB (eds). *The prevention and control of iodine deficiency disorders*. Amsterdam: Elsevier, 1987.
18 Davidson S, Passmore R, Brock JF, Truswell AS. *Human nutrition and dietetics*, 7th edn. Edinburgh: Churchill Livingstone, 1979.
19 Hetzel BS. Iodine deficiency disorders (IDD) and their eradication. *Lancet* 1983; **2**: 1126-9.

9 Other nutritional deficiencies in affluent communities

Protein-energy malnutrition in Britain and other Western industrial countries is almost always secondary to disease—for example, it may be due to diseases of the gastrointestinal tract (persistent vomiting, dysphagia, upper intestinal obstruction, malabsorption) or wasting diseases (some cancers, HIV infection, metabolic disorders) or to radiotherapy. It is briefly described in chapter 17 (enteral and parenteral nutrition management to prevent or treat malnutrition in hospital). How to assess the degree of malnutrition is dealt with in chapter 12 (Measuring nutrition). For a background, famine and childhood protein malnutrition are described in chapter 8 (Malnutrition in developing countries).

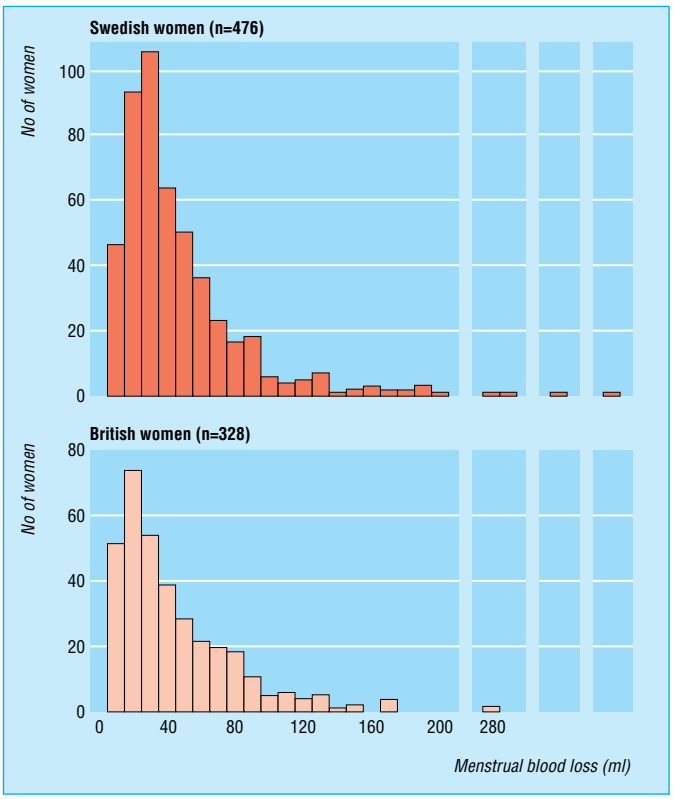
In Britain vitamin D deficiency is still a public health issue, requiring preventive measures, though rickets is not common now. Its pathogenesis is discussed in chapter 10 (Vitamins). Among hospital patients in developed countries, however, the most common vitamin deficiency is probably folate deficiency, of which there is an account in chapter 10. The most serious nutritional deficiency in alcoholics is the complex of Wernicke's encephalopathy and Korsakoff's psychosis, which is summarised under thiamin in chapter 10. The following nutritional problems have not been discussed anywhere else in this book.

Iron deficiency

Iron deficiency is probably the commonest micronutrient deficiency in the world. Over 400 million women are estimated to have subnormal haemoglobins and over one billion people to have iron deficiency.[1] People in industrialised countries are affected as well as those in developing countries, especially women aged 15 to 50. In a survey in Vanuatu, however, most people with hypochromic anaemia were found to have normal plasma ferritin but, by DNA analysis, a unsuspectedly high prevalence of α-thalassaemias.[2] Hypochromic anaemias in other tropical countries cannot all be assumed to be due to iron deficiency unless biochemical studies of iron status are done. In Britain 33% of women had serum ferritin under 25 μg/l, indicating low stores and 14% had very low ferritins (below 13 μg/l). In the same survey only 4% of women had haemoglobins below 11 g/dl.[3] In British children serum ferritins are more likely to be low in youngest boys (13% were <20 μg/l) and in adolescent girls (27% were 15 μg/l).[4]

Iron is the second most abundant metal in the earth's crust but it is virtually insoluble in the complexes of ferric iron usual in foods at neutral pH, and absorption is difficult. There is no physiological mechanism for secretion of iron so maintenance of iron homeostasis depends on its absorption. Normally in men, children, and postmenopausal women iron is lost only in desquamated surface cells from gut and skin at an estimated rate of 1 mg/day or less.

Blood is by far the richest tissue in iron; 1 ml contains 0.5 mg so that a regular loss of only 2 ml/day—for example, from epistaxis or haemorrhoids—doubles the iron requirements. Women of reproductive age lose an average of 30 ml blood per period, corresponding to 0.5 mg iron per day over the month, so they need more iron than men. An important minority of women lose considerably more.



Menstrual blood loss per period in 476 Swedish and 328 British women[5]

Iron content of some foods[6] (mg/100 g) (in descending order of total iron per usual serving)

Food	Iron content
Cockles (boiled)	26.0*
Black pudding (blood sausage)	20.0*
Liver, cooked: ox – pig	7.8-17.0*
Beef, rump steak, grilled, lean	3.5*
Lamb, leg, roast, lean	2.7*
Oatmeal	3.8
Legumes, cooked: baked beans – tofu	1.4-3.5
Green leafy vegetables: lettuce – watercress	0.7-2.2
Wines, white or red	0.6-1.2*
Egg, boiled	1.9
Dried fruit: dates – figs	1.1-4.2
Fish: cod-anchovies	0.4-4.1*
Chicken, roast, meat	0.8*
Nuts: chestnuts – cashew	0.9-6.2
Chocolate, plain, dark	2.4
Potato, baked	0.7
Bread: white-wholemeal	1.6-2.7
Fresh fruit: apples – passion fruit	0.1-1.3
Milk, cows', whole	0.05

* Higher availability—haem and muscle iron or alcohol
Other foods with unusually high Fe content: curry powder (58 mg/100 g), All Bran (20 mg/100 g), wheat bran (13 mg/100 g), some other fortified breakfast cereals (13 mg/100 g), venison (8 mg/100 g), pigeon (15 mg/100 g), hare (11 mg/100 g), grouse (8 mg/100 g), hearts (5-8 mg/100 g), kidneys (5-8 mg/100 g)

These women, pregnant women, children growing fast, and anyone with chronic bleeding all need to absorb extra iron or they will use up what tissue stores they have.

Absorption of iron is inefficient. It averages roughly 10% from a mixed diet but is much less from many plant foods and from eggs and dairy foods. Haem iron is better absorbed than non-haem iron. Absorption of the latter is enhanced by animal flesh and by ascorbic and other organic acids (for example citric and lactic) and reduced by phytates and polyphenols (as in tea). Iron absorption is being studied by extrinsic labelling of foods with isotopes of iron (radio-iron was used earlier, but stable isotopes are now available and preferred). The different factors in foods combine algebraically to produce a high, medium, or low absorption from the meal. For example, tea reduces iron absorption from a meal but orange juice enhances it.

Iron is essential for haemoglobin formation. It is also part of myoglobin, of some enzymes required for neurotransmitter synthesis, and of an important enzyme in DNA synthesis. In deficiency, sometimes even before there is anaemia, adults have decreased capacity for heavy work, pregnant women have an increased risk of low birth weight or premature babies, and children do not concentrate or learn as well as they can after iron treatment.[7]

In people at risk of iron deficiency the haemoglobin (at least) should be checked, in:

- infants at the age of 1 year
- children and adolescents during phases of rapid growth
- through and after pregnancy
- after gastric surgery at least once a year
- women with heavy periods (direct questions may need to be asked)
- patients presenting with gastrointestinal symptoms or disease
- anyone with a history of recurrent bleeding, with a positive occult blood test, or a woman who is a frequent blood donor
- people (for example, with arthritis) taking aspirin or NSAIDs regularly.

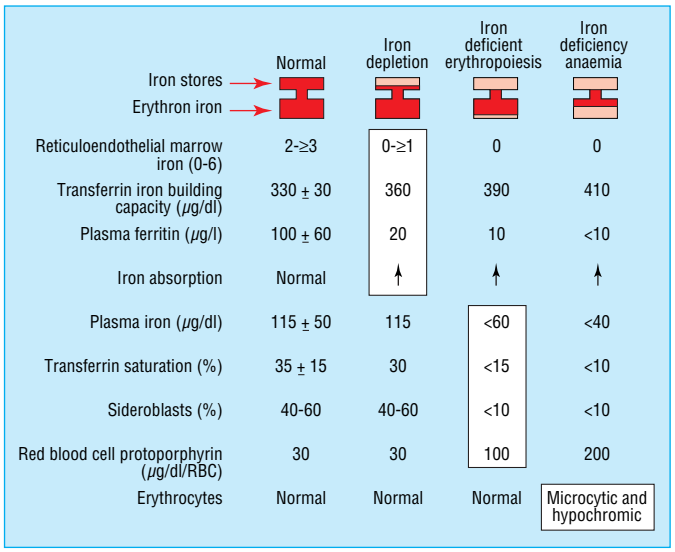
Ferrous sulphate is standard **therapy** for iron deficiency. The traditional dose for anaemia has been three tablets per day (3×60 mg Fe) but gastrointestinal side effects are common, related to the dose of iron and can lead to poor compliance. Absorption of iron is enhanced when there is deficiency and it may be practical to prescribe a more modest dose of 60 mg iron (one tablet) per day (given between meals). Iron is as well absorbed from ferrous fumarate (60 mg/tablet) or gluconate, but the latter only contains 35 mg per tablet. There is no advantage to more complex and expensive tablets containing iron. In developing countries trials are underway in which iron tablets are given only weekly for mild to moderate chronic iron deficiency.

Prevention

A common cause of iron deficiency is that women, who require twice as much as men, consume only half as much. They often eat less liver, meat, and fish, the best sources of available iron. People who take no exercise or are on a weight reducing diet have such a low calorie intake that it is difficult for them to eat enough iron.[8] Iron absorption is increased in deficient individuals but this may not be enough to compensate for low intake or increased losses. Staple foods are fortified with iron in some countries (for example, many breakfast cereals in Britain) and old fashioned iron cooking pots add some iron to the food.

Too much iron?

- There is no homeostatic mechanism for disposing of excess iron; if too much is absorbed it accumulates in liver, heart and other organs. Chronic iron overload occurs in three situations. In **hereditary haemochromatosis** absorption is enhanced on an ordinary diet. One in 300 men of North West European descent are affected and fewer women, who have a mutation of the HFE gene, discovered in 1996. Patients with intractable anaemias, (for example, thalassaemia) **and repeated blood transfusions** receive excess iron parenterally, and people who regularly imbibe **alcohol beverages with high iron content** receive excess absorbable iron by mouth (for example, Bantu siderosis).
- Free iron can damage tissues by catalysing the conversion of hydrogen peroxide to the destructive hydroxyl free radical ($OH^{\bullet}$). This is why in the body iron (and copper) are almost always bound to carrier proteins or locked away in storage proteins.



Sequential changes in the development of iron deficiency[5]

Calcium and the bones

Calcium is the most obvious and persistent of the micronutrients, the fifth most abundant element (and the most

abundant cation) in the body, yet it is more difficult to measure adequacy of intake for calcium than for other nutrients. There are two major questions about calcium.

(1) Will a generous intake during childhood and adolescence contribute to taller adult height or heavier bones, or both? If so how much is needed?
(2) Will a generous intake from about 45 years onwards delay the onset of osteoporosis, especially in women, who are more likely to be affected? If so how much is best and in what form?

Over 99% of body calcium is in the skeleton. Here it not only provides structural support but is a large reservoir for maintaining the plasma calcium concentration at very stable concentrations. Any reduction of absorbed calcium does not show in the plasma concentration, which is immediately reset by an increased parathyroid hormone concentration and the formation of active 1,25 dihydroxyvitamin D (in the kidney from 25 hydroxyvitamin D). Ionised calcium **in the plasma** has many vital functions—muscle contractility, neuromuscular irritability, blood coagulation, etc—which would be disturbed if its concentration fell. By different modulators the ionised calcium **inside cells** is also tightly controlled. An inadequate dietary calcium will therefore not normally be allowed to lower plasma or intracellular calcium concentrations and disturb their numerous important soft tissue functions; instead a little less will go into the bones in children or a little will be removed from the bones in adults.

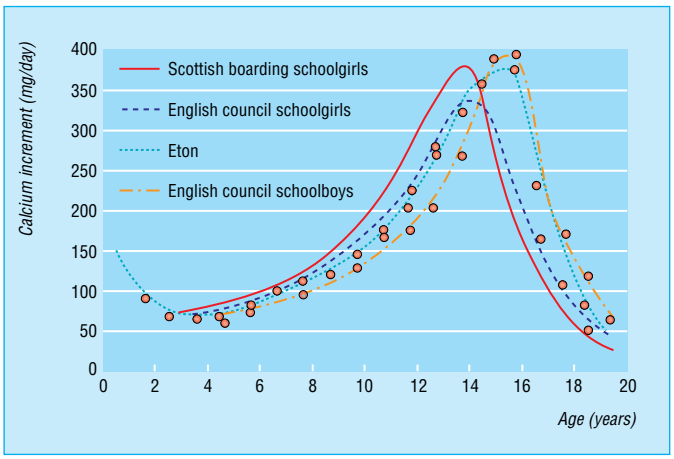
Calcium in the bones amounts to 25 g at birth and builds up to about 1200 g in an adult. From the indices of growth and skeletal composition the amount of calcium which is being added to the bones each day in growing children can be calculated. It averages about 180 mg calcium per day but reaches 400 mg per day at the peak of adolescent linear growth. This amount is the required positive balance. Calcium absorption is inefficient—faecal calcium is about 70% of intake in adult calcium balances—and there is an obligatory loss of calcium in urine as well. The diet must therefore supply substantially more than the daily skeletal increment, and the recommended daily amount in Britain is 1000 mg in boys and 800 mg in girls aged 11-18 years. There is evidence that in some conditions a supplement of milk (the best source of calcium) has improved the growth rate of children. The great increase in the height of young Japanese adults from 1950 to 1970 coincided with a tripling of the national calcium intake from about the lowest in the world to 600 mg/day.

Yet it is hard to understand how poor children in developing countries on diets of cereals and vegetables and no milk obtain enough calcium to grow. Their final adult height is usually lower than in industrial countries, but osteoporosis in older people seems to be less of a problem. There are several possibilities to explain the more economical handling of calcium in people in countries where there is little or no milk.

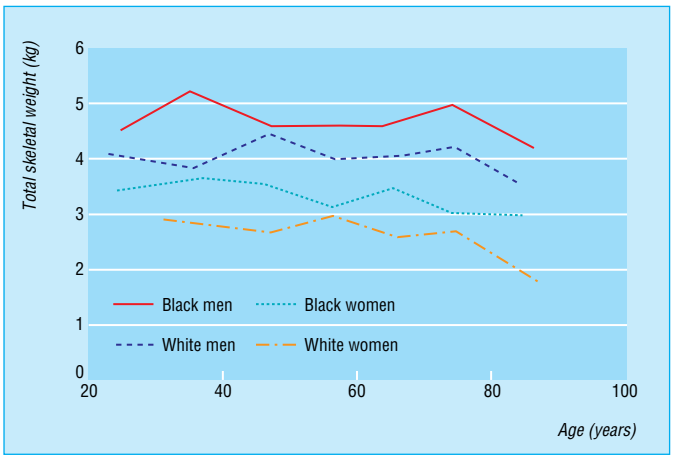
- More skin synthesis of vitamin D from sunshine in the tropics.
- Regular weight-bearing exercise which promotes bone formation and muscle strength.
- Lower intake of animal protein (which is known to increase the obligatory urinary calcium).
- Lower intake of salt (which also increases urinary calcium loss).
- Genetic selection.

Osteoporosis

Osteoporosis is porosis of the bones. There are more holes in trabecular (inside) bone and thinning of cortical bone. The ratio of mineral (mostly $CaPO_4$) to organic matter (collagen



Estimated daily increment in skeletal calcium of children at various ages[9]



Age related changes in total skeletal weight of black and white people in the United States of America[10]

In 1927 a series of trials was carried out in Scotland in which about 1500 children in the ordinary elementary schools in the seven largest towns were given additional milk at school for a period of seven months. Periodic measurements of the children showed that the rate of growth in those getting the additional milk was about 20% greater than in those not getting additional milk. The increased rate of growth was accompanied by a noticeable improvement in health and vigour. This experiment was twice repeated by different observers who obtained substantially the same results on numbers up to 20 000 children.[11]

and cells) is normal but the amount of total bone is reduced. In other words osteoporosis is atrophy of bones. It causes no symptoms unless/until there is the only complication—a fracture associated with inadequate trauma.

The commonest sites of osteoporotic fracture are:

- neck of femur (hip) fracture, the most serious
- vertebral collapse fracture(s)
- arm fracture—Colles' wrist fracture.

Osteoporosis cannot be recognised by clinical examination. It can be diagnosed with DEXA (dual *x*-ray absorptiometry) of a femoral neck and/or lumbar spine and/or a distal forearm. WHO defines osteoporosis as a bone mineral density less than 2.5 standard deviations below the mean for young adults (separate standards for women and men; for Caucasians and African Americans).

Osteoporosis is thus analogous to atherosclerosis. They are both important pathological states because they are precursors of common and serious clinical diseases, but they can only be recognised by special techniques (special imaging for osteoporosis, invasive for atherosclerosis). There are two main types of **primary** osteoporosis: postmenopausal and senile. **Secondary** osteoporosis is due to conditions such as long-term corticosteroid treatment. Bone density is also reduced by vitamin D deficiency (disproportionate loss of mineral) or infiltration by different types of cancer.

Dieting can be dangerous

Malnutrition can result from dietary regimens which happen to be very unbalanced nutritionally. Some of these were introduced by medical graduates, others are unscientific. The following are only a sample.

"Liquid protein" combined with fasting

In the United States of America in the late 1970s this diet led to at least 60 deaths from cardiac arrhythmias in people with no history of heart disease. The product "Prolinn", an extract from beef hides, lacked several essential amino acids. It was withdrawn, but prolonged fasting with or without protein supplements (even those of good biological value) carries the risk of sudden fatal arrhythmias and has been criticised authoritatively.

Zen macrobiotic diets

These diets consist of 10 levels. The highest level is 100% cereals and prescribes a very low fluid intake. These diets have led to scurvy and/or impaired renal function, anaemia, hypocalcaemia, and emaciation. In some cases these have been fatal. These diets have been condemned by the American Medical Association.

Dr Atkins's diet revolution

This weight reducing diet in a popular paperback written in 1972 prescribed a minimal carbohydrate intake. Ketosis is inevitable; and the diet raises plasma lipid concentrations. It was condemned by the American Medical Association but the book can still be found on bookstalls, having sold millions of copies. It has been revived recently. This time it should be thoroughly tested for efficacy and safety.

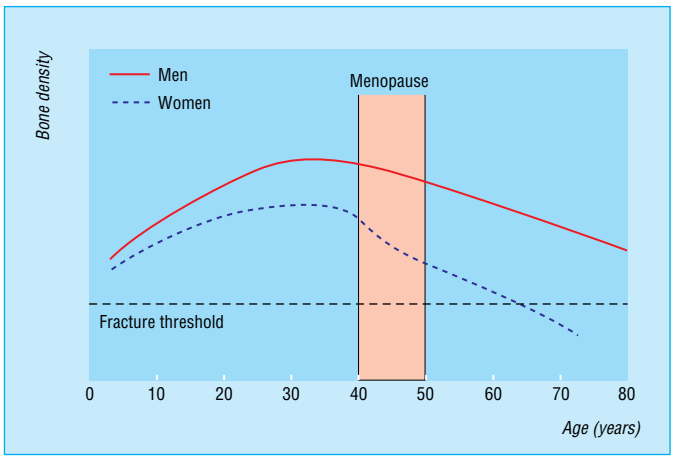
Strict vegan diets for infants

Plant foods contain no vitamins B-12 or D. The latter can be synthesised in the skin if a child is exposed to sunlight, but the most serious nutritional complication of strict vegetarian diets is vitamin B-12 deficiency in infants. The milk of vegan mothers

Calcium and osteoporosis

- There have been several controlled trials of the effect in adolescent girls of extra milk or calcium tablets on bone density. Most reported small extra gains in the supplement group[12] but the trials were only of 1 or $1\frac{1}{2}$ year's duration and it is not known whether the effects persist.
- In postmenopausal women 20 trials have been reported with calcium tablets or extra milk (usually 1000 mg Ca for two years) added to usual diet. In most of the trials bone density increased 1-3% (not always significantly) in the treated group.[13] Presumably extra calcium has most benefit where the usual dietary calcium was low. The US Committee on Dietary Reference Intakes has set the dietary reference intake for people over 70 years at 1200 mg (30 mmol) Ca/day[14] but in the UK a COMA subcommittee's re-examination did not consider there was evidence for increase of the dietary reference value of 700 mg/day.[15]
- A good calcium intake is part of preventing (delaying) osteoporosis, along with regular weight bearing exercise and vitamin D from sun or diet. For treating osteoporosis calcium is less effective than hormone replacement but should be prescribed alongside the latter.

Osteoporosis is only important because of these fractures. The most serious is hip fracture. Its pathogenesis is multifactorial. As well as low bone mineral density of the proximal femur, instability, muscle weakness and lack of adipose "padding" all contribute.[16]



Average bone density with age and fracture threshold

Articles which show the dangers of some diets

- *Liquid protein and fasting*—Isner JM *et al. Circulation* 1979; **60:** 1401-12.
- *Zen macrobiotic diets*—AMA Council on Foods and Nutrition. *JAMA* 1971; **218**: 397-8.
- *Dr Atkins's diet revolution*—AMA Council on Foods and Nutrition. *JAMA* 1973; **224**: 1415-19.
- Vitamin B-12 deficiency in vegan infants—Wighton MC *et al. Med J Aust* 1979; **ii**: 1-3 (see next box).
- *Beverly Hills diet*—Mirkin GB, Shore RN. *JAMA* 1981; **246**: 2235-7.

contains insufficient vitamin B-12 unless she takes a supplement. This vitamin is required for normal myelin formation, and infants' nervous systems are specially susceptible to deficiency. They can show impaired mental development, involuntary movements, and even coma responsive to vitamin B-12, as well as megaloblastic anaemia.

The Beverley Hills diet

This weight reducing diet requires consumption of nothing but fruit (all in a certain order and only the designated fruits) for the first 10 days. Some bread, salad, and meat are added later. The theory behind this diet is unscientific, and it has been criticised in detail in the *Journal of the American Medical Association.*

An exclusively breast fed infant of a vegan mother

"Neurological deterioration commenced between 3 and 6 months of age and progressed to a comatose premoribund state by the age of 9 months. Investigations revealed a mild nutritional vitamin B-12 deficiency in the mother and a very severe nutritional B-12 deficiency in the infant with severe megaloblastic anaemia. Treatment of the infant with vitamin B-12 resulted in a rapid clinical and haematological improvement but neurological recovery was incomplete …"

By 17 months of age his general level of motor, social, and intellectual development was that of a child of 11 months.

Wighton *et al.,* 1979
(reference in previous box)

Eating disorders[17]

Anorexia nervosa is an illness of our time, although it was first described and named in 1874. Many teenage girls and young women say they are dieting to stay or become slim but most are not very successful. They do not stick to their diets. The young woman with anorexia nervosa is unusual: without talking about dieting she succeeds in losing a lot of the weight that the others say they would like to. But then she cannot stop. By rigid control of her eating she avoids foods that she understands to be fattening (appetite is usually there, but suppressed). She has a phobia of being fat and a distorted body image, seeing herself fatter than she really is. Amenorrhoea is characteristic.

Up to one in 100 middle class women aged 15 to 25 may be affected. Before this loss of weight the young woman with anorexia nervosa was often a model of good behaviour, conformism, and achievement though this probably concealed a sense of ineffectiveness and self doubt.

Some women not only abstain: they have learnt to induce vomiting or purging and may have eating binges between. When habitual this behaviour is **bulimia nervosa**.

The physical effects of a young woman starving herself to 45 kg and below are similar to those described for famine in chapter 8. But there are differences. The anorectic patient usually eats adequate protein and micronutrients and is restless and overactive. There is skinniness or emaciation, cold extremities, lanugo hair, bradycardia, low blood pressure, and normal pubic and axillary hair. Plasma potassium concentrations may be low if there has been vomiting or purging, and plasma cholesterol or carotene values are sometimes raised. The amenorrhoea is similar to that which occurs in starvation.

Early diagnosis is important because a long illness, and severe weight loss are all bad prognostic features. The young woman denies she is too thin or that she is dieting. Amenorrhoea may appear to have preceded the weight loss. Other organic diseases that can lead to emaciation and amenorrhoea have to be excluded—for example, thyrotoxicosis, malabsorption, and hypopituitarism. But the bigger challenge for the family doctor is to detect the characteristic features of anorexia nervosa and to convince the young woman that she needs treatment.

Treatment is usually best managed by a specialised team, most often a psychiatrist with special experience in anorexia nervosa working with a dietitian. Moderate cases are treated as out patients. Admission to a special anorexia unit is indicated if weight loss is severe (BMI 16.5) or there is some medical or psychological complication. The two lines of treatment are to persuade her to increase food intake and to get her weight up to a target figure (usually a compromise) and at the same time

Criteria for diagnosis

Anorexia nervosa

1 Loss of weight to body mass index* under 17.5.
2 Disturbance of body image.
3 Refusal to maintain normal weight.
4 Intense fear of becoming fat.
5 Amenorrhoea.
6 No known medical illness leading to weight loss.

Bulimia

1 Recurrent episodes of binge eating, often secretive with rapid consumption of high calorie foods.
2 Binge followed by self-induced vomiting or purging.
3 Strict dieting between binges.
4 Body weight may be within the normal range.
5 Awareness of abnormal eating patterns and fear of not being able to stop voluntarily.
6 Not due to any physical disorder.

In practice intermediate and atypical forms are common
*Weight (kg)/height $(m)^2$

Treatment in a special anorexia unit

to provide emotional support. Eating is tactfully supervised by nurses, and behavioural principles can be effective, more privileges for each step of weight gain, previously agreed by the patient. As she is refed and becomes physically stronger it becomes easier for the patient to tackle the changes needed so that she can adapt to society and develop a healthier body image.

In a 5-year follow up study Ben Tovin *et al.* found that 3/95 anorexia patients died and 20 had persistent anorexia nervosa. Prognosis was better for bulimia: none died and only 8% still had bulimia. Many patients made a good recovery without accessing specialised treatments. Those with lengthy admissions did not necessarily have better long term outcome.[18]

Interactions of food, nutrition, and drugs

Particular drugs can affect the nutritional state, changing the results of biochemical tests or even leading on occasions to clinical undernutrition, overnutrition, or malnutrition.

Appetite may be decreased by anorectic drugs, bulking agents, dexamphetamine, metformin, cardiac glycosides, glucagon, morphine, phenylbutazone, indometacin, cyclophosphamide, fluorouracil, methylphenidate, salbutamol, levodopa, etc, and by drugs that alter taste (griseofulvin, penicillamine, and lincomycin).

Appetite may be increased by sulphonylureas, oral contraceptives, cyproheptadine, chlorpromazine, androgens, anabolic steroids, corticosteroids, insulin, lithium, amitriptyline, pizotifen, clomipramine, benzodiazepines, and metoclopramide.

Malabsorption for more than one nutrient may be induced by neomycin, kanamycin, paromomycin, colchicine, phenindione, chlortetracycline, cholestyramine, colestipol, cyclophosphamide, indometacin, liquid paraffin (fat soluble vitamins), methotrextate, and methyldopa.

Energy metabolism may be stimulated by caffeine, smoking, and some sympathomimetic drugs.

Carbohydrates—Increased blood glucose concentrations may be produced by corticosteroids, thiazide diuretics, diazoxide, oral contraceptives, and phenytoin. Hypoglycaemia may be produced by propranolol and by alcohol (as well as by sulphonylureas, metformin and insulin).

Lipids—Plasma total cholesterol may be raised by thiazides (for example, chlorthalidone, hydrochlorothiazide), by phenobarbital, chlorpromazine, some oral contraceptives, and large intakes of boiled coffee. As well as specific cholesterol lowering drugs, aspirin, colchicine, prazosin, clonidine, neomycin, phenformin, and sulphinpyrazone may lower total cholesterol. **Plasma high density lipoprotein cholesterol** may be raised by phenytoin, ethanol, cimetidine, valproate, carbamazine, terbutaline, and prazosin. It may be lowered by danazol, propranolol, and oxprenolol.

Plasma triglycerides may be raised by propranolol, ethanol, and (oestrogenic) oral contraceptives. They may be lowered by norethisterone.

Protein—Nitrogen balance may be made negative by corticosteroids, vaccines, and tetracyclines. It may be made positive by insulin or anabolic steroids. **Plasma amino acids** may be increased by tranylcypromine and lowered by oral contraceptives. Plasma phenylalanine may be raised by trimethoprim and methotrexate.

Thiamin absorption can be reduced by ethanol.

Riboflavin status may be lowered by oral contraceptives and by chlorpromazine.

Niacin may be antagonised by isoniazid.

Vitamin B-6 may be antagonised by isoniazid, hydralazine, cycloserine, penicillamine, oral contraceptives, oestrogens, hydrocortisone, imipramine, levodopa, piperazine, and pyrazinamide.

Nutrients, foods, and drugs can interact in several ways

- **Foods can affect drugs**, for example, by affecting absorption, an acute effect of single meals. Grapefruit juice inhibits one of the cytochrome P450s that metabolises drugs such as calcium-channel blockers, statins, carbamazepine, and terfenadine.[19]
- **Nutrition can affect drugs**. The nutritional state can affect drug metabolism and hence dosage and toxicity, for example, in Kwashiorkor.
- **Particular drugs can affect the nutritional state**. Appetite, absorption, metabolism, and concentration of nutrients can be affected, positively or negatively, by different drugs (see Left).
- **Drugs can cause unpleasant reactions to minor components in some foods** whose metabolism we normally take for granted—for example, hypertension from tyramine in cheese in patients taking monoamine oxidase inhibitors.
- **A few drugs are used as drinks**, as part of the usual diet: alcoholic drinks, coffee, tea, and carbonated cola beverages.
- **Some nutrients are used as drugs**. The nutrients are all obtainable in pure form. They may, in doses above the nutrient requirement, sometimes have a useful pharmacological action—for example, nicotinic acid for hyperlipidaemia.

Should I take the medicine before or after meals, doctor?

- Most drugs are best taken with or just after meals, because this is the easiest way to remember to take any drug and some are gastric irritants. Absorption of several drugs is a little delayed but this is unimportant and a few are better absorbed when taken with meals—for example, griseofulvin, metoprolol, and labetalol.
- Plenty of water should be taken with uricosurics (to prevent renal precipitation) and with cholestyramine and bulk formers like methyl cellulose.
- A few drugs should be taken **half an hour before meals**: antibiotics which are labile in acid—ampicillin, benzylpenicillin, cloxacillin, erythromycin, lincomycin, tetracycline, rifampicin, and isoniazid. So should one of the antidiabetic agents—glipizide—and, of course, appetite suppressant drugs.

Folate may be antagonised by ethanol, phenytoin, oral contraceptives (uncommonly), cycloserine, triamterine, and cholestyramine. In addition several drugs owe their antibacterial action to antagonism of folate metabolism—more in microbial than mammalian cells—pyrimethamine, trimethoprin, and pentamidine. Methotrexate is a potent folate antagonist which has more effect on rapidly dividing cells—for example, cancer cells.

Vitamin B-12 absorption may be impaired by slow K, cimetidine, ranitidine, metformin, colchicine, trifluroperazine, and by high doses of vitamin C, cholestyramine, and methotrexate. Prolonged nitrous oxide anaesthesia oxidises vitamin B-12 *in vivo*. Smoking and oral contraceptives reduce the plasma concentration.

Vitamin C—Plasma concentrations are lowered by oral contraceptives, smoking, aspirin, and tetracycline. Ascorbate excretion is increased by corticosteroids, phenylbutazone and sulfinpyrazone.

Vitamin A plasma concentration is increased by oral contraceptives. Absorption may be reduced by liquid paraffin and cholestyramine.

Vitamin D status is lowered by anticonvulsants—for example, phenytoin, phenobarbitone, and when these are taken in high dose for long periods rickets can occur.

Vitamin E is antagonised by iron in premature newborns. Fish oils (refined) increase requirements.

Vitamin K—Coumarin drugs—for example warfarin—are antimetabolites. Purgatives and intestinal antibiotics, such as neomycin, tetracyclines, and sulphonamides, may remove the contribution from colonic bacteria. Salicylates and cholestyramine may reduce absorption, and some cephalosporin antibiotics antagonise the vitamin K-epoxide cycle.

Potassium—Drugs are important causes of potassium depletion: purgatives and laxatives increase faecal loss; thiazide diuretics and frusemide and ethacrynic acid increase renal loss. Other drugs that may increase urinary potassium are penicillin, glucocorticoids, liquorice, outdated tetracycline, gentamicin, and alcohol. Insulin can lower serum potassium. Drugs that raise serum potassium include ACE inhibitors, spironolactone, succinylcholine, triamterene, and potassium compounds.

Calcium—Absorption may be increased by aluminium hydroxide or by cholestyramine and decreased by phosphates and corticosteroids. Thiazide diuretics decrease urinary calcium excretion. Gentamicin, dactinomycin, frusemide, and ethacrynic acid increase it.

Iron—Gastrointestinal bleeding from aspirin and NSAIDS depletes the body's iron. Allopurinol, fructose, and ascorbic acid increase absorption. Antacids, phosphates, and tetracycline decrease it. Oral contraceptives tend to increase serum iron.

Iodine—Sulphonylureas, phenylbutazone, amiodorone, and lithium can cause goitre; they interfere with iodine uptake in the gland. Serum protein bound iodine is increased by oral contraceptives, *x*-ray contrast media, and potassium iodide, and decreased by phenytoin.

Phosphate absorption is decreased by aluminium or calcium compounds.

Zinc depletion from increased urinary excretion may be produced by thiazide diuretics and frusemide, by cisplatin, penicillamine, and alcohol.

Magnesium depletion from increased urinary loss may be produced by thiazides and frusemide, cisplatin, alcohol, aminoglycosides, amphotericin, ciclosporin, and gentamicin.

References

1 United National Subcommittee on Nutrition, Focus on micronutrients. *SCN News* no 9: mid 1993; 1-34.
2 Bowden DK, Hill AVS, Higgs DR, Weatherall DJ, Clegg JB. Relative roles of genetic factors, dietary deficiency, and infection in anaemia in Vanuatu, south-west Pacific. *Lancet* 1985; **ii**: 1025-8.
3 Gregory J, Foster K, Tyler H, Wiseman M. *The dietary and nutritional survey of British adults*. London: HMSO, 1990.
4 Gregory J, Lowe S, Bates CT *et al*. National Diet and Nutrition Survey: young people aged 4 to 18 years. London: Stationery Office, 2000.
5 Bothwell TH, Charlton RW, Cook JD, Finch CA. *Iron metabolism in man*. Oxford: Blackwell, 1979, p 251.
6 Holland B, Welch AA, Unwin ID, Buss DA, Paul AA, Southgate DAT. *McCance & Widdowson's the composition of foods*, 5th edn. Cambridge: Royal Society of Chemistry, 1991.
7 Brunner AB, Joffe A, Duggan AK, Casella JF, Brandt J. Randomised study of cognitive effects of iron supplementation in iron-anaemic iron-deficient adolescent girls. *Lancet* 1996; **348**: 929-96.
8 Barber SA, Bull NL, Buss DH. Low iron intakes among young women in Britain. *BMJ* 1985; **290**: 743-4.
9 Leitch I, Aitken FC. The estimation of calcium requirements a re-examination. *Nutr Rev* 1959; **29**: 393-407.
10 Merz AL, Trotter M, Peterson RR. Estimation of skeletal weight in the living. *Am J Phys Anthropol* 1956; **14**: 589-610.
11 Orr JB. *Food, health and income. Report of a survey of adequacy of diet in relation to income*. London: Macmillan, 1936.
12 Cadogan J, Eastell R, Jones N, Barker ME. Milk intake and bone mineral acquisition in adolescent girls: randomised, controlled intervention trial. *BMJ* 1997; **317**: 1255-60.
13 Nordin BEC. Calcium and osteoporosis. *Nutrition (Elsevier)* 1997; **13**: 664-86.
14 Standing Committee on the Scientific Evaluation of Dietary Reference Intakes, Food and Nutrition Board, Institute of Medicine. *Dietary reference intakes for calcium, phosphorus, magnesium, vitamin D and fluoride*. Washington, DC: National Academy Press, 1997.
15 Report of the Subgroup on Bone Health, Working Group on the Nutritional Status of the Population of the Committee on Medical Aspects of Food and Nutrition Policy. *Nutrition and Bone Health: with particular reference to calcium and vitamin D*. London: Stationery Office, 1998.
16 Cummings SR, Melton LJ. Osteoporosis 1. Epidemiology and outcomes of osteoporotic fractures. *Lancet* 2002; **359**: 1761-7.
17 Fairburn CG, Harrison PJ. Eating disorders. *Lancet* 2003; **361**: 407–16.
18 Ben Tovin DI, Walker K, Gilchrist P, Freemen R, Kalucy R, Esterman A. Outcome in patients with eating disorders: a 5-year study. *Lancet* 2001; **357**: 1254-7.
19 Spence JD. Drug interactions with grapefruit: whose responsibility is it to warn the public? *Clin Pharm Therap* 1997; **61**: 395-400.

10 Vitamins and some minerals

> No animal can live on a mixture of pure protein, fat and carbohydrate, and even when the necessary inorganic material is carefully supplied the animal still cannot flourish. The animal body is adjusted to live upon plant tissues or the tissues of other animals and these contain countless substances other than the proteins, carbohydrates and fats.
>
> Sir Frederick Gowland Hopkins (1906)

Deficiencies of vitamins still occur in affluent countries: folate, thiamin, and vitamins D and C. Some of these deficiencies are induced by diseases or drugs. In developing countries deficiency diseases are more prevalent. Vitamin A deficiency (xerophthalmia), for example, is a major cause of blindness. Some vitamins may have useful actions above the dose that prevents classic deficiency disease—for example, vitamins A, C, and B-6; nicotinic acid has been used to treat hyperlipidaemia.

Vitamins have caught the popular imagination, and they are also big business. Many people take over the counter vitamins without medical advice and a few unorthodox practitioners prescribe "megavitamin therapy". Doctors therefore need to know the symptoms of overdosage.

Daily requirements for healthy adults*

Vitamin A	1 mg
Thiamin	1 mg
Riboflavin	1.5 mg
Niacin	15-20 mg**
Vitamin B-6	1.5 mg
Pantothenic acid	5 mg
Biotin	30 μg
Folate	200 μg†
Vitamin B-12	1.5 μg
Vitamin C	40-60 mg
Vitamin D	5 μg‡
Vitamin E	10 mg
Vitamin K	70 μg

*Based on DHSS 1991[1] and rounded
**Part replaceable by tryptophan in proteins
†Double this in pregnancy
‡More for growth: no dietary requirement if adequate exposure to sunlight

Definition

Vitamins are:

(*a*) Organic substances or groups of related substances
(*b*) found in some foods
(*c*) substances with specific biochemical functions in the human body
(*d*) not made in the body (or not in sufficient quantity)
(*e*) required in **very small** amounts.

Many people seem to have lost sight of point (*e*), but it appears in all dictionary definitions and can be seen in the table of requirements. The daily requirement of most vitamins is around 1 mg, the weight of one grain of raw sugar. There are no exceptions to points (*a*), (*b*), and (*e*). On point (*c*), the biochemical action of most vitamins can now be visualised, but those of vitamins A and C are not yet explained fully, and the active metabolite of vitamin D acts as a hormone. Exceptions to point (*d*) are that certain carotenoids can replace vitamin A; proteins (through the amino acid, tryptophan) can replace niacin; and exposure to sunlight can replace vitamin D.

Recommended names for vitamins

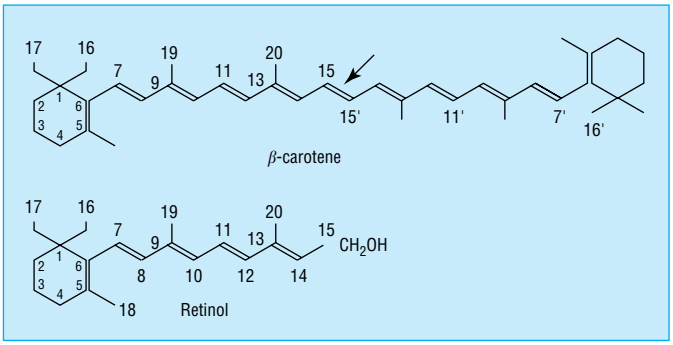
Recommended name*	Alternative name	Usual pharmaceutical preparation
Vitamin A	Retinol	Retinol palmitate
Thiamin†	Vitamin B-1	Thiamine hydrochloride
Riboflavin	Vitamin B-2	Riboflavin
Niacin	Nicotinic acid and nicotinamide	Nicotinamide
Vitamin B-6	Pyridoxine	Pyridoxine hydrochloride
Pantothenic acid		Calcium pantothenate
Biotin		Biotin
Folate	Folacin	Folic acid
Vitamin B-12	Cobalamin	Hydroxocobalamin or Cyanocobalamin
Vitamin C	Ascorbic acid	Ascorbic acid
Vitamin D	Vitamins D_2 and D_3	(Ergo) calciferol
Vitamin E		α-Tocopheryl acetate
Vitamin K		Vitamin K_1

* International Union of Nutritional Sciences
† Spelt 'thiamine' in the pharmaceutical literature

Vitamin A

Best understood of the actions of vitamin A is its role in night vision; 11-*cis* retinaldehyde is combined with a specific protein in the light-sensitive pigment, rhodopsin, in the rods of the retina. Night blindness occurs in children deficient in vitamin A in some developing countries, and in affluent countries it is seen rarely in patients with chronic biliary obstruction or malabsorption.

More recently discovered functions of vitamin A affect many different cell types. Retinol, carried in plasma on retinol binding protein, is taken up in cells by cellular retinol binding protein, oxidised to retinoic acid or 9-*cis* retinoic acid. These are transported to the nucleus where they are bound to specific receptors and initiate genetic transcription. In vitamin A deficiency there is metaplasia of conjunctival epithelium and



Formation of retinol from β-carotene

loss of mucus production, leading to xerophthalmia (chapter 8). Other epithelia, for example, respiratory, are similarly affected and their resistance to infection is lowered.

Preformed vitamin A (retinol) is found in animal foods: liver is the richest source, but about a quarter of vitamin A intake in Britain comes from carotenes, yellow and orange pigments in the leaves of vegetables and in some fruits, chiefly β-carotene. One molecule of β-carotene can be cleaved by a specific intestinal enzyme into two molecules of vitamin A. But this conversion is not very efficient, 6 μg β-carotene is assumed to be equivalent to 1 μg retinol. Vitamin A is stored in the liver; stores are enough for one to two years in most British adults (see chapter 8). Retinol is transported from the liver to the rest of the body on retinol binding protein, part of the pre-albumin complex. Its concentration is normally held constant and does not reflect vitamin A intake except when this is very low or high.

[International units of vitamin A can be confusing if used for β-carotene and are best avoided, but some pharmaceuticals continue to use them. For retinol 1 IU = 0.3 μg retinol, so the UK reference nutrient intake for men is either 700 μg or 2333 IU.]

In supranutritional amounts vitamin A reduces both keratinisation of skin and sebum production. 13-*cis* retinoic acid (tretinoin) and its isomer isotretinoin, are used either topically in creams or orally in capsules, but oral retinoids are teratogenic and must not be prescribed for women in whom there is any possibility of pregnancy.

Because vitamin A is teratogenic there is no role for megadosage of this vitamin. Regular intakes should not exceed 3.3 mg in early pregnancy (4.5 times the UK reference nutrient intake of 0.7 mg). Liver contains 13-40 mg vitamin A per 100 g (depending on species), so women who are or might become pregnant are advised not to eat liver or products made from it. High doses of vitamin A are toxic to non-pregnant people and fatalities have occurred. Acute hypervitaminosis A causes raised intracranial pressure and skin desequamation. Chronic overdosage is more common and can occur after long term intakes of 10 times the nutritional requirement or more. Symptoms include headache, alopecia, dry itchy skin, hepatomegaly, bone and joint pains. A high plasma vitamin A confirms the diagnosis. High intakes of β-carotene, on the other hand, colour the plasma and skin (hypercarotenaemia) but are not dangerous.

Food sources of vitamin A

Preformed vitamin A (retinol)
Liver, fish, liver oils (very rich sources)
kidney, dairy produce
eggs, fortified margarine

β-carotene
Carrots, red palm oil
apricots, melon, pumpkin
dark green leafy vegetables
(spinach, broccoli, sprouts, etc)

In Britain the main sources in the diet are liver, margarine, butter and dairy products.

Carotenoids

600 colours of flowers, autumn leaves and birds, yellow to red, have chemically similar structure to β-carotene, with variations usually at the end rings. Of the three carotenoids in higher concentration in human plasma (reflecting dietary intake) only β-carotene is pro-vitamin A. The other two are lycopene, the red colour of tomatoes, which has antioxidant properties, and lutein/zeaxanthin (stereoisomers) which give the yellow colour to the macula lutea in the retina.

Thiamin (vitamin B-1)

Thiamin plays a part in the metabolism of carbohydrates, alcohol, and branched chain amino acids. The body contains only 30 mg—30 times the daily nutrient requirement—and deficiency starts after about a month, on a thiamin free diet sooner than for any other vitamin. The requirements are proportional to the non-fat energy intake. The two principal deficiency diseases are beriberi and Wernicke–Korsakoff syndrome.

Beriberi is now rare in the countries where it was originally described—Japan, Indonesia, and Malaysia. In Western countries occasional cases are seen in alcoholics: clinical features are a high output cardiac failure with few electrocardiograph changes and a prompt response to thiamin treatment alone.

Wernicke–Korsakoff syndrome is usually seen in alcoholics: it can also occur in people who fast (such as hunger strikers) or who have persistent vomiting (as in hyperemesis gravidarum). Early recognition is important. The ophthalmoplegia and lowered consciousness respond to thiamin (50 mg intramuscularly) in two days, but if treatment is delayed the

Food sources of thiamin

- whole wheat and wholemeal breads
- wheat germ (richest source) bran
- yeast, mycoprotein, nuts
- pork, bacon, Marmite
- fortified breakfast cereals
- oatmeal, potatoes, and peas

In Britain the main sources in the diet are bread and cereal products, potatoes, and meat.

memory may never recover. Red cell transketolase and the effect on it of thiamin pyrophosphate (TPP) *in vitro* are used to confirm thiamin deficiency, but fresh whole blood is needed and must, if it is to be meaningful, be taken before thiamin treatment is started. If thiamin deficiency is suspected treatment should be started without waiting for the laboratory result. Two days later there will either have been a clinical response and a positive laboratory report of high TPP effect (indicates deficiency) or the provisional diagnosis will not have been confirmed.

Patients on regular haemodialysis should routinely be given small supplements of thiamin and other water-soluble vitamins. Thiamin should also be given prophylactically to people with persistent vomiting or prolonged gastric aspiration and those who go on long fasts, as well as alcoholics. The toxicity of thiamin is very low, though occasional cases of anaphylaxis have been reported after intravenous injection.

Wernicke–Korsakoff syndrome

In 1880 Wernicke first described an encephalopathy. Characteristic features are:

- stupor or apathy
- ophthalmoplegia (lateral or vertical)
- nystagmus
- ataxia.

With treatment most patients pass through a phase in which they show the memory disorder, first described by Korsakoff (1887), which consists of inability to retain new memories and confabulation.

The pathological findings in Wernicke's encephalopathy and Korsakoff's psychosis are similar: capillary haemorrhages in the mamillary bodies and round the aqueduct in the mid brain.

Wernicke's encephalopathy responds rapidly to thiamin but Korsakoff's psychosis responds slowly or not at all.

Riboflavin (vitamin B-2)

Riboflavin, a yellow substance with green fluorescence in the coenzymes flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD), has vital roles in cellular oxidation. Its biochemical functions do not easily explain the clinical manifestations that have been recorded in volunteers on a riboflavin deficient diet: angular stomatitis, cheilosis, atrophic papillae on the tongue, nasolabial dyssebacea, and anaemia. There are no real body stores of riboflavin, but the liver contains enough (in coenzyme form) to withstand depletion for about three months.

Most of the features of riboflavin deficiency have more than one cause. Angular stomatitis, for example, may occur with deficiencies of niacin, pyridoxine, or iron; after herpes febrilis; or with ill fitting dentures. Clinical riboflavin deficiency is very uncommon in milk drinking countries like Britain. Pregnant women, people with thyrotoxicosis, and those taking chlorpromazine, imipramine, and amitriptyline have increased requirements. Riboflavin has to be included in total diets: infant formulas, fluids for total parenteral nutrition, and supplements for patients on dialysis.

Food sources of riboflavin

- liver, kidney (richest sources)
- milk, yoghurt
- cheese, Marmite
- fortified cereals
- eggs, beef
- wheat bran
- mushrooms, wheat germ

In Britain the main sources in the diet are milk, meat, fortified breakfast cereals, and vegetables.

Niacin

Niacin (nicotinamide and nicotinic acid) is the part of the coenzymes, nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinulceotide phosphate (NADP), that has to be supplied by the diet. In addition the amino acid, tryptophan has a minor metabolic pathway via kynurenine to nicotinamide; about $\frac{1}{60}$ of ingested tryptophan goes this way. Tryptophan makes up about 1% of dietary proteins, so 70 g protein a day provides about 12 mg niacin equivalents towards the total niacin requirement of 15-18 mg a day (for adults).

Pellagra, caused by niacin deficiency, is now rare except in areas, such as parts of Africa, where people subsist on maize and little else. In maize the niacin is in a bound form, biologically unavailable (except when cooked after pre-treatment with calcium hydroxide water, the traditional Central American way), and tryptophan is its limiting amino acid (unlike other cereals). Secondary pellagra may occur in patients with chronic renal failure on low protein diets or dialysis, if niacin is not included in the regimen. Another rare cause is Hartnup disease, a recessive inborn error of tryptophan absorption.

Above the nutrient dose nicotinic acid (not the amide) produces cutaneous flushing from histamine release at doses of 100 mg/day or more; it has been used for chilblains.

Food sources of niacin

- liver, kidneys (richest source)
- meat, poultry
- fish
- brewer's yeast, Marmite
- peanuts
- bran, pulses
- wholemeal wheat
- coffee

The above are foods with useful amounts of niacin. Protein-rich foods also provide niacin equivalents via their tryptophan content.

At doses of 3 g/day or more it inhibits lipolysis in adipose tissue and lowers plasma cholesterol and triglyceride concentrations. It has been one of the treatments for combined hyperlipidaemia—hypercholesterolaemia plus hypertriglyceridaemia. Patients often develop tolerance to the flushing. Its other side effects (in high dosage) include gastric irritation, hyperuricaemia, impaired glucose tolerance and liver function tests, and occasionally cholestatic jaundice.

Vitamin B-6

The term vitamin B-6 includes five closely related substances that all occur in foods and in the body: pyridoxal and pyridoxamine, their 5′ phosphates, and pyridoxine, best known to doctors as the pharmaceutical form. Pyridoxal 5′ phosphate is coenzyme for over 100 reactions in the body involving amino acids. Many foods contain moderate amounts. Primary dietary deficiency is rare. An outbreak of convulsions in infants in 1954 was traced to insufficient vitamin B-6 in milk formula because of a manufacturing error. Several drugs interfere with vitamin B-6: hydralazine, penicillamine, and possibly oestrogens. Peripheral neuropathy from high dose isoniazid is prevented with pyridoxine. There are several conditions for which pharmacological doses of 50 to 100 mg pyridoxine are probably beneficial. These include homocystinuria, hyperoxaluria, gyrate atrophy of the choroid, hypochromic sideroblastic anaemia, and radiation sickness. Some biochemical indices of vitamin B-6 state may be abnormal in women taking some oral contraceptives, but these are indirect indices. The more specific plasma pyridoxal phosphate concentration is usually normal. Premenstrual tension is a very variable condition: prescribed or self medication with pyridoxine has no physiological basis and has never been subjected to a convincing double-blind trial.

Above 100 mg/day pyridoxine in repeated dosage may cause severe sensory neuropathy.[2] All seven patients in the first report of this side effect were taking pyridoxine for an inadequate indication—mostly for premenstrual oedema—and most had increased the dosage on their own. Pyridoxine should not be available over the counter at tablet size above 50 mg (which is already 33 times the nutrient requirement).

Food sources of vitamin B-6

- wheat germ and bran
- potatoes
- nuts and seeds, peanut butter
- meat, fatty fish, and offal
- other fish
- fortified breakfast cereals
- banana, avocado, dried fruits
- vegetables (especially raw), baked beans
- milk

Vitamin B-12

The red vitamin was the last to be isolated (1948). Humans eat it preformed in animal foods including fish and milk. It is synthesised by some micro-organisms—for example, in the rumen of cows and sheep (which require traces of cobalt in the pasture). No vegetable food has been shown to contain vitamin B-12 consistently unless it is contaminated—for example, by manure. Humans excrete in the faeces vitamin B-12 that has been synthesised by the colonic bacteria. Vitamin B-12 is the largest of the nutrients, with a molecular weight of about 1350. The physiological mechanism for its absorption requires intrinsic factor from the stomach, and the complex is absorbed only at a special site, in the terminal ileum. Deficiency occurs in several gastric, intestinal, and ileal diseases, including pernicious anaemia (gastric atrophy; no intrinsic factor), and in vegans (pure vegetarians). Adult body stores of vitamin B-12 in the liver last longer than those for any other vitamin, but deficiency occurs more quickly in infants. Vitamin B-12 cooperates with folate in DNA synthesis, so deficiency of either leads to megaloblastosis (anaemia and infertility). Vitamin B-12 has a separate biochemical role, unrelated to folate, in synthesising fatty acids in myelin,

Food sources of vitamin B-12

- liver (richest source)
- kidney
- sardines, oysters
- heart, rabbit
- other meats, fish
- eggs
- cheese
- milk
- some fortified breakfast cereals

so deficiency can present with neurological symptoms. Deficiency is diagnosed by measuring the serum vitamin B-12 and/or the concentration of methylmalonate which requires vitamin B-12 for its metabolism.

Supplementation with hydroxocobalamin is desirable for adult vegans and essential for their young children. Several drugs, such as colchicine and metformin, and prolonged anaesthesia with nitrous oxide, can interfere with absorption of vitamin B-12. Hydroxocobalamin can improve some cases of optic neuritis, possibly by detoxifying accumulated cyanide. Apart from rare hypersensitivity reactions there are no known toxic effects from vitamin B-12. It thus makes an ideal placebo, which may still be the commonest reason for its prescription!

Folate

Folic acid (pteroylglutamic acid) is the primary vitamin from the chemical point of view, and this is the pharmaceutical form because of its stability. But it is rare in foods or in the body. Most folates are in the reduced form (tetrahydrofolate); they have one-carbon components attached to the pteroyl ring, and up to seven (instead of one) glutamic acid residues in a row at one end. Folate is the group name for all these compounds with vitamin activity. These folates have many essential roles in one-carbon transfers in the body, including one of the steps in DNA synthesis.

In folate deficiency there is first a reduction of serum folate below 3 ng/ml (7 nmol/l) and later megaloblastosis of blood cells and other cells with a rapid turnover because cells cannot double their DNA to enable nuclear division. As well as anaemia, diarrhoea is common when the deficiency results from antagonism (due to drugs) rather than dietary lack.

Folate deficiency may occur simply from a poor diet, but it is usually seen when there is malabsorption or increased requirements because of pregnancy (chapter 4) or increased cell proliferation (haemopoiesis, lymphoproliferative disorders) or antagonism from a number of drugs. Methotrexate, pyrimethamine, and co-trimoxazole act preferentially in cancer cells or micro-organisms by inhibiting the complete reduction of folate to the active form, tetrahydrofolate, preferentially in cancer cells or micro-organisms. Alcohol is the commonest antagonist.

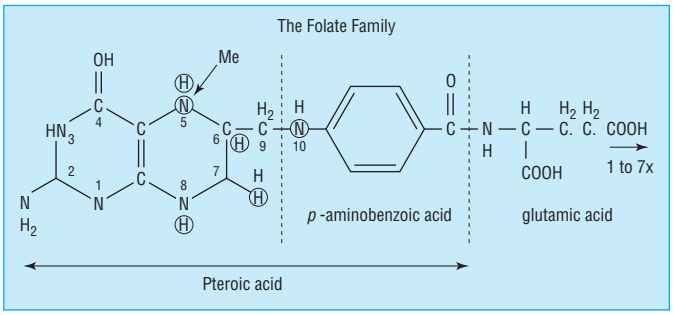
Body stores of folate are not large and deficiency can develop quickly in patients on intensive therapy. In some cases this can be ascribed to intravenous alcohol or particular parenteral amino acid mixtures. Trauma, infection, uraemia, increased haemopoiesis, dialysis, vomiting, or diarrhoea may also be partly responsible. Folate deficiency appears to be the most common vitamin deficiency among adult hospital patients in countries such as Britain, so supplements should be prescribed whenever patients are fed intravenously for more than a few days.

The name comes from the Latin *folia* (= leaf), but liver, legumes, nuts, and even wholemeal bread are as good dietary sources as leafy vegetables. Prolonged boiling destroys much of the vitamin in hospital cabbage. No toxic effects are known from moderate doses up to 1000 μg/day; 200 μg folate/day is more than enough to prevent folate deficiency megaloblastic anaemia and had been accepted worldwide as the reference nutrient intake (or recommended dietary allowance). Two research developments in the late 1990s changed this concept. Firstly, maternal folate intakes at the start of pregnancy above this level have been shown to greatly reduce the risk of neural tube effects in the fetus (chapter 4). Secondly, extra folate can reduce raised plasma homocysteine levels which have been

Food sources of folate

- liver (especially chicken)
- fortified breakfast cereals
- wheat germ, bran, soya flour
- blackeye beans (boiled)
- Brussels sprouts, peanuts
- kidney, other nuts and seeds
- broccoli, lettuce, peas, etc
- wholemeal bread, eggs
- citrus fruits, blackberries, potatoes
- cheese
- beef

In Britain the foods that provide most folate are potatoes, fortified breakfast cereals, bread, and fresh vegetables and some from beer



This is folic acid with extra hydrogens at 5, 6, 7 and 8. Tetrahydrofolate (pteroyl glutamic acid) monoglutamate.

Folic acid (pteroyl glutamic acid) is the primary vitamin from the chemical point of view, and it is the pharmaceutical form because of its stability. But it is rare in foods and in the body. Most folates are in the reduced form, tetrahydrofolate (THF); they also have 1-carbon components (methyl or formyl) attached to nitrogen atom 5 or 10, or bridging between them (5,10-methylene-tetrahydrofolate). In addition, they have up to 7 glutamic acid residues in a row (at the right in the figure)

shown to be an independent risk factor for cardiovascular disease (chapter 1).

In the USA fortification of cereal foods with folic acid was made mandatory in 1998 and the recommended intake has been increased to 400 μg per day for men and women (600 μg in pregnancy).

Vitamin C

Ascorbic acid is the major dietary antioxidant in the aqueous phase of the body. The best established biochemical consequence of its deficiency is impaired reduction of the amino acid, proline to hydroxyproline. Hydroxyproline is an uncommon amino acid, except in collagen, of which it makes up an indispensable 12%. Impaired collagen formation is the biochemical basis of scurvy.

Small doses of vitamin C will cure scurvy. Lind achieved this with two oranges and a lemon in the first controlled trial on HMS *Salisbury* in 1747; 30 mg of vitamin C is more than enough to prevent scurvy. Desirable intakes of vitamin C can be thought of at three levels.

(1) The official reference nutrient intake for adults—40 mg/day in Britain and 75 to 90 mg/day in the United States—is for healthy people. This is more than enough ascorbate to prevent scurvy.
(2) In hospital patients this is not enough. Absorption of the vitamin may be reduced or its catabolism increased by disease. Trauma and surgery increase the need for vitamin C for collagen synthesis. Several drugs antagonise vitamin C: adrenal corticosteroids, aspirin, indometacin, phenylbutazone, and tetracycline, together with smoking. Hence it is advisable to give a supplement of up to 250 mg ascorbic acid a day to cover major surgery.
(3) The third level is the great vitamin C debate: megadoses (up to 10 g/day) proposed by the late Linus Pauling for superhealth—or not? The best known claim for large intakes of vitamin C is that they prevent common colds. At least 31 controlled trials have been reported and in 23 of them (including the largest and best designed ones) there was no significant preventive effect. The eight supportive trials all had qualifications—for example, they were not double blind, had tiny groups, or showed an effect only in a subgroup.[4]

The present evolution of Pauling's hypothesis is that high intakes of ascorbate increase antioxidant capacity in the body and may help to prevent (or delay) degenerative diseases, such as cataract.[5] High intakes of fruits and vegetables appear epidemiologically to reduce the risk of stomach cancer. The active protective factor might be vitamin C.

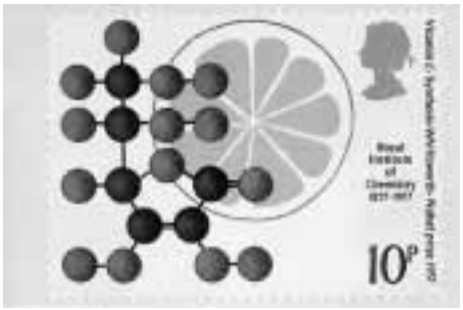
At megadosage the law of diminishing returns comes in. The small intestine has limited capacity to absorb ascorbate (so it should be eaten, or taken as tablets, tds). When blood levels increase, the vitamin is excreted in the urine. Careful pharmacokinetic experiments show that little further increase of plasma ascorbate can be achieved above intakes of 250 mg/ day.[6] At high intakes urinary excretion of oxalate (which is a metabolite of ascorbate) increases somewhat. Although oxalate is a common component of urinary tract stones, large epidemiological studies have not found more cases of stones in people who take vitamin C supplements.[7] People with a history of nephrolithiasis should nevertheless avoid these supplements, so should patients with chronic renal failure whose plasma ascorbate can rise to levels above the normal range.

Vitamin C enhances the absorption of non-haem iron taken at the same time. This works with mixtures of foods or juices and is a benefit from generous intakes for 299/300 people.

Serum folate results 1994–1998 in a large clinical laboratory in California, USA

Year	Number of tests	Rate of serum folate testing*	Test results < 2.7 μg/l	Test results ≥ 20 μg/l	Median serum folate value μg/l
1994	14 493	7.3	183 (1.3)	3709 (25.6)	12.6
1995	14 750	6.7	186 (1.3)	3652 (24.8)	12.7
1996	17 642	7.5	223 (1.3)	4130 (23.4)	11.7
1997	22 805	8.9	134 (0.6)	7185 (34.3)	14.9
1998	26 662	10.2	89 (0.3)	12 990 (45.3)	18.7
Total	98 351	8.3	815 (0.8)	31 666 (32.9)	14.7

*Per 1000 members
Since mandatory fortification with folate in the USA (Jan 1998) and Canada (Nov 1998) serum folates have increased, homocysteines have declined and the incidence of neural tube defects is lower[3]



Food sources of vitamin C

- blackcurrants, guavas
- rosehip syrup, green peppers
- oranges, other citrus fruit, strawberries
- cauliflower, broccoli
- sprouts, cabbage, watercress
- potatoes
- (liver and milk)

In Britain the main sources of vitamin C are fruit juices, potatoes, and other vegetables.

But for the 1/300 people with homozygous haemochromatosis genes vitamin C supplements are contraindicated.

Vitamin C is easily destroyed by cooking (aggravated by alkaline conditions, for example, sodium bicarbonate), so fresh fruit and juices and salads should be encouraged and vegetables cooked lightly and quickly.

Vitamin D

Cholecalciferol is hydroxylated in the liver to 25-OHD_3, the plasma concentration of which is a good index of vitamin D status. In the kidney 25-OHD_3 is further hydroxylated either to 1,25$(OH_2)D_3$ (calcitriol) or to an inactive metabolite. Calcitriol functions as a hormone whose best known action is to stimulate the synthesis of a calcium transport protein in the epithelium of the small intestines.

The natural substance cholecalciferol was originally called vitamin D_3. Vitamin D_2 is the artificially produced ergocalciferol. The natural and usual source of cholecalciferol is by the action of short wavelength ultraviolet light from the sun on a companion of cholesterol in the skin, 7-dehydrocholesterol. Cholecalciferol also occurs in a small minority of our foods. When people live in high latitudes, wear clothes, and spend nearly all the time indoors and the sky is polluted with smoke they have insufficient exposure in the winter to ultraviolet light to make the required amount of this substance; under these conditions dietary intake becomes critical and cholecalciferol assumes the role of a vitamin.

In rickets and osteomalacia there is reduced calcification of growing and mature bones respectively. These diseases have been more prevalent in Britain than in other Western countries. They tend to affect adolescents and the elderly, especially Asians in northern cities. In Britons with normal levels plasma 25-OHD_3 concentrations show annual fluctuations, with their trough in late winter and their peak after the summer holidays. It is not clear whether the lower prevalence of rickets in Canada and Sweden is because milk is fortified with vitamin D or because people receive more ultraviolet radiation of their skin over the year in these other northern countries.

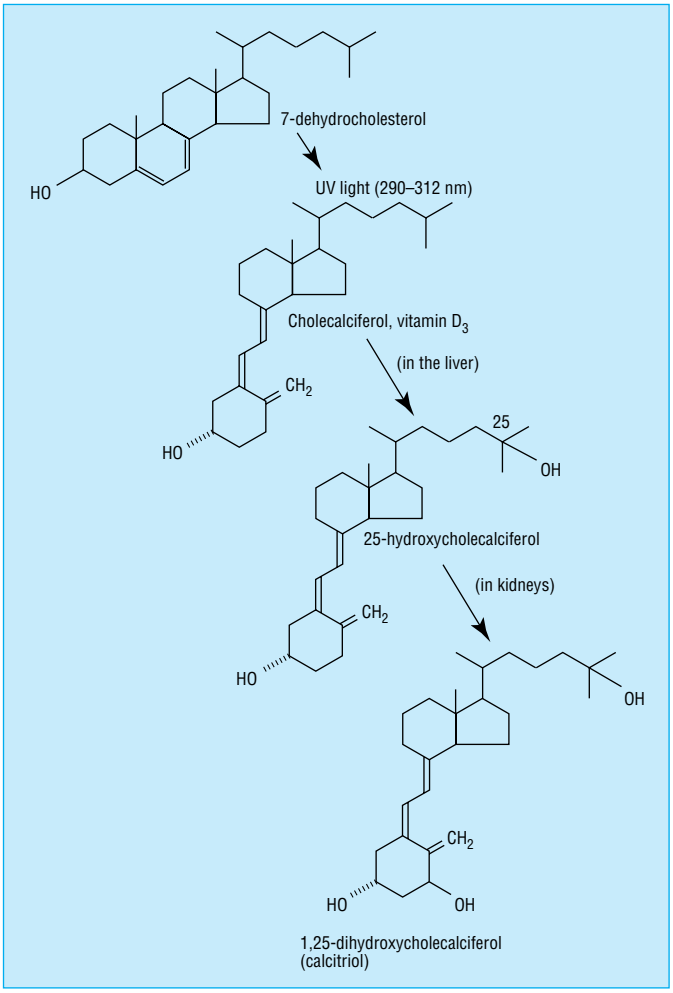
The small dietary contribution of vitamin D is lost in malabsorption and chronic biliary obstruction. Long term anticonvulsants, phenobarbitone, and phenytoin, increase metabolic losses. Vitamin D is indicated in these conditions. In chronic renal failure and hypoparathyroidism 1 α-hydroxylation to the active metabolite is impaired and renal bone disease responds only to 1,25$(OH)_2D_3$ (calcitriol) or 1α-OHD_3 (alfacalcidol), a synthetic derivative.

Irradiation of the skin may cause sunburn but does not lead to vitamin D toxicity. On the other hand, the margin of safety with oral vitamin D, between the nutrient requirements of up to 10 μg and toxic intakes, is narrow. Overdose with vitamin D causes hypercalcaemia, with thirst, anorexia, polyuria, and the risk of metastatic calcification. Some children have developed hypercalcaemia on vitamin D intakes only five times the recommended nutrient intake. More than this should not be taken except for rickets or osteomalacia. Here 25 to 100 μg vitamin D—for example, as ergocalciferol—is the usual therapeutic dose.

[One international unit (IU, obsolescent) of vitamin D = 0.025 μg of cholecalciferol or ergocalciferol—to convert IU to micrograms, divide by 40.]

Vitamin E

α-Tocopherol is the most active of eight very similar compounds with vitamin E activity. Being fat soluble, vitamin E



It may be difficult to remember which foods contain useful amounts of which vitamins but it's a good general rule that highly refined foods and drinks like fats, oils, sugar, cornflour, and alcoholic spirits contain little or no vitamins.

Food sources of vitamin D

- fish liver oils
- fatty fish (sardines, herring, mackerel, tuna, salmon, pilchards)
- margarine (fortified)
- infant milk formulas (fortified)
- eggs, liver

In Britain the main sources in the diet are margarines, fatty fish, dairy spreads, breakfast cereals, and eggs

is present in all cell membranes where, being an antioxidant, it is thought to reduce peroxidation of unsaturated fatty acids by free oxygen radicals.

The nutritional requirement for vitamin E is roughly proportional to the intake of polyunsaturated fat. Vitamin E is not easily transported across the placenta, and signs of deficiency, mild haemolytic anaemia, are sometimes found in premature infants.

The most severe cases of deficiency occur in patients with chronic fat malabsorption, especially fibrocystic disease of the pancreas and abetalipoproteinaemia. As well as mild anaemia, in these conditions ataxia, loss of tendon jerks, and pigmentary retinopathy have been reported, which respond to long term vitamin E treatment.

Many people take vitamin E supplements on their own initiative in large doses. Earlier it was rumoured to enhance virility (infertility had been the first reported effect of deficiency in rats) but double-blind trial did not confirm this. In the 1990s the focus is on whether vitamin E's antioxidant activity *in vitro* can reach sufficient concentrations inside the body to reduce atherogenesis. Two large cohort studies in the United States[8] suggest that vitamin E supplements may reduce the risk of coronary heart disease, but a large 5-year randomised trial in Finland[9] found no benefit from 50 mg α-tocopherol/day (about five times average dietary intake). Two other large preventive trials in Italy and another in Canada have likewise found no benefit with even larger doses. Although it is a fat-soluble vitamin, tocopherol has a low toxicity. Few adverse effects have been reported from doses up to 3200 mg/day and none were observed consistently.[1]

Food sources of vitamin E

- vegetable oils—wheat germ oil the richest
- margarines, mayonnaise
- nuts and seeds

Small amounts in wholegrain cereals, eggs, butter, some vegetables, and some fruits.

- Four members of the vitamin E family are α-, β-, γ- and δ-tocopherol. These differ chemically in the number and position of methyl groups on the double ring at one end of the molecule. Biologically α-tocopherol is the most potent and β, γ, and δ are each in turn less active. The tocopherols are more active than the four *tocotrienols*, which have double bonds in the side chain; α-tocotrienol is the most biopotent, next β-tocotrienol.
- The eight vitamin E compounds also show d- and l-stereoisomerism. Natural forms are d- (or RRR) and synthetic or dl- (or racemic). Both forms of α-tocopherol are available commercially. The most biologically active compound is the natural d- (or RRR) α-tocopherol, and vitamin E activity in foods or tissues is summed as d- (or RRR) α-tocopherol equivalents.

Vitamin K

The Koagulations vitamin (Dam, 1935) comes in two chemical forms. Vitamin K_1 (phytomenadione) is found mainly in vegetables. The K_2 vitamins (menaquinones) are a series produced by bacteria—for example, in the gut. Deficiency of vitamin K manifests itself as hypoprothrombinaemia and bleeding.

Cord blood levels of vitamin K are very low (evidently placental transfer is limited), and breast milk contains little of the vitamin unless the mother has been dosed with vitamin K. To prevent haemorrhagic disease of the newborn 1 mg of vitamin K_1 (by injection or by mouth) is given either to all infants or to those at increased risk (low birth weight or difficult delivery), depending on the hospital's policy. The single intramuscular injection of vitamin K_1 prevents both early and late vitamin K deficiency bleeding. In one British epidemiological study this injection at birth appeared to be associated with increased risk of childhood cancers. Subsequent studies in several countries have not confirmed this and in the United States there has been no increase of childhood cancer nationally since vitamin K injection at birth became common practice around 1961.[10] Oral vitamin K prevents early but not late haemorrhagic disease. Doubts still linger[11] and doctors should follow locally agreed policy.

In adults vitamin K deficiency is to be expected in obstructive jaundice and can occur in malabsorption syndromes. Vitamin K_1 must be given before surgery for these conditions. Anticoagulants of the warfarin group owe their therapeutic action to antagonism of vitamin K, and vitamin K_1 is the antidote for overdose.

Vitamin K promotes the synthesis of an unusual amino acid, γ-carboxyglutamic (gla) a component of coagulation proteins II, VII, IX, and X. Another protein that contains gla and

Food sources of vitamin K

- turnip greens
- broccoli
- cabbage, lettuce
- liver

These are all good sources, though there is no systematic list.

requires vitamin K for its synthesis is osteocalcin, involved in bone formation.

Vitamins that can usually be taken for granted (but are required in total parenteral nutrition)

Biotin is cofactor for several carboxylase enzymes concerned in fat synthesis and amino acid catabolism. It is widely distributed in foods, the requirement is small, and deficiency is rare. Deficiency has occurred in people who eat large amounts of raw eggs (which contain a protein that binds biotin and prevents its absorption) and in patients receiving total parenteral nutrition with biotin omitted. They suffer scaly dermatitis, loss of hair, hypercholesterolaemia, and a characteristic combination of organic acids in the urine. **Panthothenic acid** is a constituent of coenzyme A which has many functions and is widespread in the body and in foods. The name means "available everywhere". Spontaneous deficiency in man has never been proved. **Choline** is part of lecithin and of sphingomyelin, the two major phospholipids in the body, and it is also part of acetylcholine, the neurotransmitter. It is a dietary essential for the rat, but man seems to be able to synthesise it (partly from methionine) and does not have the active catabolising enzyme (choline oxidase) found in rat liver.

Multivitamins

The sensible purpose of multivitamin preparations is an insurance policy for people whose diet may be restricted or unbalanced but neither they nor their adviser is sure which vitamin may be lacking. There is a case for multivitamin supplements for people with low calorie intake because of poor appetite or a weight reducing diet or frailty, also for food faddists, the emotionally disturbed and socially disadvantaged people.[12] The dose of each vitamin should be near the nutritional requirement so a multivitamin cannot do harm, even if it does not do good. A doctor prescribing multivitamins or talking to patients who choose to take them should sometimes check the small print—as one should with an insurance policy. How many vitamins of the maximum 13, or the 11 described above are in the ingredient list? Do they contain folic acid? Multivitamin preparations have no lucrative patents, so drug companies are not very interested in keeping them up to date; reformulating is expensive. They are also rather neglected by medicine committees and by dietitians.

Some minerals

At least 13 inorganic elements *per se* are known to be essential for man (the same as the number of vitamins) while others are needed in compounds (P, S, Cl, Co). All must be provided for long term total parenteral nutrition and ensured in infant formulas. Of the nutritionally important inorganic elements, sodium and potassium are discussed in chapter 2, fluoride in chapter 3, iodine in chapter 8, calcium and iron in chapter 9. Zinc and selenium are sometimes taken as supplements and deserve mention here.

Zinc
This metal is cofactor for over 100 enzymes (including superoxide dismutase) and "zinc fingers" are part of a number

Not vitamins

The following compounds sold in "health food" shops and still included in some multivitamin pharmaceuticals are not vitamins. They are not required in infant formulas or in fluids for total parenteral nutrition:

- Bioflavonoids
- Inositol
- Orotic acid
- Aminobenzoic acid (PABA)
- Vitamin B-15 ("pangamic acid")
- Vitamin B-17 (laetrile)
- Vitamin P

of important transcription factors for DNA. Tissue concentrations of zinc are highest in the choroid of the eye, the prostate and in semen.

The first recognised human deficiency disease (1963) was adolescent growth retardation and hypogonadism in rural Iran; absorption of the small zinc intake in their mainly vegetable diet was hindered by phytate in unleavened bread. The most florid clinical features of zinc deficiency—moist facial eczema, depression, hair loss and diarrhoea have been seen with total parenteral nutrition that omitted zinc and in acrodermatitis enteropathica, a rare inborn error of zinc absorption. Zinc is predominantly intracellular and serum zinc is not a reliable indicator of deficiency. Randomised controlled trials have demonstrated benefit from zinc supplements in children in developing countries with acute diarrhoea,[13] pneumonia, and stunting.[14] Universal zinc supplements will not, however, improve the growth of stunted children except where zinc is the primary growth-limiting nutrient. Reports of benefit in cases of the common cold in industrial countries were not confirmed. The best dietary sources of zinc are meat, fish, cheese, and whole grain cereals (though absorption is reduced by phytate).

Selenium

Selenium compounds have long been known to be toxic. The content of this trace element in foods varies greatly, depending on the amount in the soil. Since the 1950s selenium deficiency in animals has been known to cause muscular dystrophy in sheep and liver necrosis in rats. Human deficiency has caused juvenile cardiomyopathy in Keshan, China (1979). In the body selenium replaces sulphur in selenomethionine and selenocysteine. It is part of glutathione peroxidase (antioxidant) and an enzyme that converts thyroxine to triiodothyronine. Blood selenium reflects intake and nutritional status and so does the Se content of toenails. Intakes have been moderately low in Finland and in South Island, New Zealand. Finland corrected this by fortifying fertilisers used for wheat fields with selenium from 1994. In New Zealand extensive research has been done at Otago University on selenium nutrition and metabolism. No clear human condition has been attributable to the local selenium level. Meanwhile intakes there are rising a little with more foods coming in from Australia. But in Britain selenium intakes have been declining[15] due to less wheat coming in from North America (high Se) and more from the continent of Europe (lower Se). Occasional cases of muscular disease have been reported in patients maintained long term on total parenteral nutrition if selenium was not included. A large randomised controlled trial with selenium in people with early skin cancer for five years, in low selenium areas in the south east United States of America showed an unexpected significantly reduced incidence of prostate cancer.[16]

References

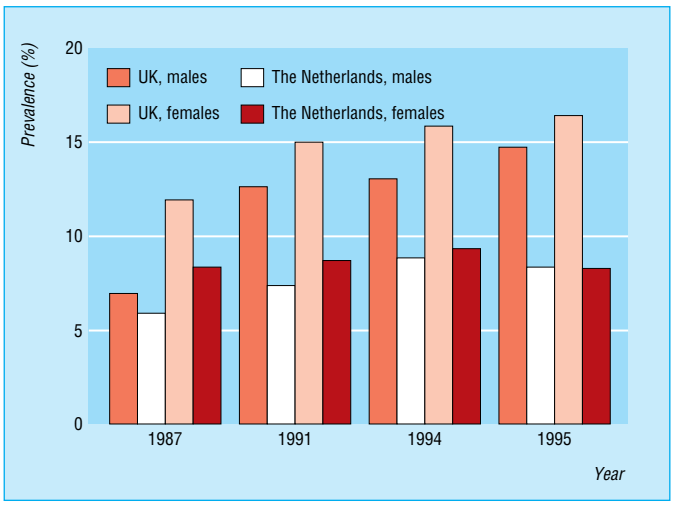
1 Department of Health. *Dietary reference values for food energy and nutrients for the United Kingdom.* Report on Health & Social Subjects no 41. London: HMSO, 1991.
2 Institute of Medicine. *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B-6, Folate, Vitamin B-12, Pantothenic Acid, Biotin and Choline.* Washington DC: National Academy Press, 1998.
3 Lawrence JM, Petitti DB, Watkins M, Umekubo MA. Trends in serum folate after food fortification. *Lancet* 1999; **354**: 915–6.
4 Truswell AS. Ascorbic acid and colds (letter). *N Engl J Med* 1986; **315**: 709.
5 Jacques PF, Taylor A, Hankinson SE *et al.* Long-term vitamin C supplement use and prevalence of early age-related lens opacities. *Am J Clin Nutr* 1997; **66**: 911-16.
6 Levine M, Corry-Cantilena C, Wang Y *et al.* Vitamin C pharmacokinetics in healthy volunteers: evidence for a recommended dietary allowance. *Proc Natl Acad Sci USA* 1996; **93**: 3704-9.
7 Gerster H. No contribution of ascorbic acid to renal calcium oxalate stones. *Ann Nutr Metab* 1997; **41**: 269-82.
8 Rimm EB, Stampfer MJ, Ascherio A, Giovannucci E, Colditz GA, Willett WC. Vitamin E consumption and the risk of coronary heart disease in men. *N Engl J Med* 1993; **328**: 1450-6.
9 Rapola JM, Virtamo J, Ripatti S *et al.* Randomised trial of α-tocopherol and β-carotene supplements on incidence of major coronary events in men with previous myocardial infarction. *Lancet* 1997; **349**: 1715-20.
10 American Academy of Pediatrics, Vitamin K ad hoc Task Force. Controversies concerning vitamin K and the newborn. *Pediatrics* 1993; **91**: 1001-3.
11 The 17 January 1998 issue of the BMJ contains an editorial, four research papers and a letter with a meta-analysis—a total of 24 pages—on neonatal vitamin K prophlaxis. *BMJ* 1998; **316**: 161, 173-93.
12 Truswell AS. Who should take vitamin supplements? *BMJ* 1990; **301**: 135-6.
13 Sazawal S, Black RB, Bhan MK, Bhandari N, Sinha A, Jalk S. Zinc supplementation in young children with acute diarrhoea in India. *N Engl J Med* 1995; **333**: 839-44.
14 Umeta M, West CE, Haidar J, Deurenberg P, Hautvast JGAJ. Zinc supplementation and stunted infants in Ethiopia: a randomised controlled trial. *Lancet* 2000; **355**: 2021-6.
15 Rayman MP. Dietary selenium: time to act. *BMJ* 1997; **314**: 387-8.
16 Clark LC, Combs GF, Turnbull BW *et al.* Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin. A randomised controlled trial. *JAMA* 1996; **276**: 1957-63.

11 Overweight and obesity

Obesity is an increasing problem

Obesity is rising to alarming levels around the world.[1]
In England during the 15 years 1980 to 1995 some important health statistics improved and life expectancy lengthened by three years. But the numbers of obese men (16-64 years) increased from 6% to 15% and obese women from 8% to 16.5%. Obesity is one of only four of the 25 targeted statistics in the Health of the Nation strategy that has moved in the wrong direction. Similar trends of increasing obesity have occurred in almost all industrialised countries and in the prosperous class in developing countries.

There have been recent advances in our understanding of genetics, endocrinology and metabolism related to adiposity. However treatment options have not kept pace; fenfluramine and dexfenfluramine, sadly, in 1997 joined the list of disappointing and/or potentially dangerous treatments for obesity.[1] Resources for the management of obesity—dietetic, pharmaceutical, and surgical—have been likened to treatment options for hypertension 40 years ago.[3]



Prevalence of obesity in the UK and the Netherlands 1987-1995[2]

Why is there an epidemic of increasing obesity?
Although the external influences and internal processes are both very complex the ultimate causes of obesity are under-exercising and/or over-eating: energy intake > energy expenditure. But the development of obesity, via overweight is usually so slow and insidious that people hardly notice it is happening. The body's homeostatic energy regulation is better able to defend against insufficient food than against a little more food, a little less energy expenditure (see box opposite).

Energy balance

1 kg body weight gained has energy of approx 7000 kcal
10 kg weight gain over 5 years = $\frac{70\,0000}{5\times365}$ = 38 kcal/day
This is a daily error of energy balance of +1.5%
OR 10 minutes' walk
OR one square (1/8) of a 2oz milk chocolate bar
OR half a digestive biscuit

Possible reasons for over-eating these days:

- cheaper foods (relative to incomes)
- more varied foods; supermarkets
- advertising and promotion of foods
- more eating outside the home (pubs, ethnic restaurants, fast food chains, etc)
- more fatty food; more snack foods
- over-eating because of anxiety (for example, work stress) or depression (for example, unemployment)
- grazing and irregular meal times
- fewer people now smoke (smoking suppresses appetite)

Possible reasons for under-exercising these days:

- more labour-saving machinery at work and at home (fork-lift trucks, power tools, washing machines, even automatic doors and lifts)
- television (couch potatoes)
- personal computers ("mouse potatoes") and email
- more cars; less walking and cycling
- less open space for recreation
- fear of violence in the streets
- central heating might also reduce energy expenditure

Prentice and Jebb[4] considered the case for Britain. The continuing series of food consumption data in households shows a reduction in average total energy intake per head since 1980. But this is food disappearance data (not food actually eaten) and does not include food eaten outside the home. The National Food Surveys, year by year, also show that consumption of fat (the most concentrated food energy) has

Email is more fattening than ice cream (Courtesy Roche with permission)

declined slightly since 1980. The evidence for reduced energy expenditure is more persuasive, though based on indirect indices. Car ownership and television viewing were strikingly related to the prevalence of obesity whether plotted against year or against socioeconomic status.

Facing moderate obesity

It is easy to diagnose moderate and gross obesity before the patient undresses. The difficulty with these patients is to organise the time and summon up the enthusiasm to embark on management which will be lengthy and may be unsuccessful. While our profession can transplant a human heart, manage *in vitro* fertilisation, and eliminate smallpox, to manage obesity is a challenge for a conscientious practitioner. A health service, like the British NHS, which provides an annual fee per patient is a suitable framework.

Obesity is different from most other diseases. If the general practitioner feels they are not the best person to look after a patient's haemorrhoids or backache or poor vision there is usually a specialist at the nearest big hospital who has the skills and equipment and will be happy to manage this part of the patient. But there is unlikely to be a consultant with special skills or equipment for looking after obesity. Obesity is a continuum and all but the most complicated or severe cases would not be welcome in the district hospital.

Many obese patients are referred to dietitians but there are too many obese people for the number of dietitians in Britain, and hospital staff are not well placed for looking after obese "outpatients"—that is, people who live at home and have to go out to work each weekday.

To put in the hard work of treating the obese people on the practice's list reduces the likelihood of having to treat them later for complications of obesity. The best conditions for managing obesity are:

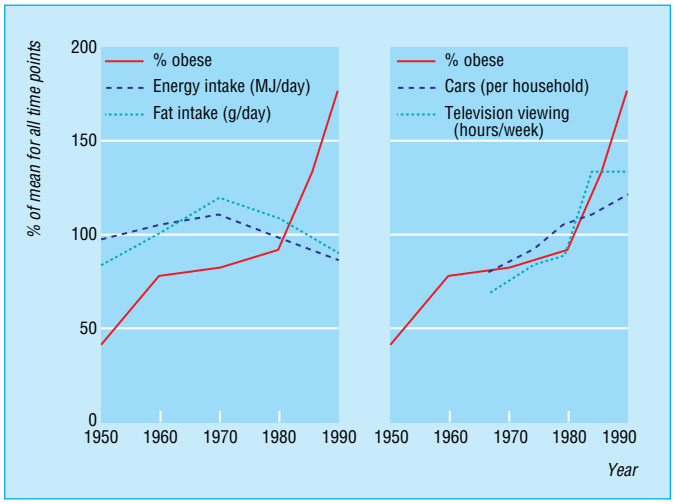
- regular visits, at least once a fortnight
- weighing the patient under the same conditions on the same scale
- about quarter of an hour's talk with the same practitioner each visit
- opportunity to bring wife, or husband
- the therapist is not obese.

Complications of obesity

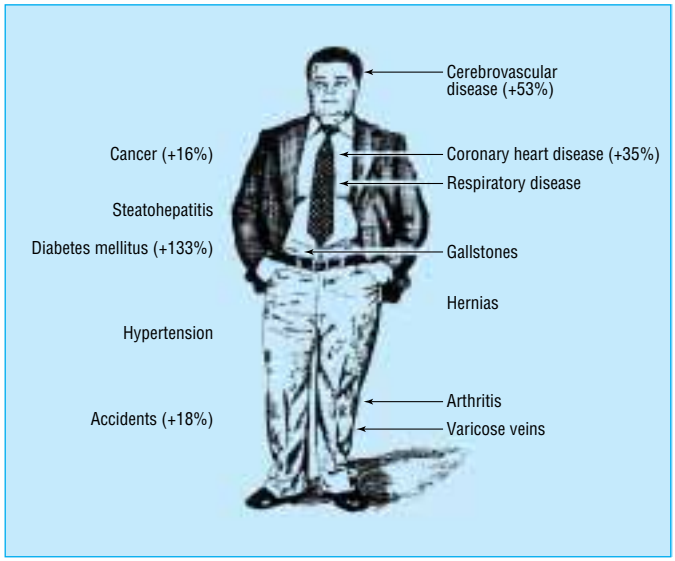
Most of the medical complications of obesity are well known. Risks of cardiovascular complications and diabetes are greater in people with abdominal obesity. This can be assessed clinically by measuring waist circumference. In men the normal measurement is up to 94 cm and metabolic complications are substantially increased at above 102 cm. The corresponding waist circumferences in women are: healthy <80 cm, increased risk 80-88 cm, substantially increased risk 88 cm.[5] Considerable weight gain in a short time carries greater risks than reaching the same weight slowly.[6]

The social complications are more immediate and may be very painful. Religious, racial, and gender prejudice are now unacceptable in mainstream Western societies, as is prejudice against the disabled. But prejudice against fat people is undisguised. It starts at school, affects the job opportunities and social status of obese people—and it is so pervasive that it probably colours many doctors' attitudes towards their obviously obese patients.

> Make less thy body hence, and more thy grace;
> Leave gormandising; know the grave doth gape
> For thee thrice wider than for other men.
>
> Shakespeare, *King Henry IV* Part II, V, v.



Secular trends in diet (left) and activity (right) in relation to obesity in Britain. (Adapted from Prentice and Jebb[4])



Obesity increases the risk of mortality from some of these diseases (notably diabetes) more than others

Substances secreted by adipose tissue[7]

Adipose tissue liberates free fatty acids and glycerol under the stimulus of catecholamines on lipase. When it loses fat in this way some of the stored cholesterol and accumulated non-polar foreign substances, for example DDT, also go into the circulation.

Since the discovery (1994) of **leptin**—the chemical signal to the brain of the amount of adipose tissue—it is realised that adipocytes are more than simple fat stores. They also secrete:

- tumour necrosis factor α, TNF α
- interleukin 6, IL6
- resistin
- angiotensinogen
- adipsin (complement factor D)
- tissue factor (initiator of coagulation)
- adiponectin
- aromatase, etc

TNF α, IL6 and resistin appear to increase insulin resistance,[8] the extra angiotensinogen makes hypertension more likely and aromatase converts androstenedione to the oestrogen estrone.

Recognising mild obesity/overweight

By contrast, for overweight or mild obesity diagnosis and recognition are where the management goes wrong. Yet there are several people with mild obesity for every case of obvious obesity in a practice. Mild obesity is much easier to treat, and prevention of gross obesity is much easier than cure.

People with mild obesity do not usually come to their practitioner complaining that they are too fat. The adiposity has been slowly creeping up on them and does not cause pain, distress, or fear. They come to the doctor with any other symptom and disease.

Patients may become mildly obese under their doctor's eyes, during the course of an illness which the practitioner is following and concentrating on: bed rest after myocardial infarction; giving up smoking; pregnancy and lactation; anxiety or depression from stress at work or at home. Recognition of overweight can be more difficult too in a patient in bed.

Have you got a minute doctor to sign this for me?

Measuring obesity

It is part of good practice to have a system so that all patients on the list are weighed at regular intervals by the receptionist or nurse. The weight is entered on the record. Knowing the patient's height (without shoes) the doctor can then decide whether he or she is: underweight, in the normal range, overweight, or obese. Comparison with previous weights shows if the patient is putting on weight.

Weight measurements are objective. To show the patient his or her weight against the standards for his or her height is impressive and, if outside the normal range, the basis for action. Obesity is an excess of adipose tissue. There are two methods for deciding whether someone is too fat. One is **social** or according to fashion. The doctor's role here is to advise patients not to make themselves too thin, not to take unphysiological diets or drugs to lose weight unnecessarily, and not to start on the road to anorexia nervosa.

The other method for deciding whether someone is too fat (or too thin) is **actuarial**. Above (or below) a range of weights for a given height the risk of developing illness or of mortality increases. Weight/height^2 (in kg/m^2) is a convenient index of relative weight. It gives a single number and is almost independent of height. This is the body mass index (BMI) or Quetelet's index. Note that smoking carries a higher risk of mortality than overweight or the milder degrees of obesity.

The lower and upper ends of the acceptable range here are close to, respectively, the lowest weight for small frame and highest weight for large frame for men in the earlier "desirable weights" of the Metropolitan Life Insurance Company of New York (1959).

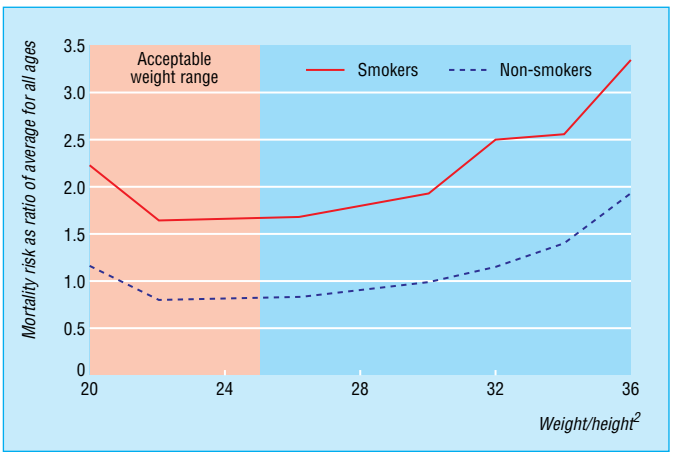
The unisex table on page 72 was introduced in the early 1980s. It is simpler than older standards. They had somewhat lower cut off points for women than men, but the numbers of women taking out life insurance were probably inadequate. Modern prospective data show no significant differences between men and women.

Measuring the patient

- Beam and lever scales are more reliable but take more space and are slower to use.
- For screening it is a good idea to have a lever scale in the reception area and for the receptionist or nurse to weigh the patient with shoes and coat removed.
- Obese patients being treated can be weighed (ideally in their underclothes) at each visit by the doctor in his consulting room on the same platform scale with a quick reading dial.
- The patient's height can be measured with a rule or tape measure attached to a flat wall and recorded in the notes. Heights are lower in the evening than in the morning. If the patient has any weakness or deformity an assistant is needed to get them as straight as possible for the measurement.

BMI cut offs in different ethnic groups

The range 18.5 to 24.9 may not be optimal for the whole world. Somewhat higher ranges probably suit Polynesians who tend to have more muscle, less fat. Somewhat lower ranges are probably optimal for south Asians who have insulin resistance at lower BMIs than Caucasians.[10]



Variations in mortality by weight among 750 000 men and women.[9] Cut off for obesity starts at BMI of 30

Obesity is taken in this table as generally starting at a BMI of 30 kg/m^2, which is 20% above the upper end of the acceptable range of weights for height. Between the top of acceptable weight and start of obesity is "overweight", which is the BMI range 25-29.9. The column at the far right shows weights for BMI of 40 kg/m^2, the start of "gross obesity". Adiposity is thus graded; this grading is needed for deciding management, and reflects prognosis.

The cut off values in the table have to be read from the viewpoint of the individual patient, like normal ranges for laboratory tests. People with gracile bones should weigh less than heavy boned people. Unusual muscle development (as in weight lifters) and oedema increase body weight. The waist circumference or waist/hip ratio should also be included in the assessment.

In children there are no actuarial data on which to base cut off weights for obesity. However, the upper percentiles of BMI references for British boys and girls by Cole *et al.*[11] can be used. These are based on a total sample of 30 500 children and adolescents (also see chapter 6, page 24).

Guidelines for body weight in adults

Height (without shoes)		Weight (kg) (minimal clothes)			
ft, in	(m)	Normal BMI 18.5-25	Overweight BMI 25-30	Obese BMI 30+	Grossly obese BMI 40+
4,9	1.45	39-53	53-64	>64	>85
4,10	1.48	40-55	55-65	>65	>88
4,11	1.50	42-55	55-68	>68	>90
5,0	1.52	43-58	58-69	>69	>93
5,1	1.54	44-59	59-71	>71	>95
5,1	1.56	45-61	61-73	>73	>97
5,2	1.58	46-63	63-75	>75	>100
5,3	1.60	47-64	64-78	>78	>103
5,4	1.62	48-66	66-79	>79	>105
5,5	1.64	50-67	67-81	>81	>108
5,5	1.66	51-69	69-83	>83	>111
5,6	1.68	52-71	71-85	>85	>113
5,7	1.70	53-73	73-87	>87	>116
5,8	1.72	55-74	74-89	>89	>119
5,9	1.74	56-76	76-91	>91	>121
5,9	1.76	57-78	78-93	>93	>124
5,10	1.78	59-79	79-95	>95	>127
5,11	1.80	60-81	81-97	>97	>130
6,0	1.82	61-83	83-99	>99	>133
6,0	1.84	63-85	85-102	>102	>136
6,1	1.86	64-87	87-104	>104	>139
6,2	1.88	65-88	88-106	>106	>141
6,3	1.90	67-90	90-108	>108	>145
6,4	1.92	68-92	92-111	>111	>148

Causes

Obesity **secondary** to hypothalamic conditions that increase appetite is rare, and to endocrine disorders uncommon.

Obesity may follow (*a*) enforced inactivity such as bed rest, arthritis, stroke, change to a less active job, sports injury, or (*b*) over-eating associated with psychological disturbances, for example, depression or anxiety, or some drugs that increase the appetite. Pregnancy and stopping smoking contribute to overweight.

Drugs that can increase appetite and promote weight gain

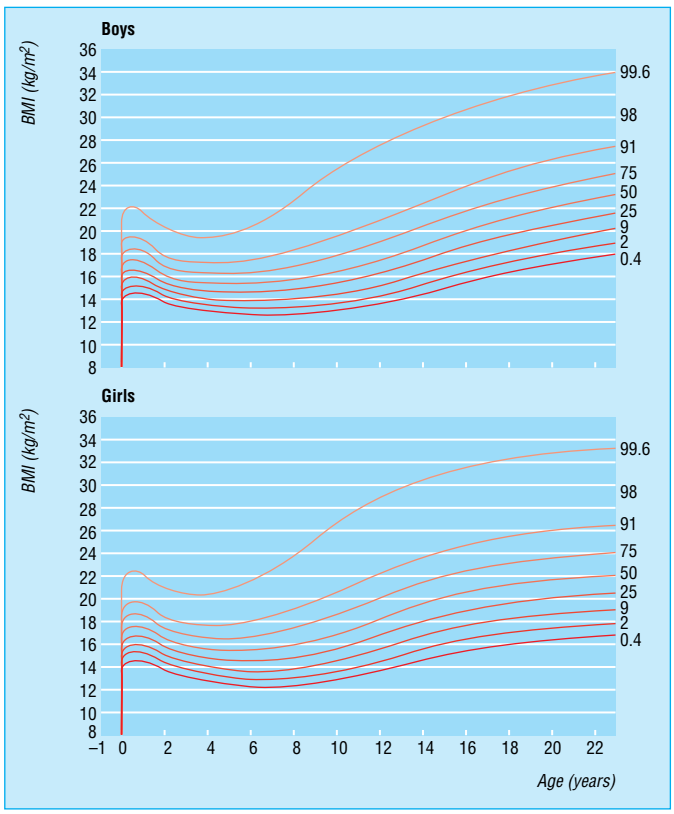
Corticosteroids, anabolic agents, some oral contraceptives, sulphonylureas, cyproheptidine, amitryptyline, clomipramine, lithium, pizotifen, metoclopramide, clozapine, olanzapine, some benzodiazepines

The genetic influence on obesity was clearly shown in a study using the Danish Adoption Register. A strong relation existed between the weight class of the adoptees (thin, acceptable, overweight, or obese) and the body mass index of their **biological**, but not their adoptive, parents. In twins body mass index is more strongly correlated between monozygotic than dizygotic twins, even when they are reared apart.[12]

Wide searches are going on to see if mutations of candidate genes for peptides and receptors involved in energy regulation are associated with obesity. Mutations of MC4R the melanocortin receptor, of β_1, β_2 and β_3 adrenoreceptors, of the leptin receptor and of leptin have been found associated with obesity. All are very rare except the MC4R mutation which is uncommon. In the great majority of obese people the weight gain is polygenic.

In most people obesity is **primary**. There is no obvious predisposing condition. If the patient says (s)he eats little: (1) the weight gain may have been in the past (2) some people undoubtedly need less food than others apparently comparable; they are efficient metabolisers with a low basal metabolic rate (but normal thyroid) and may feel the cold sooner than others, (3) repeated periods on low calorie diets, weight cycling, may lead to further adaptive lowering of the basal metabolic rate.

In surveys, obese people have often reported that they do not eat more than thin people. But the doubly labelled water technique (that measures rates of disappearance of the stable



BMI reference curves for boys and girls: 1990. Figures on right indicate percentiles (Adapted from Cole *et al.*[11])

isotopes, ^{2}H and ^{18}O) showed that the energy expenditure of a group of obese people was higher than their self-recorded food energy intakes. In other words, obese people tend to **under-report** their food intake.[13]

Regulation of energy balance

Many outside influences and numerous complex biological mechanisms inside the body affect energy balance. No diagram or model can adequately describe how most people's energy metabolism and fat stores are unconsciously regulated with little or no cumulative error. The system can be visualised at four levels: cerebral cortex, satiety, control centres in the hypothalamus, and thermogenesis.

Cerebral cortex
Eating behaviour is initiated by cultural and psychological influences: meal times, habitual diet, olfactory and visual stimuli, perception of appetite, social norms, foods available, personal relationships, etc.

Satiety[14]
Signals to stop eating come from distension and from chemical stimuli in the stomach and upper small intestine. These send afferent impulses via the vagus and cause secretion of gut hormones, for example cholecystokinin (CCK) is the most potent. Satiety signals go to the tractus solitarus nucleus in the brain stem, which has connections to the hypothalamus. Serotonergic receptors are involved in the brain.

Thermogenesis[17]
Animals that hibernate or adapt to cold have brown fat in which (unlike white fat) noradrenaline stimulates uncoupling of energy flow in the mitochondria, with production of heat instead of ATP. After infancy humans possess little or no distinct area of brown fat but are capable of some thermogenesis (increased metabolic rate)—on exposure to cold and with overfeeding. There appear to be varying numbers of brown adipose cells dispersed in white fat depots. The agent that catalyses the proton leak in these cells is Uncoupling Protein 1 (UCP1) which is triggered by β_3 adrenergic receptors and influenced by thyroid hormone. Two other uncoupling proteins UCP2 and UCP3 have been recently discovered in humans.

> It is not possible to conjure a unifying theory which will explain why some people become fat while others remain lean. This is too much to expect; economists have no simple explanation for the commercial success or failure of businesses or nations, and the energy economy of a human being is subject to as many influences as any financial economic model.[18]

Control centres in the hypothalamus[15]

Circulating leptin conveys information about the size and fluctuation of the adipose tissue stores to receptors in the hypothalamus. There are also receptors there for insulin. If leptin or insulin is injected directly into the brain it reduces food intake: deficiency of either hormone increases intake. Decreased leptin (reduced energy stores) activates neurones in the arcuate nucleus that express two neuropeptides that increase appetite: **neuropeptide Y** (NPY) and **agouti-related protein** (AGRP). Increased leptin inhibits these neurones and instead activates neurones that express **melanocortin**, which suppresses appetite. These different neuropeptides act on specific receptors in the paraventricular nucleus, notably Y1 and Y5 for NPY and MC4 for melanocortin. (Drugs that could block NPY receptors are already being studied.)

Ghrelin is a newly discovered peptide secreted by the stomach during fasting.[16] It acts on the hypothalamus to stimulate the neurones that express NPY and AGRP.

"Obesity cannot be prevented or managed solely at the individual level. Communities, governments, the media and the food industry need to work together to modify the environment so that it is less conducive to weight gain. Such partnerships are required to ensure that effective and sustainable changes in diet and everyday levels of physical activity can be achieved throughout the community."[19]

Management

There are three phases in the management.

- First is for the obese patient to reach an agreement with the health professional about what they believe they can achieve.
- Second is the difficult phase of the patient modifying eating and exercise habits to lose this much weight.
- Third is the longer and sometimes more difficult phase of the patient maintaining the new lower weight.

Three phases in management of obesity

- Agreement on what can be achieved
- Modification of eating and exercise habits
- Maintenance of new lower weight

First: agreeing on the objective
To reduce weight to the acceptable range is a reasonable aim that should be achievable for someone who is in the overweight category (BMI 25-30, or grade 1 obesity). With this grade of adiposity the weight loss required is only about 10 kg (22 lb).

But for someone considerably obese, BMI 35 or more, we know that most people cannot manage, with present resources, to reduce their food intake and increase their exercise sufficiently for long enough to bring their weight down all the way to the acceptable range. For someone with a BMI of 35 and a height of 175 cm this means taking off 31 kg (68 lb). It is better to set a **more realistic, intermediate goal** (which could always be extended later). A weight loss of 10 kg brings with it real health benefits.

For any overweight or obese patient the medical adviser can at least hope to help them to **hold their weight stable** and not let it creep up further. To achieve this may involve discovering why the patient has been putting on weight and discussing how these conditions can be modified.

Because there are now so many people in any general practice with overweight and obesity and these increase risks for many degenerative diseases, it is important work for the doctor (1) to recognise overweight and give early warning, and (2) in patients with obesity to take a structured therapeutic role.[20] For patients with type 2 diabetes (especially), hypertension, coronary disease, hernias, etc, action to reduce weight is not just preventive medicine, it is a central part of the treatment. The Nutrition Task Force and the Physical Activity Task Force of the UK Health of the Nation strategy proposed that prevention of obesity rather than its treatment should be a national priority.

Second: reducing weight

Two main treatments are indispensable: eating less dietary energy and increasing energy expenditure. Both are needed; they are synergistic. The cupboard of drugs for obesity is rather bare at present, and surgery—perhaps gastric stapling—should only be considered for the very fat minority.

You can lose weight only by achieving a cumulative negative energy balance. Calories in must be less than calories expended. An average loss of 2000 kJ (500 kcal) per day—14 000 kJ (3500 kcal) over a week—is equivalent to a loss of about 0.5 kg (1 lb) per week or more at the start, when water is lost.

The new way of eating

- Try to avoid the word "diet" that has discouraging overtones and suggests temporary hardship rather than a lifestyle that can last. Habits are more important than diets.
- It does not matter what you eat as long as it is less. This may sound irresponsible and obviously could not be taken to extremes (but there is sometimes too much technical detail in weight-loss programmes and malnutrition is a very rare complication).
- Eat smaller portions of what you usually eat, no second helpings, no snacks between meals.
- Have three regular meals each day. It is more difficult to control one's weight when meal occasions are irregular.
- Eat more low energy (calorie) foods—salads, fruits, vegetables, low fat milk, and yoghurt (left side of "energy values of foods" table on page 75).
- Avoid or ration energy (calorie)-dense foods with a high fat content—and fats in some foods are invisible; cakes, biscuits, cheese, chocolate, nuts, potato crisps, standard meats, chicken skin, and of course fried foods.
- The household should be involved and supportive. Husband and wife or partners should share in discussions with the doctor and dietitian.

Likely benefits of 10 kg weight loss[20]

- *Blood pressure*
  Fall of 10 mmHg systolic — Fall of 20 mmHg diastolic
- *Diabetes*
  Fall of 50% in fasting blood glucose
- *Plasma lipids*
  Fall of 10% total cholesterol — Fall of 15% LDL
  Fall of 30% triglycerides — Rise of 8% HDL
- *Mortality*
  Fall of >20% total mortality

Starting measurements

If an agreement has been reached between patient and doctor that they will start a weight-loss programme this is a good time to make baseline measurements that are likely to improve. As well as accurate height and unclothed weight, hence BMI, waist circumference should be measured and fasting blood taken for plasma lipids (including triglycerides) and glucose.
Risks of metabolic complications are greater in people with abdominal obesity[21]

- For women: increased risk above 80 cm waist circumference (32 inches) substantial risk above 88 cm (35 inches)
- For men: increased risk above 94 cm waist circumference (37 inches) substantial risk above 102 cm (40 inches)

Successes

Klem *et al.*[22] collected records of 629 women and 155 men in the United States who successfully lost at least 30 lb (13.6 kg) and maintained the weight loss for at least one year. Their average starting BMI was 35, they lost an average of 30 kg and their new BMI was 24.5 kg/m^2.

A little over half of this sample lost weight through formal programmes; the rest lost weight on their own. All of them used diet **and** exercise. Three-quarters of them reported that a triggering event had preceded their successful weight loss. Some limited classes of foods, some limited quantities of foods, some counted calories. They said that their weight loss led to improvements in energy, physical mobility, mood, self-confidence, and physical health.

Satiety

Satiety is a function that has been neglected in weight reduction programmes. People stop eating because of internal feelings of satiety, "I've had enough, thank you". But this feeling of having eaten enough is not directly related to the energy (calories) of a meal or food. Some foods give stronger satiety per MJ than others. High fat foods have weak satiety for their energy values compared to predominantly carbohydrate foods.[23] Short-term studies indicate for example that boiled potatoes have twice the satiety effects of chips for the same calories and fruits have greater satiety effect than biscuits or confectionery.[24]

Representative energy values of (stated) typical servings of some common foods

	kJ	kcal		kJ	kcal		kJ	kcal
Lettuce (30 g)	17	4	Bread (1 slice, 30 g)	280	67	Biscuits (2 digestive)	670	160
Cucumber (45 g)	21	5	Egg (1 boiled)	305	73	Peanuts (30 g)	711	170
Carrot (50 g)	42	10	Banana (1 fruit)	334	80	Avocado (1/2 fruit)	711	170
Cauliflower (90 g, raw)	50	12	Beer (1/2 pint)	360	86	Chicken roast, meat only, 120 g)	744	178
Tomato (90 g)	54	13	Wine (125 g)	393	94	Cheese (Cheddar, 45 g)	752	180
Grapefruit (1/2, no sugar)	75	18	Yoghurt (carton, low fat)	418	100	Chocolate biscuits (2)	794	190
Milk (full cream, 30 g, in tea)	84	20	Cornflakes (30 g)	460	110	Beef steak (grilled, 150 g)	1003	240
Sugar (1 level teaspoon)	84	20	Butter (15 g)	460	110	Potato crisps (50 g)	1108	265
Crispbread (10 g)	134	32	Fish (cod, grilled, 120 g)	481	115	Rice (75 g, raw)	1129	270
Jam (15 g)	168	40	Potatoes (boiled, 150 g)	481	115	Macaroni (75 g, raw)	1170	280
Orange juice (120 g)	192	46	Carbonated soft drink (323 ml)	543	130	Sponge cake (65 g)	1212	290
Apple (100 g)	192	46	Dates (60 g)	585	140	Fish (fried in batter, 120 g)	1338	320
Peas (90 g)	209	50	Baked beans (240 g)	606	145	Chips (fried, 180 g)	1902	455
Whisky (25 g)	234	56	Milk chocolate (30 g)	648	155	Pork chops (fried, including fat, 210 g with bone)	2257	540

- Simple food tables can help in planning lifestyle, shopping, meals, and menus. But note that they give kilojoules (kilocalories) and fat per 100 g. Typical servings of some foods are less, of other foods more than this (see "energy values of foods" table).
- A dietitian, preferably nearby and linked with the practice, can help with detailed suggestions.

Crash diets do not work.

A weight loss of 1 kg (2 lbs) per week is the most that can be expected but half this is useful (and all that some can manage).

A valuable technique for managing obesity is modification of eating behaviour. Its introduction by RB Stuart in 1967 changed the expectation of treatment from poor to fair and refinements are improving prospects further.

Firstly the patient makes notes of everything he or she eats for a week and where they were at the time, what they were doing, and how they felt. The calories can be worked out later. The therapist guides and encourages the patient to fill in the form. In the process the patient discovers in what circumstances he or she eats most. The doctor or dietitian discusses the completed form with the patient and suggests behaviour modifications. Obese people do not eat so much because they are hungry, but more in response to external cues—boredom, anger, delicious taste, other people eating, food that would be wasted, etc. The patient can make arrangements to minimise these cues.

Time	Food	Taste rating	Amount	kcal	Where?	Who with?	Mood	Hunger	Associated activity	Why eaten?
4·45pm	Chocolate Cake	My weakness V. Good	3 slices	510	Kitchen	Children	Fed up	No	Giving children their tea	Irritated
5·00pm	Egg Sandwich	Nothing special	1½ sandwiches	200	Kitchen	Alone	OK	No	Clearing	Leftovers

Record of what is eaten, where and when

Rules for modifying eating behaviour

- Buy non-fattening foods. Do not buy foods that specially tempt you. Use a shopping list and stick to it. Do not shop when you are hungry.
- Always eat in one room, in only one place in that room—for example, seated at the dining table—and avoid other activities (except conversation). Make eating a pure experience.
- Look for times when you are most likely to eat unnecessarily—for example, when giving children their tea, or because you cannot bear to throw food away—and take steps to change your routine.
- Always have nearby a variety of low calorie foods to use as snacks—like raw vegetables.
- Recruit others to help you curb your eating—spouse, coworkers, friends—they can help most by praising when you do not over-eat.

Fatty and thinny eating side by side

- Build in rewards for sticking to the programme—something you would like to do, or a present. Family and friends are usually happy to cooperate. Of course, the reward cannot be a meal or food.
- Make small portions of food appear to be large (small plate, food cut up and spread all over it). Make second helpings hard to get; do not keep serving dishes on the table. Leave the table as soon as you have eaten.
- Slow down the rate at which you eat. Chew each mouthful for longer. Always use a knife and fork or a spoon and put them down between mouthfuls. Swallow one mouthful before the next.
- Take steps to minimise hunger, loneliness, depression, boredom, anger, and fatigue, each of which can set off a bout of overeating. This needs discussion and planning. Hunger is minimised by three regular meals daily.
- Increase the exercise you take each day.
- Keep a record of how much you eat and exercise, and of your weight.

Exercise

During an hour's walk at ordinary speed, mostly on the level, 21 kJ/min × 60 = 1260 kJ (or 5 × 60 = 300 kcal) of energy are used up. But at rest about 4.2 × 60 = 252 kJ (1 × 60 = 60 kcal) would be spent per hour. The energy used by going for an hour's walk is therefore the difference between 1260 and 252 = 1008 kJ (300 and 60 = 240 kcal). This is equivalent to about two slices of bread and butter (60 g bread + 15 g butter, see "energy value of foods" table). People can be discouraged by the small amount of food which is directly equivalent to the use of quite a lot of precious time taking exercise.

There are, however, additional benefits from increased regular exercise.

(1) Obese people, with a heavier body to move, use more energy than in the table for the same amount of work.
(2) Exercise can be valuable as a diversion from sitting indoors and being tempted to eat.
(3) Exercise is more likely to reduce than increase appetite.
(4) After exercise, the resting metabolic rate may increase for some hours (though the effect is evidently small, and some experimenters have not been able to measure it).
(5) When exercise is taken after meals the thermic effect of the meal may be increased.

Exercise also limits the proportion of lean tissues lost in slimming programmes and helps in weight maintenance. Fat people cannot usually manage vigorous sports and may be embarrassed to dress in sports gear. Walking in all its varieties is the most important exercise.

Exercise advice could include[20]:

- Walking is the key to controlling your weight
- Travel whenever possible by foot and aim for 30 minutes' brisk walking per day
- Walk all or part of your journey to work or the shops
- If you usually travel by bus get off a stop earlier
- Use the stairs instead of a lift
- Avoid sitting for long periods; be active during TV adverts
- If you have a garden, spend more time working in it
- If you have a dog, take it for more frequent or longer walks

Third: maintaining the lower body weight

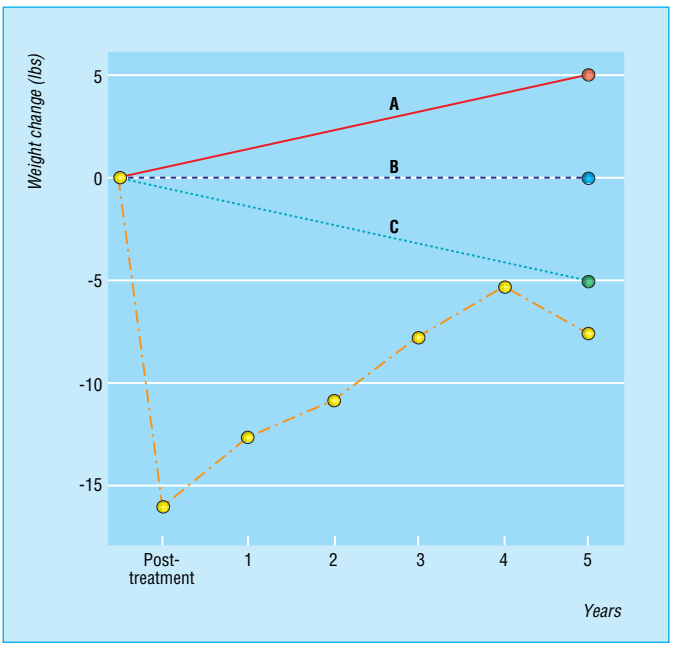
After a weight-loss programme with the new eating plan, and increased exercise and behaviour modification for 12-20 weeks, the motivated and well-supported patient should have lost near to 10 kg. Now starts the longer phase of keeping the weight off. This is where temptations and disappointment lie in wait.

Walking the dog

Energy/minute used in activities
(rounded approximate figures)

	kJ	kcal
At rest (Men +10%, women −10%)	4	1
Moderate exercise For example, walking, gardening, golf	21	5
Intermediate For example, cycling, swimming, tennis	29	7
Strenuous For example, squash, jogging, hill climbing, heavy work	42	10

In SI units: 1 kcal/min = 4.2 kJ/min = 70 watts



Mean weight loss of behavioral programmes post-treatment and at 1, 2, 3, 4, and 5 years follow up. Lines A, B, and C represent hypothetical groups of people against which to compare the results of behavioral programmes. Line A, natural history of obese persons receiving no treatment as gradually weight is gained; line B, stable weight; line C, gradual loss. (Adapted from Brownell and Kramer[25])

The graph on page 76 shows combined average results of several behavioural programmes.

Weight loss was satisfactory but gradually the subjects' weight slowly crept up so that after five years about half the weight lost had been put back on. The same slow regain of weight has been reported whatever treatment was used for the weight loss phase (very low calorie diet, appetite suppressant drugs, etc). The challenge, therefore, is for therapist and subject to be aware of this tendency to re-gain and work to minimise it.

If the weight-loss diet was drastic and incompatible with a normal lifestyle, food habits will rebound. Some restriction of diet has to continue in the maintenance phase. An *ad lib* low fat (high carbohydrate) diet has been shown to be more effective for weight maintenance than a fixed energy intake.[26] An active lifestyle with regular adequate exercise seems to be essential to keep body weight down.[27]

Other treatments

Drug therapy

It is futile to prescribe drugs unless the patient is committed to eating less calories and/or exercising more. Only a limited range of drugs are available at present, though pharmaceutical companies are researching hard to find safe additional types.

Orlistat ("Xenical") is an inhibitor of pancreatic lipase. It reduces digestion of fat (triglycerides) and increases faecal fat. Oily spotting and faecal urgency can be controlled by eating a low fat (so lower energy) diet.

Sibutramine inhibits the reuptake of serotonin and noradrenaline in the central nervous system. It decreases appetite and may increase thermogenesis (metabolic rate). Blood pressure can rise.

The serotonergic appetite suppressants fenfluramine and dexfenfluramine have been withdrawn because of rare but severe cardiac valve disease and pulmonary hypertension. Of the catecholaminergic drugs, amphetamines cannot be used because of their potential for abuse. Only phentermine is in the *British National Formulary*. It should only be used to support someone with moderate to severe obesity who is restricted to a 12-week prescription.

Commercial organisations, for example, Weight Watchers or similar organisations, use group therapy and can help some people with mild to moderate obesity. The Scottish Obesity guidelines[20] have a set of useful criteria for evaluating the local organisation of this type. Other patients prefer a more individual approach.

More radical treatments sometimes used are **very low calorie diets** and **gastric stapling**. They are in the specialist's domain. Ideally patients with a BMI over 35 kg/m^2 should be referred to a specialist centre,[20] but these are not well developed in all areas.

References

1 McLellan F. Obesity rising to alarming levels round the world. *Lancet* 2002; **359**: 1412.
2 Caterson I. Obesity 1998 – has anything changed? *Modern Med* 1998; **41**: 52-66.
3 Bray G. Drug treatment of obesity: don't throw the baby out with the bath water. *Am J Clin Nutr* 1998; **67**: 1-4.
4 Prentice A, Jebb SA. Obesity in Britain: gluttony or sloth? *BMJ* 1995; **311**: 437-9.
5 Report of a WHO Consultation. *Obesity: Preventing and managing the Global Epidemic*. Geneva: WHO, 2000.
6 Wannamethee G, Shaper AG. Weight change in middle-aged British men: implications for health. *Eur J Clin Nutr* 1990; **44**: 133-42.
7 Trayhurn P, Beattie JH. Physiological role of adipose tissue: white adipose tissue as an endocrine and secretory organ. *Proc Nutr Soc* 2001; **60**: 329-39.
8 Flier JS. The missing link with obesity. *Nature* 2001; **409**: 292-3.
9 Lew EA, Garfinkel L. Variations in mortality by weight among 750 000 men and women. *J Chron Dis* 1979; **32**: 563-76.
10 Whincup PH, Gilg JA, Papacosta O *et al*. Early evidence of ethnic differences in cardiovascular risk: cross sectional comparison of British South Asian and white children. *BMJ* 2002; **324**: 635-8.
11 Cole TJ, Freeman JV, Preece MA. Body mass index reference curves for the UK, 1990. *Arch Dis Childhood* 1995; **73**: 25-9.
12 Stunkard A, Harris JR, Pedersen NL, McClearn GE. The body mass index of twins who have been reared apart. *N Engl J Med* 1990; **332**: 1483-7.
13 Prentice AM, Black AE, Coward WA. High levels of energy expenditure in obese women. *BMJ* 1986; **292**: 483-7.
14 Blundell JE, Goodson S, Halford JCG. Regulation of appetite: role of leptin in signalling systems for drive and satiety. *Int J Obesity* 2001; **25** (Suppl 1): S29-34.
15 Schwartz MW, Woods SC, Porte D, Seeley RJ, Baskin DG. Central nervous system control of food intake. *Nature* 2000; **404**: 661-71.
16 Pinkney J, Williams G. Ghrelin gets hungry. *Lancet* 2002; **359**: 1360-1.
17 Lowell BB, Spiegelman BM. Towards a molecular understanding of adaptive thermogenesis. *Nature* 2001; **404**: 652-9.
18 Garrow J. Luxuskonsumption, brown fat, and human obesity. *BMJ* 1983; **286**: 1684-6.
19 Report of a WHO Consultation. *Obesity: Preventing and Managing the Global Epidemic*. Geneva: WHO, 2000: p241.
20 Obesity in Scotland. *Integrating prevention with weight management. A National Clinical Guideline recommended for use in Scotland*. Edinburgh: Scottish Intercollegiate Guidelines Network (SIGN), 1996.
21 Després J-P, Lemieux I, Prud'homme D. Treatment of obesity: need to focus on high risk abdominally obese patients. *BMJ* 2001; **322**: 16-20.
22 Klem ML, Wing RR, McGuire MT, Seagle HM, Hill JO. A descriptive study of individuals successful at long-term maintenance of substantial weight loss. *Am J Clin Nutr* 1997; **66**: 236-46.
23 Blundell JE, MacDiarmid JI. Fat as a risk factor for overconsumption: satiation, satiety and patterns of eating. *J Am Dietetic Assoc* 1997; **97** (Suppl): S63-9.
24 Holt SHA, Brand Miller JC, Petocz P, Farmakalides E. A satiety index of common foods. *Eur J Clin Nutr* 1995; **49**: 675-90.
25 Brownell KD, Kramer FM. Behavioural management of obesity. In: Blackburn GL, Kander BS (eds). *Obesity: pathophysiology, psychology and treatment*. New York: Chapman & Hall, 1994.
26 Toubro S, Astrup A. Randomised comparison of diets for maintaining subjects' weight after major weight loss: ad lib, low fat, high carbohydrate diet v. fixed energy intake. *BMJ* 1997; **314**: 29-34.
27 Anderson JW, Konz EC, Frederich RC, Wood CL. Long-term weight-loss maintenance: a meta-analysis of US studies. *Am J Clin Nutr* 2001; 74: 579-84.

12 Measuring nutrition

Energy (calories)

Methods of measuring calories (food energy) are different from those for the other essential nutrients.

1 kcal = 4.184 kJ
kilojoules and megajoules are the SI units

Energy expenditure

Measurement of energy balance is technically difficult. *Energy expenditure* can be measured in one of five ways.

(1) By measuring a subject's heat output in a special insulated room. This is **direct calorimetry**, a costly and complicated experiment. There are very few direct calorimeter rooms anywhere in the world. The method has been replaced by (2).

(2) **Indirect calorimetry** measures oxygen consumption and the production of CO_2 and urinary nitrogen. From these the mixture of carbohydrate, fat (exogenous or endogenous), and protein metabolised can be calculated. Urinary N indicates protein catabolism and CO_2/O_2 (the respiratory quotient) indicates the ratio of carbohydrate to fat metabolised. Energy production per gramme differs with the metabolic fuel (greater when fat is oxidised). These measurements can be made for up to 24 hours in a *respiration chamber*, in which a subject can live and carry out various somewhat restricted activities.[1]
The energy produced per litre of oxygen consumed happens to be much the same whichever of the three macronutrients is oxidised. Oxygen consumption alone gives an adequate quantification of energy expenditure *over short periods*. It can be measured with different types of apparatus: Douglas bag, or a portable electronic respirometer carried on the back, a ventilated hood or a bedside metabolic monitor. The former are used to estimate the energy cost of different activities. The latter apparatus is used clinically to measure resting or **basal metabolic rate**, which can be increased in patients—for example, after burns or trauma or with infections.

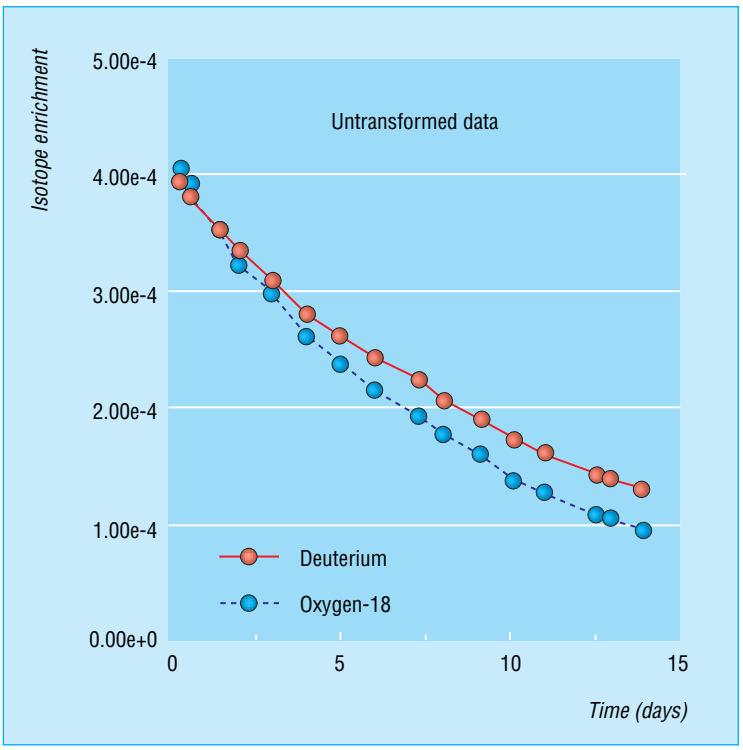
(3) The **doubly labelled water** method measures the decay of body water concentrations of the stable isotopes ^{2}H and ^{18}O. It is an ingenious method, developed for human work in the 1980s, suitable for measuring total energy expenditure over 7-10 days. It depends on the principle that labelled oxygen is lost partly as carbon dioxide and partly as water while the rate of water loss is given by the decay of labelled hydrogen.[2] The isotopes are given as a very accurately weighed dose of water. Disappearance is measured (by dedicated mass spectrometer) in repeat timed urine specimens. From CO_2 production (knowing the average respiratory quotient from the diet record and estimate of any changes in body fat) energy expenditure can be calculated as for indirect calorimetry. Unfortunately the heavy oxygen is very expensive.

(4) It ought to be possible to assess energy expenditure from an estimate of basal metabolic rate (from the patient's age, sex, and weight) and from recording all his activities (lying, sitting, walking upstairs, etc) throughout the day. Energy values for the different activities have to be assumed from reported values. The method is tedious and inaccurate.

(5) The possibilities for estimating energy expenditure are better by counting the heart rate for a day or more with a small portable cardiac monitor. (Practical details for this and the other methods are reviewed by Murgatroyd *et al.*[2])

The nutritional state for energy, or for any essential nutrient, depends on the balance (B) between intake (I) (dietary or parenteral) and output (O) or expenditure: B = I − O.

When the balance is negative the nutritional state tends to go down towards depletion, but there may be an adaptive reduction in output (losses). When the balance is positive the nutritional state tends to go up: the nutrient may be stored somewhere in the body but some nutrients can start to become toxic. "You can have too much of a good thing."



Doubly labelled water measurement of energy expenditure over 14 days. Isotope disappearances from a typical adult subject[2]

Energy intake

For *energy intake* the energy (calorie) contents of foods eaten should ideally be analysed chemically by measuring the protein, available carbohydrate, fat, and alcohol contents and multiplying each of these by their metabolisable energy values. Energy values in food tables are only estimated averages.

The usual way of estimating *energy balance* is, of course, from the resultant **changes in body weight**. A gain or loss of energy by the body of about 25-29 MJ (6000-7000 kcal) should, respectively, increase or reduce the weight by 1 kg. Most of this weight change is in fat, with a variable amount of water and a minority of muscle. Gain or loss of tissue needs to be over 1 kg to be detectable, because, even with accurate weighing and on a constant regimen, healthy people's weights fluctuate within the day and from day to day. Furthermore, many scales and weighing techniques are not accurate.[4]

Reference standards for body weights at different heights based on body mass index (BMI) are given for adults in chapters 8 and 11 and for children in chapter 6.

It is difficult to weigh deformed or paralysed people. Very sick bedfast people cannot be weighed unless they are in a specially designed weighing bed. It is nearly always easy to weigh an ambulant patient—it just takes a little time and trouble. But in hospital patients confined to bed—for example, those with fluid lines or splints—some idea of loss of tissue can be obtained by measuring the arm circumference with or without one or more skinfold thicknesses.

Measuring arm circumference needs only a tape measure, put round the arm (preferably left) midway between the tip of the acromion and the olecranon. The enclosed area is made up of muscles and subcutaneous fat over the constant humerus.

Reference standards for mid-upperarm circumference (mm)[5,6]

	Men			Women		
		Centiles			Centiles	
Age	50th	10th	5th	50th	10th	5th
19-24	308	272	262	265	230	221
25-34	319	282	271	277	240	233
35-44	326	287	278	290	251	241
45-54	322	281	267	299	256	242
55-64	317	273	258	303	254	243
65-74	307	263	248	299	252	240

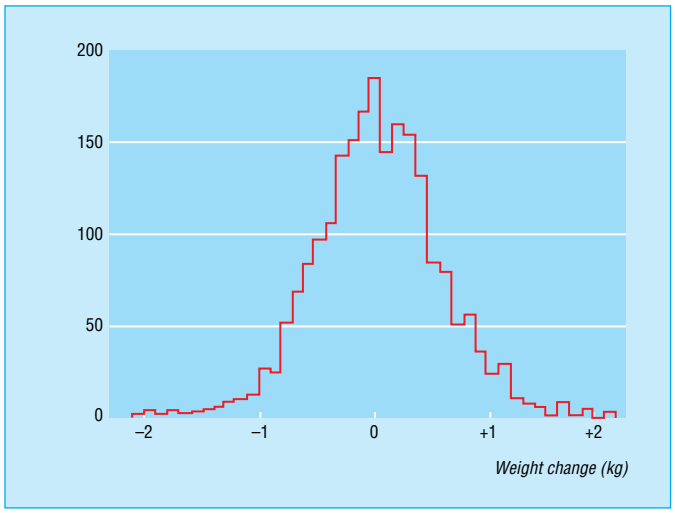
Measuring skinfold thickness requires special calipers. The best are obtainable from Holtain Ltd, Crymych, Dyfed SA41 3UF. Cheap plastic ones are unreliable. The most common sites for measuring skinfold thickness are over the *mid-triceps*, halfway between the acromion and the olecranon, and the *subscapular* skinfold, 1 cm below the inferior angle of the scapula. Reference standards are available for skinfolds at these sites (see table opposite).

The triceps skinfold measurement can be used to calculate the mid-upperarm muscle circumference, AMC.

$$\text{AMC} = \text{arm circumference} - \pi \times \text{triceps skinfold (mm)}$$

Before substantial weight changes are detectable, or if weight cannot be measured, a rough idea of energy balance can be obtained by calculating the energy per day in recent food intake from food tables and comparing this with the estimated average requirements for energy published by the DoH.[7]

These reference values for energy intake are estimates and averages. They differ with gender, age, body weight, and



Distribution of day to day weight change (kg) on 2078 occasions on consecutive days in healthy young men[3]

Weight measurement

Beam or lever balances are most reliable, not the usual bathroom scales. The best scales are not easily portable: they are heavy and the knife edge balancing part can be damaged. The patient should if possible be brought to the scale. It should be serviced regularly (as recommended by the manufacturer) and checked frequently against known weights, such as a heavy weight kept nearby. Subjects should be weighed in light underclothing and no shoes and with any heavy jewellery removed. A meal or full bladder increases the reading and a bowel action reduces it.

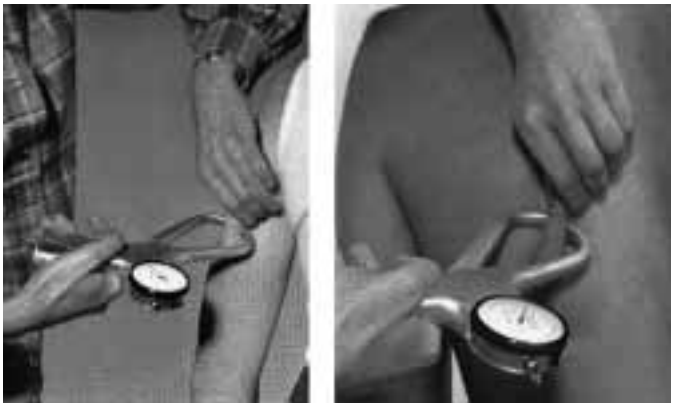
Adults and older children should stand straight on both feet on the centre of the scale's platform without touching anything else.

Pictures of how to weigh young children are shown in Valman B, Thomas R. *ABC of the First Year*, 5th edn. London: BMJ Books, 2001.

Reference standards for triceps skinfold thickness (mm)[5,6]

	Men			Women		
		Centiles			Centiles	
Age	50th	10th	5th	50th	10th	5th
18-24	9.5	5	4	18	11.5	10
25-34	12	6	4.5	21	12	10
35-44	12	6	5	23	14	12
45-54	12	6	6	25	16	12
55-64	11	6	5	25	16	12
65-74	11	6	4	24	14	12

*Figures based on a large sample of healthy US citizens from Frisancho[5] and Bishop *et al.*[6]



Measuring skinfold thickness with Holtain calipers: left—triceps site; right—subscapular site

especially with activity level.[7] About half of any group of similar people can be expected to need more, and half to need less, than these values. Individual requirements might well range from 50% to 150% of these averages. Furthermore, on any single day an individual might consume considerably less or more than their average daily intake.

There are no biochemical tests that reliably indicate energy balance. Acetone appears on the breath in people who fast for longer than 12 hours and β hydroxybutyrate concentrations increase in body fluids but these cannot be used to show the extent of energy deficit.

Protein

Tests for protein status estimate two variables: total body protein and visceral protein.

Total body protein

This is predominantly muscle and can be estimated in several ways. At the clinical level:

- Body weight for height —that is, BMI, partly reflects body protein. Though the usual cause of a high BMI is excess fat, a less common cause is unusually well developed muscles in body builders and weight lifters.
- Total protein may also be estimated approximately from the mid-upperarm muscle circumference (mid-upperarm circumference − $\pi \times$ triceps skinfold).
- Twenty-four-hour urinary creatinine gives a biochemical measure of muscle mass because creatinine is a metabolite from the turnover of muscle creatine. One gram of creatinine/day comes from about 20 kg of muscle but urinary creatinine shows quite large day to day fluctuations even if collection is complete, and it is spuriously increased by eating meat and by exercise.

Three research procedures are potentially available to give the most reliable estimates of lean body mass.

- Total body nitrogen may be measured by *in vivo* neutron activation with simultaneous counting of the 10.8 MeV gamma rays produced from the protein nitrogen[9]; then $N \times 6.25$ = total body protein.
- Potassium is predominantly intracellular and normally proportional to body protein. Total body potassium can be measured non-invasively by placing a subject in a heavily screened whole-body counter and counting the natural weak gamma-ray emissions from the subject's own ^{40}K which is mixed as 0.012% with the stable ^{39}K throughout our bodies.
- Body density can be measured by underwater weighing in a special bath (a harmless procedure but only available in research centres). As the density of body fat is 0.9 and of lean body mass 1.1, the proportions of fat and lean can be calculated.

Other methods are available which are more accessible investigations but tend to be less precise.

- Use of bioelectrical impedance depends on the greater electrical conductivity of lean body mass than fat. A weak current is passed from ankle to hand. The equipment is portable and relatively inexpensive.
- Imaging methods such as dual *x*-ray absorptiometry can also provide estimates of body fat and lean.
- From total body water, by dilution of deuterium, lean body mass can be estimated because body water is in the lean not the fat.
- Total body fat can be estimated approximately by applying a formula to the sum of four skinfolds.[10]

Estimated average requirements of food energy for people in the United Kingdom[7]

Age	Males MJ/d (kcal/d)	Females MJ/d (kcal/d)
0-3 months	2.28 (545)	2.16 (515)
4-6 months	2.89 (690)	2.69 (645)
7-9 months	3.44 (825)	3.20 (765)
10-12 months	3.85 (920)	3.61 (865)
1-3 years	5.15 (1230)	4.86 (1165)
4-6 years	7.16 (1715)	6.46 (1545)
7-10 years	8.24 (1970)	7.28 (1740)
11-14 years	9.27 (2220)	7.92 (1845)
15-18 years	11.51 (2755)	8.83 (2110)
19-59 years	**10.60 (2550)**	**8.05 (1920)**
60-74 years	9.82 (2355)	7.98 (1900)
75+ years	8.77 (2100)	7.61 (1810)

Values from 0 to 18 years are based on average energy intakes. From 19 years on, however, they are based on measured energy expenditure, assuming low physical activity levels at work and leisure (physical activity level = 1.4 × BMR (basal metabolic rate)). More active people need to consume more than the figures here. In the light of recent measurements, the figures for children under 5 years may be too high by about 10%[8]

Reference nutrient intake for protein*, based on British dietary reference values[7]

Age and sex	Body weight (kg)	Protein (g/kg body weight)	Reference nutrient intake (RNI) (g/day)
0-3 months	5.9	—	13
4-6 months	7.7	1.7	13
7-9 months	8.8	1.6	14
10-12 months	9.7	1.55	15
1-3 years	12.5	1.2	15
4-6 years	17.8	1.1	20
7-10 years	28.3	1.0	28
Males			
11-14 years	43.0	1.0	42
15-18 years	64.5	0.85	55
19-50 years	74.0	0.76	56
50+ years	71.0	0.75	53
Females			
11-14 years	43.8	0.94	41
15-18 years	55.5	0.81	45
19-50 years	60.0	0.75	45
50+ years	62.0	0.76	47
Pregnancy			+6
Lactation			+11

British dietary reference values are based on FAO/WHO (1984) safe levels of protein intake.
*These recommended intakes assume that the protein comes from a mixed diet, as in the average British diet. Requirements may be higher if digestibility of the protein is incomplete or if one (or more) of the indispensable amino acids is poorly represented

Visceral protein

Visceral protein is sometimes disproportionately reduced in protein deficiency, as seen most strikingly in kwashiorkor. There is fatty liver, intestinal mucosal and pancreatic atrophy, and impaired lymphocyte function. The usual tests measure concentrations of plasma albumin or transferrin, proteins synthesised in the liver. These are disproportionately reduced when there is visceral protein depletion and often within normal limits in total body protein depletion. Plasma albumin concentrations are also moderately reduced in the metabolic response to severe injury or infection as well as in liver cirrhosis and nephrotic syndrome. Transferrin concentrations are increased in iron deficiency.

Other nutrients

Other nutrients do not affect anthropometric measurements directly. Inadequate intake can be suspected from the dietary intake or shown by specific biochemical tests.

Assessment of dietary intake

The sequence of assessing nutrient intake is to:

- estimate **food intake** as g/day of different foods
- use **food tables** to convert g/day of each individual food to g, mg, or μg/day of various nutrients. These calculations can, and always used to be done manually but computers are generally used
- compare this patient's intake of one or several nutrients likely to be inadequate with the dietary reference values[7] of food energy and nutrients. Note: the reference nutrient intake[7] covers the individual nutrient requirements of the great majority of normal people, which means that an intake somewhat below the recommended daily amount would be adequate for most people (see page 85). The other consideration is whether the day(s) on which food intake was measured were typical.

Food intake measurements

Food intake data may be obtained at the national, household or individual level. **National figures** come from food production plus imports minus exports, divided by the estimated national population. They report **apparent consumption** because some of the food is wasted, fed to tourists or pets, used in industry, etc. All developed countries publish annual figures for average consumption. **Household food consumption** has been measured in large samples across England, Wales and Scotland every year since 1940 and the National Food Survey provides a unique continuous series of national consumption of different foods.[11] The survey cannot show average individual intakes because until recently food and drink consumed outside the home was not estimated, and distribution of foods and nutrients within families varies between its members.

Although these macrofigures provide a background of trends in food consumption for clinical work and analytical epidemiology, we want to know what an **individual** has been eating and drinking. There are four types of method used to estimate individual food intake.

- **Dietary history—**"What do you eat on a typical day?" This is a good method in the hands of a skilled and patient interviewer. Food models, cups, plates, and spoons are used to estimate portion sizes.
- **24-hour recall—**"Tell me everything you've had to eat and drink in the last 24 hours." This is less subject to wishful thinking about what the person ought to have eaten. The weakness is that yesterday may have been atypical; 24-hour recalls can, however, be repeated.

Checklist on patient's diet

In a busy practice a short checklist about the patient's eating habits is a useful screening method.

- What is your appetite like?
- Do you eat more or less than other comparable people?
- Has what you eat changed—type of food or amount?
- Are you on a special diet?
- Is there any food you can't eat because it doesn't agree with you?
- Are you losing or gaining body weight?
- What do you usually have for the main meal of the day?
- Do you eat meat/fruit/fat on meat/salt/etc?
- What sort of bread do you eat?
- What sort of alcohol do you drink, how many drinks per week or per day?
- What do you have for breakfast and lunch?
- Do you take vitamin or mineral tablets?

Individual food intake

There are four types of method used to estimate individual food intake, and these are:

- Dietary history
- 24-hour recall
- Food dietary or record
- Food frequency questionnaire

3. (CONTINUED) PLEASE FILL IN YOUR **AVERAGE** USE, **DURING THE PAST YEAR**, OF EACH SPECIFIED FOOD

	FOODS AND AMOUNTS	AVERAGE USE LAST YEAR								
		NEVER, OR LESS THAN ONCE A MONTH	1–3 PER MO.	1 PER WEEK	2–4 PER WEEK	5–6 PER WEEK	1 PER DAY	2–3 PER DAY	4–5 PER DAY	6+ PER DAY
	YOGHURT (1 CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	COTTAGE OR RICOTTA CHEESE ($^1/_2$ CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	CREAM CHEESE (1oz)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	OTHER CHEESE, eg AMERICAN, CHEDDAR, etc, PLAIN OR AS A PART OF A DISH (1 SLICE OR 1 oz SERVING)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	MARGARINE (PAT), ADDED TO FOOD OR BREAD; EXCLUDE USE IN COOKING	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	BUTTER (PAT), ADDED TO FOOD OR BREAD; EXCLUDE USE IN COOKING	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
FRUITS	RAISINS (1oz OR SMALL PACK) OR GRAPES	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
PLEASE TRY TO AVERAGE YOUR SEASONAL USE OF FOODS OVER THE ENTIRE YEAR. FOR EXAMPLE, IF A FOOD SUCH AS CANTALOUPE IS EATEN 4 TIMES A WEEK DURING THE APPROXIMATE 3 MONTHS THAT IT IS IN SEASON, THEN THE **AVERAGE** USE WOULD BE ONCE PER WEEK.	PRUNES ($^1/_2$ CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	BANANAS (1)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	PEACHES, APRICOTS OR PLUMS (1 FRESH, OR $^1/_2$ CUP CANNED)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
		NEVER	1–3/M	1/W	2–4/W	5–6/W	1/D	2–3/D	4–5/D	6+/D
	FRESH APPLES OR PEARS (1)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	APPLE JUICE OR CIDER (SMALL GLASS)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	ORANGES (1)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	ORANGE JUICE (SMALL GLASS)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	GRAPEFRUIT ($^1/_2$)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	GRAPEFRUIT JUICE (SMALL GLASS)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	OTHER FRUIT JUICES (SMALL GLASS)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	STRAWBERRIES, FRESH, FROZEN OR CANNED ($^1/_2$ CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	BLUEBERRIES, FRESH, FROZEN OR CANNED ($^1/_2$ CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	CANTALOUPE ($^1/_4$ MELON)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	WATERMELON (1 SLICE)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
VEGETABLES	BROCCOLI ($^1/_2$ CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	CABBAGE OR COLESLAW ($^1/_2$ CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	CAULIFLOWER ($^1/_2$ CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	BRUSSELS SPROUTS ($^1/_2$ CUP)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯
	CARROTS (1 WHOLE OR $^1/_2$ CUP COOKED)	◯	◯	Ⓦ	◯	◯	Ⓓ	◯	◯	◯

Example of a page from a food frequency questionnaire

- **Food diary or record—**"Please write down (and describe) **everything** you eat and drink (and estimate the amount) for the next 3 (4 or 7) days." Amounts are usually recorded in household measures but for more accuracy subjects can be provided with quick reading scales to weigh food before it goes on the plate (and any leftovers).
- **Food frequency questionnaire—**"Do you eat meat/fish/bread/milk … on average: more than once a day/2 or 3 times a week/once a week/once a month etc?" (see example above).

The self-administered food frequency questionnaire is the newest of the four methods. It is much more economical of investigator's time and suitable for computer analysis. This method has made possible cohort studies with thousands of subjects' food intakes related to disease outcome.[12] It asks for usual, typical intake.

Estimating food intake quantitatively is labour intensive (and so expensive). It depends on adequate memory and the honesty and interest of the subject. Some people's food habits are very irregular and most people eat differently on Saturday and Sunday from during weekdays and on holiday.

There is no "gold standard" method for food intake. Methods that ask subjects to record or remember the food/drinks they really consumed are better for national nutrition surveys of samples of the population. Items may be forgotten (or not listed because of embarrassment) but the exact types of food, product or dish are actually captured, including unusual ones. For large epidemiological surveys food frequency questionnaires have enabled tens of thousands to be included. What these lose in precision they gain in numbers of subjects. Usually quintiles (fifths) of intake of food components

Two pages from The *Composition of Foods.*[14] Each set of foods has four pages of nutrients. Pages 3 and 4 (not shown here) report vitamins and minerals

Cereals and cereal products *continued*

				Composition of food per 100 g						
No.	**Food**	**Description and main data sources**	**Edible proportion**	**Water g**	**Protein g**	**Fat g**	**Carbo-hydrate g**	**Energy value kcal**	**kJ**
Rice									
18	Brown rice, *raw*	5 assorted samples	1.00	13.9	6.7	2.8	81.3	357	1518
19	*boiled*	Water content weighed, other nutrients calculated from raw	1.00	66.0	2.6	1.1	32.1	141	597
20	Savoury rice, *raw*	10 samples, 5 varieties, meat and vegetable	1.00	7.0	8.4	10.3	77.4	415	1755
21	*cooked*	Calculation from raw, boiled in water	1.00	68.7	2.9	3.5[a]	26.3	142	599
22	White rice, easy cook,	10 samples, 9 different brands, parboiled	1.00	11.4	7.3	3.6	85.8	383	1630
23	*raw* easy cook, *boiled*	Calculation from raw	1.00	68.0	2.6	1.3	30.9	138	587
24	fried in lard/dripping	Recipe	1.00	70.3	2.2	3.2	25.0	131	554
Pasta									
25	Macaroni, *raw*	10 samples, 7 brands; literature sources	1.00	9.7	12.0	1.8	75.8	348	1483
26	*boiled*	10 samples, 7 brands boiled in water	1.00	78.1	3.0	0.5	18.5	86	365
27	Noodles, egg, *raw*	10 samples, 8 brands	1.00	9.1	12.1	8.2	71.7	391	1656
28	egg, *boiled*	10 samples, 8 brands boiled in water	1.00	84.3	2.2	0.5	13.0	62	264
29	Spaghetti, white, *raw*	10 samples, 7 brands	1.00	9.8	12.0	1.8	74.1	342	1456
30	white, *boiled*	10 samples, 7 brands boiled in water	1.00	73.8	3.6	0.7	22.2	104	442
31	wholemeal, *raw*	10 samples, 5 brands	1.00	10.5	13.4	2.5	66.2	324	1379
32	wholemeal, *boiled*	Water content weighed, other nutrients calculated from raw	1.00	69.1	4.7	0.9	23.2	113	485

			Fatty acids						Dietary fibre	
		Total nitrogen g	**Satd g**	**Mono unsatd g**	**Poly Unsatd g**	**Cholesterol mg**	**- Starch g**	**Total sugars g**	**Southgate method g**	**Englyst method g**
Rice										
18	Brown rice, *raw*	1.10	0.7	0.7	1.0	0	80.0	1.3	3.8	1.9
19	*boiled*	0.43	0.3	0.3	0.4	0	31.6	0.5	1.5	0.8
20	Savoury rice, *raw*	1.41	3.2	3.7	1.8	1	73.8	3.6	4.0	N
21	*cooked*	0.48	1.1	1.3	0.6	Tr	25.1	1.2	1.3	1.4
22	White rice, easy cook,	1.23	0.9	0.9	1.3	0	85.8	Tr	2.7	0.4
23	*raw* easy cook, *boiled*	0.44	0.3	0.3	0.5	0	30.9	Tr	1.0	0.1
24	fried in lard/dripping	0.37	1.4	1.2	0.5	3	23.1	1.9	1.2	0.6
Pasta										
25	Macaroni, *raw*	2.11	0.3	0.1	0.8	0	73.6	2.2	5.0	3.1[b]
26	*boiled*	0.52	0.1	Tr	0.2	0	18.2	0.3	1.5	0.9[b]
27	Noodles, egg, *raw*	2.12	2.3	3.5	0.9	30	69.8	1.9	5.0	(2.9)
28	egg, *boiled*	0.40	0.1	0.2	0.1	6	12.8	0.2	1.0	(0.6)
29	Spaghetti, white, *raw*	2.11	0.2	0.2	0.8	0	70.8	3.3	5.1	2.9
30	white, *boiled*	0.63	0.1	0.1	0.3	0	21.7	0.5	1.8	1.2
31	wholemeal, *raw*	2.30	0.4	0.3	1.1	0	62.5	3.7	11.5	8.4
32	wholemeal, *boiled*	0.81	0.1	0.1	0.4	0	21.9	1.3	4.0	3.5

[a] Calculated assuming water only was added; savoury rice cooked with fat contains approximately 8.8 g fat per 100 g
[b] Wholemeal macaroni contains 8.3 g (raw) and 2.8 g (boiled) Englyst fibre per 100 g. Reproduced by permission of the Royal Society of Chemistry

are related to disease outcomes. Relative consumption across the population is the aim here, rather than precise and specific quantitative results (examples are in chapter 1).

Biomarkers of dietary intake

All estimates of food intake are subjective. Some subjects forget, some did not notice what food they ate, some do not report because they would be ashamed to admit having that food, drink, or amount. The search is on for biomarkers, which are objective biochemical indices of dietary intake. For some food components there are useful biomarkers; for others there is nothing available. Protein intake is reflected by 24-hour urinary nitrogen. Intake of non-endogenous fatty acids (18:2, 18:3, 20:4 and *trans* unsaturated) are reflected well in serum or adipose tissue fatty acid pattern but there is no biomarker for total fat or carbohydrate intake. Intake of individual carotenoids can be seen in the plasma. Plasma lycopene, for example, reflects tomato intake. Of the inorganic nutrients 24-hour urinary sodium is a much better indicator of salt intake than dietary history because the salt content of foods and dishes varies greatly and without the subject's knowledge. Vitamin A and calcium are nutrients whose plasma concentrations cannot serve here; they are held constant across the range of usual intakes. Toenail selenium has been used to reflect selenium intake.

Food tables and nutrient databases

The British food tables are among the best in the world thanks to the original work of McCance and Widdowson,[13] their continuation by Paul and Southgate and colleagues, and the support of the Medical Research Council; Ministry of Agriculture, Fisheries, and Food; the Agriculture and Food Research Council; and the Royal Society of Chemistry. The 5th edition of *McCance & Widdowson's the composition of foods* was published in 1991,[14] and there is a growing list of supplements.

In the main volume 1188 foods and drinks arranged in 14 groups are given code numbers (which can be used for computer input—software packages are also available). The total publications, including supplements, cover around 2000 foods. Figures are given (per 100 g edible portion) for over 40 constituents.

The two example pages from the *Composition of foods* show the proximate constituents or macronutrients printed in the main tables.[14] Where relevant, alcohol content is also included. The other two pages for each food show inorganic nutrients—Na, K, Ca, Mg, P, Fe, Cu, Zn, Cl, Mn, Se, I—and all the vitamins—retinol, carotene, vitamins D and E, thiamin, riboflavin, niacin, tryptophan/60, vitamins B-6 and B-12, folate, pantothenate, biotin, vitamin C.

Individual amino acids and fatty acids are in the (1978) 4th edition of *McCance & Widdowson*[14] and a 1980 supplement. For any large scale work the nutrient data bank behind *McCance & Widdowson* is used in computer software, from the Royal Society of Chemistry, Letchworth, Herts., SG6 1HN or MAFF or a specialist software company. This software should contain the latest updates, revisions and additions to the database.

Variability of nutrient content

The British tables give a single value for each nutrient in each food—an estimated average. The American tables,[16] which are published in loose leaf sections as analyses are completed, give an indication of variability to be expected round each average figure. There is no legal guarantee that any food contains what the food tables say it should. For some constituents—for example, fat—the food could easily contain 25% less or 25% more.

Food composition tables[15]

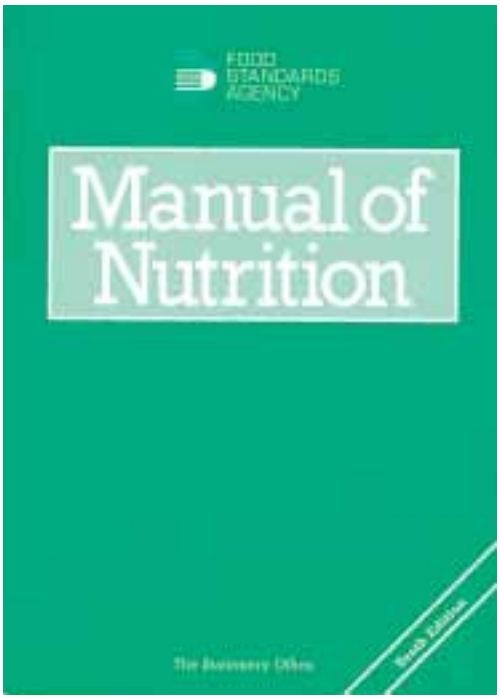
Food composition analysis is work that is never finished. It is a bottomless pit. No country has food composition tables which have kept up with all the concepts of modern nutritional science, the proliferation of foods and dishes eaten in a country and the ever-growing list of substances in foods that may or could have biological functions.

The numbers in existing food tables are derived in three ways:

- By direct analysis in the country where the food tables are produced. These may be based on multiple samples and a modern, specific analytical method or they may be antique figures based on a single sample.
- From literature. Figures may be a subjective decision (even if experienced) of the best representative one from an array of original results, or they may be "borrowed" from another country's food tables.
- Imputed figures, estimates from similar foods or from the known composition of most of the ingredients of mixed dishes.

As a rule food tables do not say where each number comes from.

Food components/nutrients are not the same if analysed by different methods. Carbohydrate can be obtained by subtracting per cent water, fat, protein and ash from 100—"by difference" or by direct chemical analysis adding up the different classes of carbohydrate. There is still no agreement on which of some four different methods should be the standard one for dietary fibre. Cholesterol may be obtained by an older colorimetric method or by more specific gas-liquid chromatography.



The *Manual of Nutrition* has a useful simplified food table, which gives calories and 11 nutrients for 150 common foods[17]

Dietary reference values and requirements

For nutrient intakes to have any meaning they have to be compared against some number representing physiological requirements for each nutrient. The major number for this is the estimated upper end of the range of individual requirements. In Britain, the name for this was changed in 1991 to *reference nutrient intake* (RNI)[7] from recommended daily amount (RDA). The latter had the same abbreviation as the American term, recommended dietary allowance.

The North American dietary reference intakes are being revised in sections.[18] These new reference values have similar terminology to the British and also have numbers for a **tolerable upper intake level** (above which adverse effects might occur).

Ten of the new North American dietary reference intakes (DRIs) are similar to the 1991 UK set but five are different. A DRI has been added for selenium on which there has been interesting research since 1991 (see chapter 10) and daily requirements for vitamin D are provided. The folate requirement has been increased because prevention of neural tube defects and hyperhomocystinaemia have emerged in the 1990s. The North American committee set higher recommendations for vitamin C by using an optimal blood level of ascorbate rather than prevention of scurvy and maintenance of measurable ascorbate in plasma. The calcium recommendation is also higher, based more on balance data than the British recommendations, based more on epidemiology.

The RNI/RDA values differ for the two sexes and several age groups and are published in tables. They are for intakes averaged over several days. Since the RNI/RDA is estimated to meet the requirements of practically all healthy people, most people's requirements are less than this upper or prescriptive reference value. For assessment of people's intakes a lower, diagnostic, reference should be used, such as the **estimated average requirement** for groups of people and the **lower reference nutrient intake** for individuals.[7] An intake below this does not necessarily mean inadequacy but the lower it is the greater that probability. As elsewhere in medicine, for complete nutritional diagnosis, findings on examination (clinical, anthropometric, and biochemical) must be considered together with the food intake history.

Sick people have requirements not covered by the reference values. For example with bed rest, energy requirements are reduced; with fever they are increased. Losses of several nutrients are increased in different ways by illness, such as protein loss in nephrotic syndrome, potassium loss in diarrhoea, iron loss with bleeding. The allowances apply to oral feeding of conventional foods. For total parenteral nutrition the requirements are different; absorption is 100%, but minor vitamins like pantothenate and biotin and trace elements like molybdenum, manganese, chromium, etc, which can usually be ignored because there is enough in the diet, cannot be taken for granted and have to be provided in the infusion solution(s).

Biochemical methods for nutritional status

Biochemical tests are an integral part of modern medical diagnosis. Plasma sodium and potassium concentrations, for example, are essential for diagnosing and treating difficult electrolyte disorders, and plasma or red cell folate and plasma vitamin B-12 should be measured before treating a patient with megaloblastic anaemia.

With most of the other nutrients also, biochemical tests have been developed which can be used (a) to confirm the diagnosis of a deficiency disease in places where it is uncommonly seen or where the clinical picture is complicated,

British dietary reference values[7]: estimated average requirement for food energy (calories) and reference nutrient intakes (RNIs) for the other nutrients

Nutrient	Men 19-50 years	Infants 4-6 months
Energy, kcal (MJ)	2550 (10.6)	670 (2.8)
Protein (g)	56	13
Vitamin A (RE, μg)	700	350
Thiamin (mg)	1.0	0.2
Riboflavin (mg)	1.3	0.4
Niacin (NE, mg)	17	3
Vitamin B−6 (mg)	1.4	0.2
Folate (total, μg)	200	50
Vitamin B−12 (μg)	1.5	0.3
Vitamin C (mg)	40	25
Vitamin D (μg)	—	8.5
Vitamin E (mg)	7	0.4/g PUF
Calcium (mg)	700	525
Iron (mg)	8.7	4.3
Magnesium (mg)	300	60
Iodine (μg)	140	60
Potassium (mmol)	90	22
Sodium (mmol)	70	12
Zinc (mg)	9.5	4

Some figures are slightly rounded. Figures for women are generally about 75% of those for men, except iron (higher until the menopause) and in pregnancy and lactation. Figures for children are interpolated between infants and adults.
RE = retinol equivalents; NE = niacin equivalents; PUF = polyunsaturated fat

Biochemical methods for diagnosing nutritional deficiencies

Nutrient	Indicating reduced intake	Indicating impaired function (IF) or cell depletion (CD)	Supplementary method
Protein	Urinary nitrogen	Plasma albumin (IF)	Fasting plasma amino acid pattern
Vitamin A	Plasma β-carotene	Plasma retinol	Relative dose response
Thiamin	Urinary thiamin	Red cell transketolase and TPP effect (IF)	
Riboflavin	Urinary riboflavin	Red cell glutathione reductase and FAD effect (IF)	
Niacin	Urinary *N′* methyl nicotinamide or 2-pyridone, or both	Red cell NAD/NADP ratio	Fasting plasma tryptophan
Vitamin B-6	Urinary 4-pyridoxic acid	Plasma pyridoxal 5′ phosphate	Urinary xanthurenic acid after tryptophan load
Folate	Plasma folate	Red cell folate (CD)	Urinary FIGLU after histidine load
Vitamin B-12	Plasma holo-transcobalamin II	Plasma vitamin B-12	Schilling test
Vitamin C	Plasma ascorbate	Leucocyte ascorbate (CD)	Urinary ascorbate
Vitamin D	Plasma 25-hydroxy-vitamin D	Raised plasma alkaline phosphatase (bone isoenzyme) (IF)	Plasma 1,25 dihydroxy-vitamin D
Vitamin E	Ratio of plasma tocopherol to cholesterol+ triglyceride	Red cell haemolysis with H_2O_2 *in vitro* (IF)	
Vitamin K	Plasma phylloquinone	Plasma prothrombin (IF)	Plasma des-γ-carboxy-prothrombin
Sodium	Urinary sodium	Plasma sodium	
Potassium	Urinary potassium	Plasma potassium	Total body potassium by counting ^{40}K
Iron	Plasma iron and transferrin	Plasma ferritin (CD)	Free erythrocyte protoporphyrin
Magnesium	Plasma magnesium	Red cell magnesium (CD)	
Iodine	Urinary (stable) iodine	Plasma thyroxine (IF)	Plasma TSH
Zinc	Plasma zinc	Red cell zinc	

TPP = thiamine pyrophosphate; FAD = flavin adenine dinucleotide;
NAD = nicotinamide-adenine-dinucleotide; NADP = NAD phosphate;
FIGLU = formiminoglutamic acid; ^{40}K = natural radioactive potassium;
TSH = thyroid stimulating hormone
There are no reliable simple methods for assessing *calcium* status (total body calcium can be measured by *in vivo* neutron activation analysis)

or (b) in community surveys and general practice to find individuals with subclinical nutrient deficiencies.

Tests which can be used for the major nutrients are listed in the table on page 85. As with other tests in chemical pathology, there may be false positive and false negative results. For example plasma vitamin B-12 concentrations are increased in acute hepatitis, and alkaline phosphatase may not be raised if rickets is accompanied by protein-energy deficiency.

When the intake of a nutrient is inadequate (less than obligatory losses) an individual generally goes through **three stages**. The first is adaptation to the low intake: urinary excretion of the nutrient or its metabolites typically falls but there is no evidence of abnormal function or of depletion of the cells.

In the second stage there are also biochemical changes indicating either impaired function or cellular depletion, but clinical manifestations of deficiency are absent or non-specific. A good example of a test showing impaired function is red cell transketolase activity. For each blood sample this enzyme is assayed in two test tubes, one with extra thiamin pyrophosphate (TPP), the other without. If the activity is more than 25% higher in the supplemented tube (TPP effect + 25%) this indicates functional thiamin deficiency. The third stage of depletion is that of clinical deficiency disease.

Most clinical biochemistry laboratories provide only some of the methods in the table as a routine but others could be set up in special circumstances or, alternatively, a laboratory specialising in nutrition research could be asked to help.

References

1 de Boer JO, van Es AJH, van Raaij JMA *et al.* Energy requirements and energy expenditure of lean and overweight women, measured by indirect calorimetry. *Am J Clin Nutr* 1987; **46**: 13-21.
2 Murgatroyd PR, Shetty PS, Prentice AM. Techniques for the measurement of human energy expenditure: a practical guide. *Int J Obesity* 1993; **17**: 549-68.
3 Edholm OG, Adam JM, Best TW. Day to day weight changes in young men. *Ann Hum Biol* 1974; **1**: 3-12.
4 McKay K, Fozdan-Faroudi S, Bowman CE. How heavy the patient? *BMJ* 1991; **303**: 1608.
5 Frisancho AR. New norms of upper limb fat and muscle areas for assessment of nutritional status. *Am J Clin Nutr* 1981; **34**: 2540-5.
6 Bishop CW, Bowen PE, Ritchey SJ. Norms for nutritional assessment of American adults by upper arm anthropometry. *Am J Clin Nutr* 1981; **34**: 2530-9.
7 Department of Health. *Dietary reference values for food energy and nutrients for the United Kingdom. (Report on Health & Social Subjects no 41.)* London: HMSO, 1991.
8 Scrimshaw NS, Waterlow JC, Schürch B. Energy and protein requirements. Proceedings of an IDECG workshop. *Eur J Clin Nutr* 1996; **50** (Suppl 1): S11-S81.
9 Allman MA, Allen BJ, Stewart PM *et al.* Body protein of patients undergoing haemodialysis. *Eur J Clin Nutr* 1990; **44**: 123-31.
10 Durnin JVGA, Womersley J. Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. *Br J Nutr* 1974; **32**: 77-97.
11 Slater JM (ed) (Ministry of Agriculture, Fisheries & Food). *Fifty years of the National Food Survey 1940-1990.* London: HMSO, 1991.
12 Willett W. *Nutritional epidemiology* 2nd edn. New York: Oxford University Press, 1998.
13 Ashwell M. *McCance and Widdowson, A Scientific Partnership of 60 Years.* London: British Nutrition Foundation.
14 Holland B, Welch AA, Unwin ID, Buss DA, Paul AA, Southgate DAT. *McCance & Widdowson's the composition of foods*, 5th edn. Cambridge: Royal Society of Chemistry, 1991.
15 Greenfield H, Southgate DAT. *Food composition data. Production, management and use.* London: Elsevier Applied Science, 1992.
16 Human Nutrition Information Service, US Department of Agriculture. *Composition of foods: raw, processed, prepared.* Washington DC: US Government Printing Office. (This is a growing set of 23 volumes for different food groups, each with looseleaf supplements, providing nutrient analyses for over 3600 foods. Handbook 8-1 is for dairy and egg products, 8-2 for spices and herbs, 8-3 for baby foods … 8-22 for mixed dishes and 8-23 for miscellaneous foods.)
17 Ministry of Agriculture, Fisheries & Food. *Manual of nutrition*, 10th edn. London: HMSO, 1995.
18 The most recent is: Institute of Medicine. *Dietary Reference Intakes for vitamin A, vitamin K, arsenic, boron, chromium, copper, iodine, iron, manganese, molybdenum, nickel, silicon, vanadium and zinc.* Washington DC: National Academy Press, 2002.

13 Therapeutic diets

A person may be advised to change his or her diet to help treat or prevent disease.

- For essential or lifesaving treatment—for example, in coeliac disease, phenylketonuria, galactosaemia, hepatic encephalopathy.
- To replenish patients who are malnourished because of diseases such as cancer, intestinal diseases, and anorexia nervosa.
- To produce a negative energy balance in obese people.
- As helpful treatment, alternative or complementary to drugs, as in diabetes mellitus, mild hypertension, dyspepsia.
- To deal with the side effects of some drugs—for example, diets with increased potassium for patients taking long term diuretics or diets with restricted tyramine for patients taking monoamine oxidase inhibiting antidepressants.
- To provide standard conditions for diagnostic tests—for example, for measuring faecal fat. Also, an elimination diet is the mainstay in diagnosis of food sensitivity.
- Prophylactic diets like those described in chapter 7 on nutrition for adults (dietary goals or guidelines for the general public) often combine mild restriction of energy, saturated fat, and sodium with a moderate increase of dietary fibre.

Therapeutic diets ask patients to make one or more of the following changes: reduce or (virtually) eliminate one or more components, increase one or more food components, change the consistency of the diet, or change the feeding pattern. These are all changes to the patient's usual diet (which, of course, varies somewhat from day to day) or in comparison with a hypothetical "normal" or average diet for the country, culture, age, and sex.

The prescription for a diet should state:

- the nature of the modification(s)
- the degree of each modification
- the planned duration of these
- any compensation for essential nutrients compromised by the modifications.

The degree of the modification is as important as the dose in pharmacotherapy. People talk loosely about a "low salt" diet but its sodium intake can range from 25 to 100 mmol/day compared with a normal British sodium intake of around 150 mmol/day. Likewise with protein, a protein restricted diet may vary from 20 to 50 g/day compared with the standard of about 75 g/day (1 g/kg).

The dietary prescription has to be adjusted for the individual patient:

- for the foods disliked and liked
- for any sensitivity or intolerance to food
- for any religious food prohibition (including Ramadan for Moslems, the month when all eating has to be after dark)
- for vegetarians
- to include foods eaten away from home
- for income, occupation, and level of education
- for cooking facilities and the patient's domestic situation
- for the need for variety in foods (some insist on variety; others like the same foods from day to day)
- for the patient's motivation and degree of obsessionality
- for calorie (energy) expenditure and needs

For the purposes of describing therapeutic alternatives to diets a "normal" or average diet provides for a hypothetical healthy 70 kg Western man something like:

- **Energy** 10.5 MJ (2500 kcal)
- **Protein** 14% of energy or 85 g
- **Fat** 35% of energy or about 100 g
- **Carbohydrates** 48% of energy or 300 g
- **Alcohol** 3% of energy or 1 drink/day.

If the percentages of energy for macronutrients omit alcohol (as some do) carbohydrates here go up to 50% of energy and fat to 36%.

A "normal" diet provides around 75% of these absolute figures for the hypothetical woman (not pregnant or lactating), but the same percentages of energy for the macronutrients.

The naming of diets

Diets are sometimes described eponymously (Giovanetti diet) or as belonging to a specific disease (renal failure diet). But neither type of name is recommended. Diets named after their (supposed) originator give no clue about their composition and particular diets do not necessarily relate to specific diseases. A "renal failure diet" may also be used for hepatic encephalopathy or rare inborn errors of the ornithine cycle. There have been many "diabetic" and "renal failure" diets.

"Cholesterol lowering diet" is ambiguous (and "low cholesterol diet" is worse). Several diets may lower the plasma concentration of cholesterol—low fat, vegetarian, or increased seed oil (ω-6 polyunsaturated fatty acids)—and it is not necessary to lower the **dietary** cholesterol.

The most reliable way of naming diets is by the major change (from an average diet) in its composition.

Degree of modification

Sodium

- 100 mmol Na (2.3 g) is a mild low sodium diet
- 50 mmol Na (1.2 g) is a moderate low sodium diet
- 25 mmol Na (0.6 g) is a strict low sodium diet

Protein

- 50 g/day (0.75 g/kg) is a mild protein restricted diet
- 30 g/day (0.5 g/kg) is a moderate protein restricted diet
- 20 g/day (0.33 g/kg) is a severe protein restricted diet

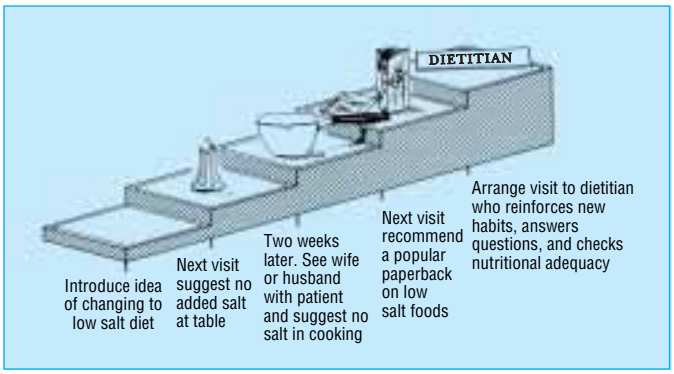
- for the duration of the diet. If a diet is necessary for only a week or two then it is not serious if it provides less than the recommended daily amount of (say) calcium or magnesium, but if the diet is continued these elements must be provided, by supplements if necessary
- for the patient's prognosis. A strict diet may not be justifiable for someone with a short life expectancy
- when two or more dietary prescriptions are combined. Sometimes these are more or less incompatible—for example, a low calorie plus high potassium diet or a high calcium plus lactose free diet (supplements would have to be used for these).

Strategy

Essential, or lifesaving, diets should be looked after in collaboration with a dietitian. For some diseases—for example, gout, mild hypertension, and hyperlipidaemia—drugs or diet are alternative or complementary treatment options. Drugs appear to act more quickly, are easier to administer and more reliable, and take less of the doctor's time, but they may cause more side effects. Diet appears more natural and safer but it will take longer to explain. Sometimes the best choice is a synergistic combination so the dose of drug can be low (hence fewer side effects) and the diet not too irksome.

We know from results with obese people who are on weight reducing diets, and from studies in diabetics, that most people do not follow the diet prescribed. It is difficult and time consuming to explain what is intended and how it may be done. It is difficult too for a patient to make major changes to his or her food habits. Minor changes are much easier to incorporate and some places in the day's food sequence are easier to change than others. Each family and each individual has different feelings and ideas about foods. Some foods are given up more readily than others.

Outside hospital a therapeutic diet ("I'm on a low salt diet") is a strange association of an occasional talk by the doctor or dietitian with daily action by the patient and his or her family in the supermarket, kitchen, dining room, works canteen, and pub. The reality is different from what is on paper.



Building up therapeutic diet at successive visits

Techniques

Essentially the doctor or dietitian has a list of foods rich in the component to be changed and of foods with medium and

Three at the table

'Salt'—Food Standards Agency leaflet

low amounts of it. The trouble with scientific food tables, such as *McCance & Widdowson's The Composition of Foods*,[1] is that they give the content of nutrients per 100 g whereas what matters is the content per usual serving or portion (see box: typical serving sizes).

The patient, with his or her spouse, can produce a list of what the family usually eats and how it is cooked. Ideally the next series of steps is for the doctor or dietitian and these two to work out the most comfortable way for them to incorporate the dietary prescription into the family's food patterns. This cannot be completed at one session. It requires trial and error, questions and compromises.

There are two fairly easy ways of changing the diet. First, a food that tastes and functions like the original but has a different composition may be substituted. Examples are: polyunsaturated margarine for butter; sunflower (etc) oil for dripping; skimmed or 2% fat milk for whole milk; salt free bread for ordinary bread; wholemeal bread for white bread; high fibre breakfast cereal for low fibre brand.

Second, a simple addition may be made to less important parts of the day's diet. Examples are: a sprinkling of bran on the breakfast cereal to increase fibre; casein powder (such as Casilan) sprinkled on to food three times a day to increase protein; spoonful(s) of fish oil (such as Maxepa) to increase long chain highly polyunsaturated fatty acids.

Diets are more likely to be followed and persisted with by patients who are well motivated, have stable mood, normal intelligence, good home support, and lead a well organised life. Indeed, in some obsessional patients there can be the opposite problem of overdoing a diet suggested long ago on thin scientific evidence or for a condition that has since disappeared.

Checking compliance and effectiveness

From an authoritarian viewpoint patients often do not properly **comply** with the doctor's instructions. This can be checked by asking revealing questions, by calling into the home at meal times, or by objective tests (for example biomarkers).

But for an intelligent patient who thinks that dieting is his or her own responsibility and that of his or her partner, with the doctor or dietitian one of their sources of information, what needs to be checked is the **effectiveness** of what they are doing.

Whatever the viewpoint or words used the same objective tests are available:

- change of body weight for reduced or increased energy diets
- increase of faecal weight for high (wheat) fibre diets
- 24-hour urinary sodium and potassium for dietary changes of sodium or potassium
- 24-hour urinary nitrogen for high or low protein diets, and also as general check on food intake (on average protein intake is 10-15% of energy intake)
- plasma fasting triglyceride fatty acid pattern to indicate consumption of polyunsaturated fat
- blood urea, urate, glucose, cholesterol, haemoglobin for respective diets prescribed to moderate these.

A Royal Pharmaceutical Society report urges clinicians to take a more egalitarian view of the relationship between prescribing and medicine taking, between patient and prescriber.[3] This is even more needed for dietary management.

Typical serving sizes, approximate weight in grams

Bread (1 slice) 30 g
Crispbread (1 slice) 10 g
Biscuits about 12 g each
Breakfast cereal 30 g
Butter or margarine (for 1 slice bread) 7 g
Oil (1 tablespoonful) 20 g
Cake (portion) 40-50 g
Jam (for 1 slice bread) 15 g
Marmite (for 1 slice bread) 2 g
Milk (for 1 cup tea) 30 g
Milk (6 oz glass) 200 g
Cream (1 tablespoonful) 20 g
Sugar (1 level teaspoon) 5 g
Yoghurt (1 carton) 125-150 g
Cheese (1 portion) 30 g
Egg (1 edible portion) 50 g
Meat (chicken or beef) little or no bone 90-120 g
Meat (with bone, for example, chop) 160-200 g
Bacon (1 strip, raw) 30-40 g
Liver 80 g
Sausage (one) 50 g (approx)
Fish (fresh and canned) 110-120 g
1 fish finger 28 g
Macaroni and other pasta (for main course) 100 g (before cooking)
Vegetables (for example, peas, cauliflower) 60-100 g (before cooking)
Potato (1 medium, raw) 90 g
Lettuce 30 g
Parsley (chopped) 3 g
Fruit (1 apple, 1 banana peeled, raw) 100 g (approx)
1 grape 5 g
Nuts 30 g
Pepper 0.2 g
Wine (glass) 110-125 g
Spirits 25 g
Beer (1/2 pint) 285 g
Carbonated soft drink 240-330 g
Coffee powder 2 g

For more detail on individual British foods, see Crawley's handbook.[2]

Diets for treating diabetes

Diabetic diets changed greatly throughout the 20th century. They have undergone further change since about 1970, as several facts emerged:

(1) Oral hypoglycaemic drugs may predispose to heart disease.
(2) There is no evidence that eating sugar **causes** diabetes.

History of diets for diabetics 1900-80

1900-25	Fasting (Naunyn, Allen); 5% *carbohydrate*, 85% fat (Newburgh 1923)
1922	Insulin discovered (but not generally available for a few years)
1930	15% *carbohydrate* and 70% fat
1940	All *carbohydrate low*, for example, 40% and fat 50%
1950	Lawrence's lines in UK
1970	*Carbohydrate round* 40%; sugar prohibited Emphasis on oral drugs or insulin rather than diet
1980	*Carbohydrates* 50% or more; emphasise "complex" CH_2O, fibre and legumes; restrict sugar
2000	Emphasis on lower glycaemic index foods

(3) Asian people with diabetes on high starch diets have fewer complications (especially atherosclerotic) than their counterparts in Western Europe and North America.
(4) Westerners with diabetes are dying of excess atherosclerotic disease, have higher plasma cholesterol values, and have been eating higher saturated fat diets than people with no diabetes.
(5) Viscous dietary fibres such as that in guar, pectin, and legumes (though carbohydrates) improve diabetic control.
(6) Increased dietary carbohydrate improves the response to a glucose tolerance test. Increasing the (complex) carbohydrate of diabetic diets is not usually followed by deteriorating control.
(7) Individual foods containing carbohydrate do not give the same glucose and other metabolic responses at a standard intake. When put to the test, in human subjects some foods give much higher blood glucose curves than others. They have a higher or lower **glycaemic index** (area under the 2-hour blood glucose curve after eating a food containing 50 g carbohydrate as percentage of the corresponding area after the same weight of glucose). This means that carbohydrate exchange lists can no longer be relied on. (It was always hard to believe that 2 oz of grapes had the same effect in the body as 7 oz of whole milk.) The main cause of a low glycaemic index is that the starch in some foods is digested slowly by pancreatic amylase.
(8) Diabetics also have an increased chance of developing hypertension. The sodium content of their diets has been largely ignored.

Principles of dietary treatment for diabetes[6,7]

Type 1: insulin dependent diabetes (IDD)

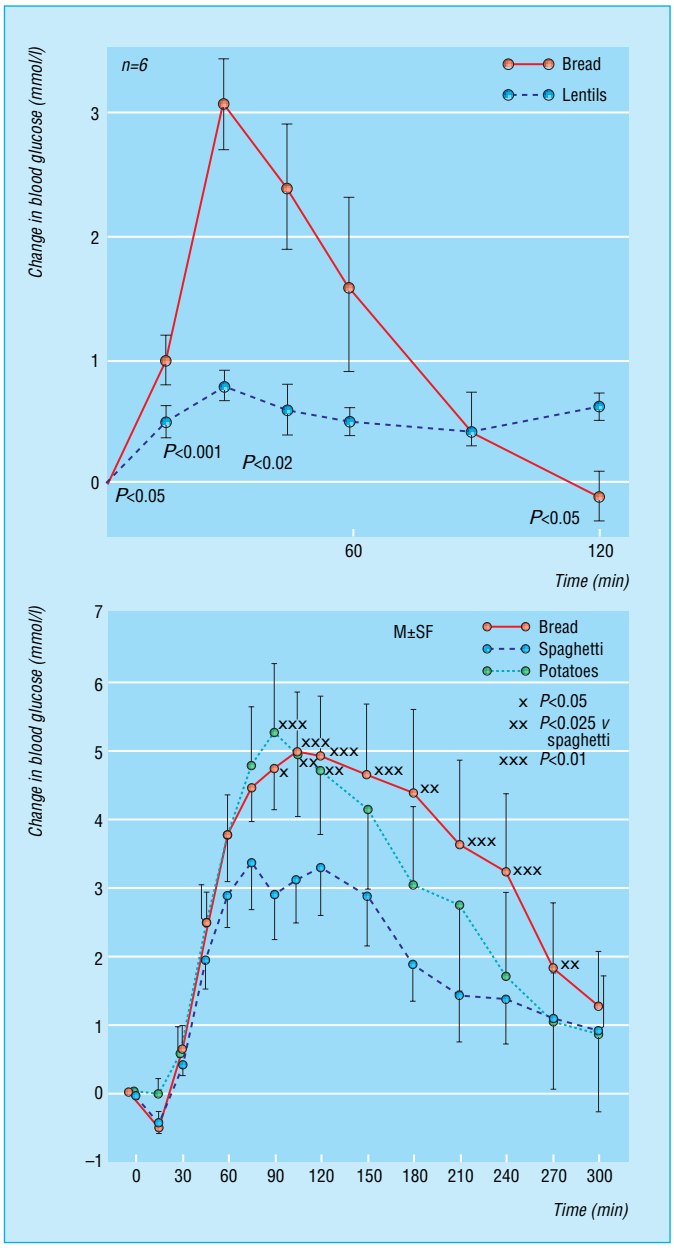
(1) **Integrate and synchronise meals (that is, the metabolic load) with the time(s) of action of the insulin treatment** to minimise high peaks of blood glucose as well as episodes of hypoglycaemia. The American Diabetes Association, for example, recommended that the individual's usual food intake is used as a basis for integrating insulin therapy into the eating and exercise patterns. Patients on insulin therapy should eat at consistent times synchronised with the time-action of the insulin preparation used.
(2) **Reduce saturated fat to 10% of total energy or less**. People with diabetes have an increased risk of coronary heart disease and this dietary change may reduce it.
(3) **Keep salt intake low**, because people with diabetes have an increased risk of hypertension.
(4) **Be very moderate with alcohol**. Large intakes carry the risk of hypoglycaemia; irregular drinking can disturb glycaemic control. But regular 1 to 2 glasses with a meal are acceptable and might be beneficial (except in pregnancy).
(5) **If still growing** make sure intakes of essential nutrients are adequate.

Type 2: non-insulin dependent diabetes (NIDD)

Dietary change has a greater potential to improve type 2 diabetes.

(1) **Reduce body weight by eating fewer kilojoules and taking regular exercise, and keep at it**! Even modest losses of weight improve metabolic control.[8] About three-quarters of type 2 diabetics are overweight or obese, and weight reduction is the first line of dietary management. To help patients lose weight and keep it off is a challenge for the physician and dietitian (cf chapter 11). Diabetics have a stronger incentive to lose weight because this improves their disease as well as their figure, but sulphonylureas or insulin (not metformin) tend to stimulate appetite. Some

Measurement of glycaemic index



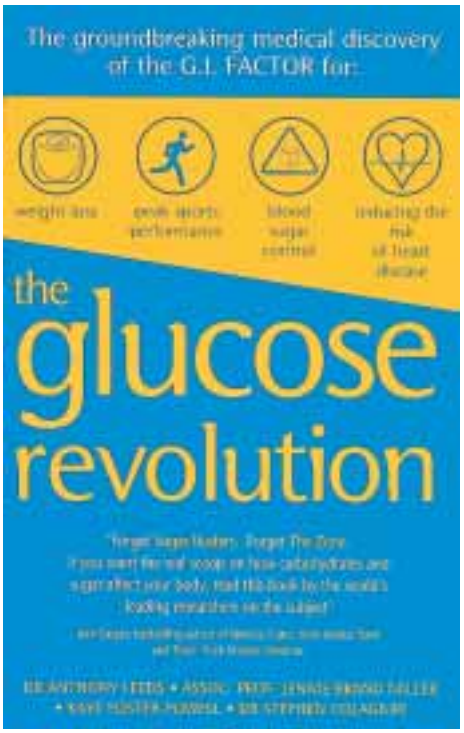
Mean rise of blood glucose after 50 g carbohydrate from four different foods. In the top graph the six subjects were not diabetic: lentils gave very low rise of blood glucose but at 2 hours lentil levels were higher than bread (now below fasting level).[4] In the lower graph the seven subjects had type 2 diabetes (two on insulin, two on glibenclamide). Blood glucose rises were higher than in the subjects with no diabetes and more sustained but over five hours area under the blood glucose curve for spaghetti was 60% that of bread[5]

who succeed in losing weight may be able to go off medications or go off insulin.

(2) **Reduce saturated fat**. Increased LDL-cholesterol may be more pathogenic in type 2 diabetes than non-diabetic people.[10]
(3) Emphasise **low glycaemic index** foods.[9]
(4) Increase intake of **vegetables, fruit, legumes, and whole grain cereals** (which increase fibre intake and mostly have low glycaemic indices).
(5) Keep **salt intake low**.
(6) **Avoid excess alcohol** but 1-2 drinks per day with meals are acceptable.
(7) Forget carbohydrate exchanges
(8) There is no need to be obsessional about reducing sucrose. The glycaemic effect of sucrose is about the same as that of most starchy foods.[7]

Complex carbohydrates

The term "complex carbohydrates" has been used since 1977 for carbohydrates other than sugars (designated "simple carbohydrates")—that is, for polysaccharides, or for carbohydrates in whole grains, vegetables, and fruits. Diabetic patients have been advised to eat foods containing complex carbohydrates rather than simple carbohydrates. But the carbohydrates in fruits and many vegetables are predominantly sugars and food starches vary in rate of digestion so that some give higher rises of blood glucose than sucrose. The term has become confusing.[11] It is better to discuss carbohydrate components by using their chemical names or the names of the foods that contain them and in general to choose those with lower glycaemic index.



An up-to-date book written for the interested diabetic patient[9]

Foods that have been shown to have low glycaemic indices (55 or less) compared to glucose = 100[12]

Soya beans (18)
Lentils (29)
Dried peas (31)
Canned baked beans (40)
Frozen peas (boiled) (48)
Other dried legumes, (around 30)

Pearl barley (28)
All pastas (40-45)
Rolled oats (55)
Oat bran (50)
All Bran (40)
Pumpernickel (rye) bread (41)

Bananas (53)
Apples (36)
Apple juice (36)
Grapefruit (25)
Oranges (43)
Orange juice (57)
Plums (24)
Peaches (28)
Milk (full cream or low fat) and yoghurt (25-35)

The glycaemic index of glucose is 100, of fructose 20, of sucrose (half way between) 60, of lactose 45.
High fat foods may also give a low glycaemic index, because of delayed gastric emptying, but are not recommended and are not on this list.

These dietary prescriptions look very similar to the general dietary guidelines in chapter 7. Indeed it can be said that a modern diabetic diet is the dietary guidelines for the general population pursued with more seriousness. At Birmingham General Hospital "Rather than tell the patients that this is a diabetic diet, we emphasise that it is a normal, healthy diet that everyone should be adhering to".[13]

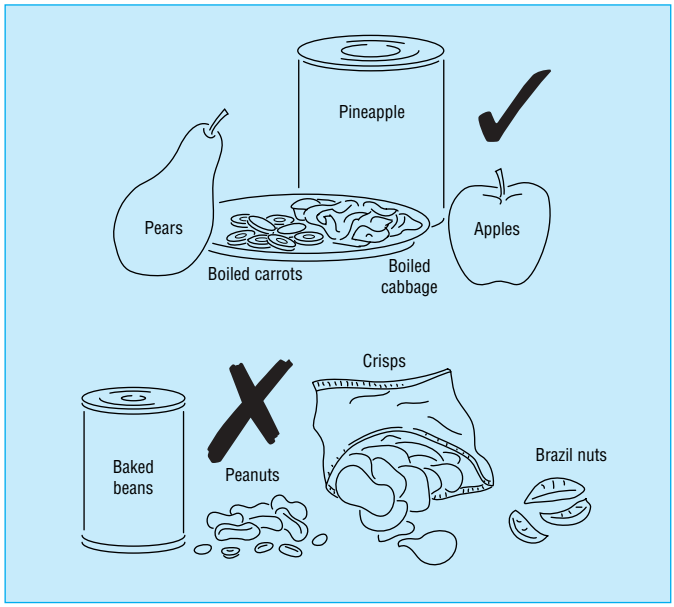
Diets for renal failure

A strict therapeutic diet is needed for patients with renal failure during the few days between diagnosis and dialysis and for the minority for whom dialysis will not be used. The diet should be low in protein (40 g/day) or very low in high biological value protein (25 g/day) with low potassium and a controlled sodium and water intake.

Most patients with chronic renal failure in Western countries however, are nowadays treated with regular dialysis while awaiting a transplant. For them the outpatient diet is not very different from a normal one. Protein should be about 1.2 g/kg body weight, a little more than the recommended daily intake for healthy people. Rather more protein is lost, and

so is needed, on continuous ambulatory peritoneal dialysis than on haemodialysis. Potassium is carefully monitored but usually needs to be only a little restricted, sometimes not at all. It is controlled by adjusting the concentration in the dialysing fluid; 50 mmol/day is an average amount for the diet. This is achieved by eating fruits which have a low potassium content (apples, pears, and canned fruits) and boiled leafy vegetable and avoiding higher potassium vegetables (legumes), nuts, dried fruits, chocolate, and potato chips and crisps.

Patients can usually take an ordinary amount of sodium (about 110 mmol/day) or need only mild restriction. Fluid intake is restricted to about 1000 ml/day. Supplements of water soluble vitamins should not be given above nutrient requirement dosage.[14] Fat soluble vitamin supplements are not required; they tend to accumulate.



Other conditions

Diets for phenylketonuria

Phenylketonuria is one of the better known examples of an inborn error of amino acid metabolism. It leads to mental retardation and other abnormalities if patients are not started on a low phenylalanine diet in the first few weeks of life. Diagnosis is by routine screening of blood phenylalanine after adequate intakes of milk, on about the seventh day of life. Bottle feeding is essential in infancy and a special low phenylalanine formula has to be used, such as Lofenalac or Minafen. When the child is weaned, the diet has to be very different from that of other children: a combination of low protein foods (controlled by phenylalanine content) and a phenylalanine free mixture of other essential amino acids—for example, Aminogran—with sugars and fats and vitamin and mineral supplements. The relative amounts of the first two are adjusted to maintain plasma phenylalanine neither too high (toxicity) nor too low (inadequate growth). The diet has to be strictly maintained and monitored until the child is about 8 years old, after which it can usually be relaxed. But it is required again in women during pregnancy.

Diets for gluten-sensitive enteropathy

Patients with gluten-sensitive enteropathy, coeliac disease, and dermatitis herpetiformis have to modify their diets to eliminate all wheat gluten, rye and barley gluten, and possibly oats. Fresh milk, fresh meat, fish and eggs, fresh vegetables and fruit, rice and maize, tea, coffee, sugar, wine, and spirits are all safe but many processed foods have wheat flour or gluten added. With many of these foods some brands contain gluten, others do not. The only thing to do is to check ingredients on the label or check brands against an up-to-date copy of Coeliac UK's list of gluten free manufactured products (PO Box 220, High Wycombe, Bucks HP11 2HY, http://www.coeliac.co.uk). Gluten free breads and pasta and other products, even communion wafers, are available, some on prescription. Unlike many other diets, even a small lapse and inclusion of the harmful component can lead to prompt return of symptoms.

Coeliac booklet (produced with permission from Coeliac UK)

Diets for dyspepsia

These diets present a contrast to the two preceding essential diets. Classic diets for peptic ulcer have not been found to accelerate healing in controlled barium meal studies. Modern drug treatment, especially with H_2 receptor antagonists and antibacterial treatment for *Helicobacter pylori* is usually effective. Diet is therefore much less emphasised than before for gastroduodenal diseases. Nevertheless, some foods are known to cause gastric irritation or stimulate acid secretion, including

chilli powder, coffee, tea, peppers, alcohol, and cola beverages. Other foods commonly cause heartburn by lowering the tone of the lower oesophageal sphincter: peppermint, garlic, onion, fatty meals. Frequent, small volume feeds are beneficial in oesophageal reflux. Traditional bland foods such as milk, chicken, mashed potatoes, bananas, apples, and ice cream usually relieve symptoms in patients with dyspepsia, though individuals vary.

Diets for diagnostic tests

For several days before a **glucose tolerance test** patients should be standardised on enough carbohydrate—that is, about 300 g or at least 50% of calories, the amount of carbohydrate in ordinary Western diets. Before a **faecal fat** test for malabsorption, patients should be on a known, controlled, and adequate fat intake, 70 to 100 g/day. Before **urinary screening for hypercalciuria** patients should be on a high normal calcium intake of about 1000 mg/day. Before urine is collected for **5-hydroxyindoleacetic acid (5-HIAA)** measurement dietary sources of it or of serotonin should be excluded: bananas, plantains, tomatoes, plums, avocadoes, pineapples, passion fruit, and walnuts. For urinary **4-hydroxy-3-methoxy-mandelic acid** (VMA) specific laboratory methods are now used, and dietary preparation should be unnecessary. Even urinary **creatinine** is affected (increased) by meat consumption.

Diet for patients taking monoamine oxidase inhibitors

The diet for depressed patients taking these antidepressants is the most striking example of dietary adjustment to prevent side effects from drugs. Monoamine oxidase inhibitors interfere with the normal breakdown of tyramine, dopamine, and other amines that occur naturally in foods in which flavour is enhanced by protein breakdown. Dangerous increases of blood pressure may follow ingestion of cheese, but other foods contain these amines and should be excluded too: wines, bananas, aged game, salami, pickled herrings, broad beans, and yeast and meat extracts, and any stale foods. The drugs which require this dietary modification include tranylcypromine, phenelzine, and isocarboxazid.

In practice the most commonly needed diet is one with reduced energy and the second most needed is a regimen with reduced alcohol intake. Both are prescribed much more often than they are successfully followed. This discrepancy remains a challenge for medical practice and research.

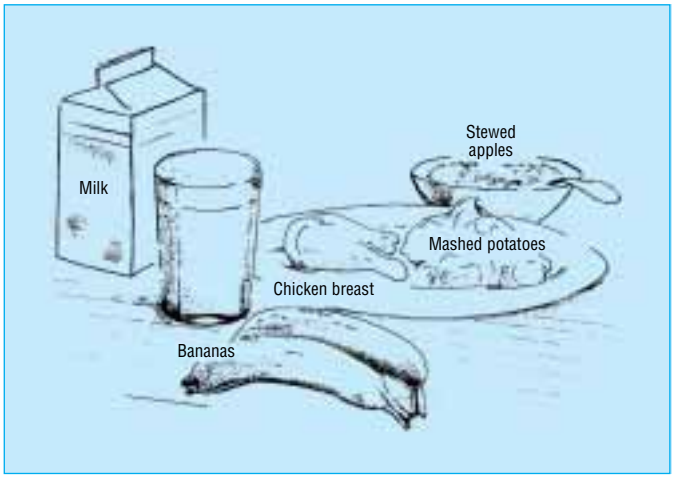
Elsewhere in this book, diets are described which are low in saturated fat, with increased polyunsaturated fat (chapter 1); low in sodium (chapter 2); low in oxalate (chapter 3); low in energy (calories) (chapter 11); and elimination diets (chapter 15).

Further reading

Thomas B (ed). *Manual of dietetic practice*, 3rd edn. Oxford: Blackwell Science, 2001

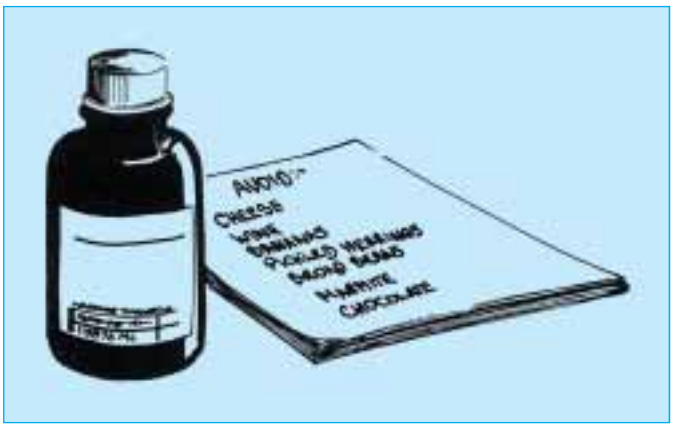
References

1 Holland B, Welch AA, Unwin ID, Buss DH, Paul AA, Southgate DAT. *McCance & Widdowson's the composition of foods*, 5th edn. Cambridge: Royal Society of Chemistry, 1991.
2 Crawley H. (Ministry of Agriculture, Fisheries and Food.) *Food portion sizes*. London: HMSO, 1988.
3 Mullen PD. Compliance becomes concordance. Making a change in terminology produces a change in behaviour. *BMJ* 1997; **314**: 691-2.
4 Jenkins DJA, Wolever TMS, Taylor RH *et al*. Rate of digestion of foods and postprandial glycaemia in normal and diabetic subjects. *BMJ* 1980; **2**: 14-17.
5 Parillo M, Giacco R, Riccardi G, Pacioni D, Rivellese A. Different glycaemic responses to pasta, bread and potatoes in diabetic patients. *Diabet Med* 1985; **2**: 374-7.



Diagnostic tests

- Glucose tolerance tests
- Faecal fat test for steatorrhoea
- Urinary screening for hypercalciuria
- 5-hydroxyindoleacetic acid (5-HIAA) for carcinoid tumours
- 4-hydroxy-3-methoxy-mandelic acid (VMA) for phaeochromocytoma
- Creatinine



6 Ha TKK, Lean MEJ. Technical review. Recommendations for the nutritional management of patients with diabetes mellitus. *Eur J Clin Nutr* 1998; **52**: 467-81.
7 Franz MJ, Bantle JP, Beebe CA *et al*. Technical review. Evidence-based nutrition principles and recommendations for the treatment and prevention of diabetes and related complications. *Diabetes Care* 2002; **25**: 148-98.
8 Jung R. Obesity as a disease. *Br Med Bull* 1997; **53**: 307-21.
9 Leeds A, Brand Miller J, Foster-Powell K, Colagiuri S. *The Glucose Revolution*. London: Hodder and Stoughton, 1998.
10 Turner RC, Millins H, Neil HAW *et al*. Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: UK prospective diabetes study (UKPDS:23) *BMJ* 1998; **316**: 823-8.
11 Report of a Joint FAO/WHO Expert Consultation (Rome 14-18 April 1987). *Carbohydrates in human nutrition*. Rome: FAO, 1998.
12 Foster-Powell K, Miller JB. International tables of glycaemic index. *Am J Clin Nutr* 1995; **62**: 871S-93S.
13 Baxter A, Wright AD. Diabetic dietary prescriptions (letter). *BMJ* 1989; **299**: 124.
14 Allman MA, Truswell AS, Stewart PM *et al*. Vitamin supplementation of patients receiving haemodialysis. *Med J Aust* 1989; **150**: 130-3.

14 Food poisoning

Patrick G Wall, Ciara E O'Reilly

Bovine spongiform encephalopathy/new variant Creutzfeldt–Jakob disease (CJD), foot and mouth disease, *E. coli* O157:H7, salmonella in poultry and eggs, antibiotic-resistant strains of bacteria and residues in food, listeria in paté and cheeses, have all damaged the public's confidence in our food supply. Coupled with this, consumers' confidence in the ability of the regulatory agencies to protect the food supply, and in the commitment of the food industry to protecting consumers' health, has also been damaged. There are many factors contributing to consumers' concerns regarding food safety as outlined in the box opposite.

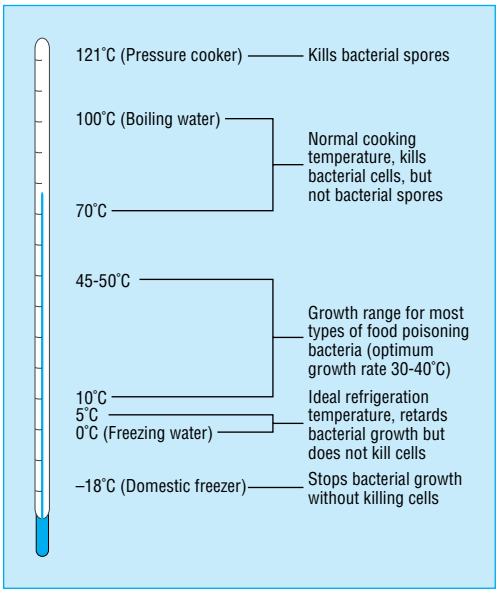
Changes in consumer behaviour including more eating out, bulk shopping, and a desire for more convenience foods have contributed to the increasing incidence of foodborne infectious disease. Busy lifestyles, two parents working outside the home and a decreasing knowledge of cooking are contributing to an increasing demand for convenience "ready to cook" and "ready to eat" food, and more and more food is being prepared outside the home. This is increasing the length of the food chain, providing more opportunities for things to go wrong. Microbes do not respect national frontiers and the increasing global distribution and mass processing of food products has given microbes the opportunity to disseminate widely.[2] In the poorer non-industrialised countries other factors that contribute to an unsafe food supply are expanding urbanisation, the increased dependence on stored foods, insufficient access to safe water and essential facilities for safe food preparation. For example an increased volume of street foods prepared under poor sanitary conditions increases the risk of foodborne illness.

Producing safe food involves a chain of responsibility including primary producers, processors, distributors, wholesalers, retailers, commercial caterers and consumers. Everyone in the chain from "plough to plate" has a role to play. See box on page 100 for details. There is no room for complacency with foodborne diseases, as what may be a mild dose of diarrhoea for a robust young adult can be a life-threatening illness for an infant, a frail elderly person or an individual suffering from some concurrent morbidity. Globally hundreds of millions of people are affected by illnesses caused by contaminated food. The toll in terms of human life and suffering is enormous, particularly among infants and young children, the elderly and other vulnerable groups. Over the next few decades the number of elderly people in the world is predicted to increase rapidly. By the year 2025 more than 1000 million of the global population will be over 60 years old and more than two-thirds will live in developing countries.[3] There are several key factors which relate to an increased risk of foodborne disease among the elderly: (a) the increase in the proportion and number of elderly people,[3] (b) physiological factors, such as changes in the functions of the gastrointestinal tract and immune systems, that predispose an elderly person to infections, (c) pathological factors associated with ageing, such as disability and losses in sensory functions, and (d) social changes in health care of the elderly.[1]

Animals and poultry are the reservoir for many foodborne infections (salmonella, campylobacter, *E. coli* O157:H7, cryptosporidia, and yersinia). Intensive farming of livestock to produce more and cheaper food has brought with it inherent dangers. It is possible to control the food, water, air and environment in these units and produce virtually disease-free

Factors contributing to food safety concerns[1]

- New food preparation
- New storage methods
- Changing lifestyles—convenience foods
- Globalisation in food trade
- Changes in food production on the farm
- New systems of food processing
- Longer distribution chains
- Increasing number of vulnerable groups (the elderly, immunocompromised individuals, etc)



Summary of the effects of temperature on bacterial growth

animals. Husbandry practices are not always ideal, however, and the high stocking density facilitates the transmission of microbes, which can result in large amounts of contaminated material entering the food chain. Farms and abattoirs are not operating theatres and even with the best operational procedures a proportion of raw meat and poultry may contain harmful microbes. Care must be taken, therefore, to avoid transferring microbes from raw product to other foods that will not be heat-treated before consumption. Raw meat and poultry should be sufficiently cooked to kill all harmful microbes and refrigerated prior to cooking to prevent multiplication of harmful microbes. Bacteria such as *Listeria monocytogenes* (found in soil and silage as well as animals) and *Bacillus* species (soil, dust, etc) are widely distributed in the environment and may thus be found in plant (for example, coleslaw, rice) as well as animal (for example, soft cheeses, unpasteurised milk) products.

Protozoal pathogens

Causative agent	Incubation period	Usual duration of symptoms	Common clinical features*	Mode of transmission*
Cryptosporidium sp	2-5 days	<3 weeks	D	W, An, X
Entamoeba histolytica	2-4 weeks	Variable	D, B	X
Giardia intestinalis	5-25 days	Variable	D, P	X, F, W

* See key in table on page 97
Adapted from Department of Health[4]

Bacterial pathogens

Causative agent	Usual incubation period	Duration of symptoms	Common clinical features	Mode of transmission
Aeromonas sp	Unknown	Varied	V, D	W, F
Bacillus sp		<36 h		F
B. cereus				
Emetic syndrome	1-5 h		N, V, D, P	
Diarrhoeal syndrome	8-16 h		D, V, N, P	
B. subtilis	1-4 h		N, V, D	
B. licheniformis	2-14 h		D, P	
Campylobacter sp	2-5 days	2 days to 1 week	D, P, Fe	F, W
Clostridium perfringens	12-18 h	24 h	D, P	F
Clostridium botulinum	12-36 h	Can be several months	Neurological signs	F, wound contamination
Escherichia coli				
Attaching and effacing (AEEC)	Unknown	Unknown	D	F
Enteroaggregative (EAggEC)	20-48 h	Unknown	D, B	F
Enteroinvasive (EIEC)	12-72 h	5-7 days	D, B	F, W
Enteropathogenic (EPEC)	12-72 h	<2 weeks	D	X, F, W
Enterotoxigenic (ETEC)	12-72 h	3-5 days	D	F, W
Verocytotoxin-producing (VTEC)	1-6 days	4-6 days (not HUS)	D, B, HUS	F, X, W
Listeria monocytogenes	1-10 weeks	Variable	Flu, meningitis, abortion	F, congenital, DC
Salmonellas (non-enteric fever)	12-72 h	<3 weeks	V, D, Fe	F, X
Salmonella typhi/paratyphi	1-3 weeks	10-14 days	N, Fe	F, X, W
Shigella sp	1-7 days	<2 weeks	D, B	X, F, W
Staphylococcus aureus	2-4 h	<12-48 h	V, P, Fe	F
Vibrio cholerae (O1, O139)	2-3 days	<7 days	D	W, F
Vibrio parahaemolyticus	12-18 h	<7 days	D	F
Vibrio vulnificus	12 h to 3 days	High mortality if concurrent liver disease or immunosuppression	Septicaemia, shock	F, DC
Yersinia sp.	3-7 days	1-3 weeks	D, P, Fe	F

Adapted from Department of Health (2000)[5]
Key to tables: Bacterial pathogens, protozoal pathogens, viral pathogens and food poisoning biological toxins associated with seafood

Clinical features:
B Blood in stool
D Diarrhoea
Fe Fever
HUS Haemolytic uraemic syndrome
N Nausea
P Abdominal pain
J Jaundice
V Vomiting

Mode of transmission:
Aer Aerosol
An Animal contact
F Food
W Water
X Person to person (faecal-oral)
DC Direct contact with infected material

Food poisoning is defined by WHO as: "Any disease of an infectious or toxic nature caused by or thought to be caused by the consumption of food and water".[5] This definition includes all food and waterborne illness regardless of the presenting symptoms and signs. Thus it includes not only acute illnesses characterised by diarrhoea and/or vomiting, but also illnesses presenting with manifestations not related to the gastrointestinal tract. Heavy metals, pesticides, and other chemicals can also cause food poisoning.

Most food poisoning whether caused by bacteria, viruses, protozoa or toxins usually results in symptoms which include diarrhoea, vomiting, nausea, abdominal pain, and occasionally fever. However, there are other agents that result in food poisoning but do not usually result in gastrointestinal symptoms. These include *Clostridium botulinum, Listeria monocytogenes*, hepatitis A virus, and poliovirus. Laboratory tests are required to identify the aetiological agent but the symptoms, incubation period, and duration of illness may suggest the likely cause.

National figures on the incidence of food poisoning in most countries arise from a combination of notifications of clinical disease, reports of laboratory-confirmed infections, and details on outbreaks. In the United Kingdom all clinicians have a statutory duty to notify to the proper officer of the local authority, cases, or suspected cases, of food poisoning. The meaning of the term "food poisoning" is not defined in the relevant UK legislation, the Public Health (Control of Diseases) Act 1984, and this has previously led to confusion as to which cases should be notified, but in 1992 the WHO definition was adopted. Notification is not contingent on laboratory confirmation of infection, and delaying notification until laboratory confirmation is available defeats the purpose of a rapid notification system designed to enable effective timely intervention at local level. An outbreak is defined as an incident in which two or more people, thought to have a common exposure, experience a similar illness or proven infection.[5] A general outbreak is defined as one that affects members of more than one household or residents of an institution, distinguishing it from an outbreak affecting one family.[5]

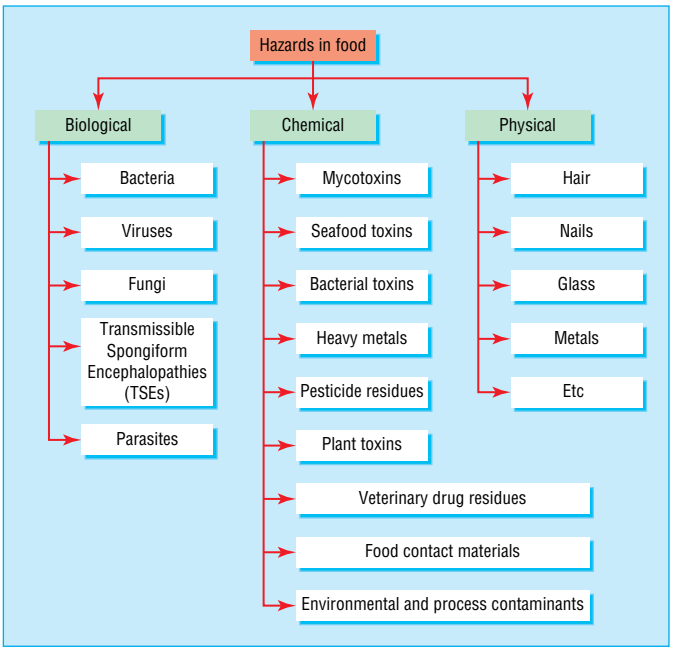
Notifications of food poisoning based on clinical diagnosis are a very crude estimation of the true extent of the burden of ill health due to food poisoning. The symptoms can occur unrelated to food poisoning and only a proportion of cases are actually notified (clinicians being unaware of the need to notify or omitting to do so). Gross under-notification of food poisoning occurs as is well documented for more serious infections.[6] Furthermore, many cases do not seek medical attention.[6]

Similarly the reports of laboratory-confirmed gastrointestinal pathogens represent just a fraction of the true incidence of these pathogens, as only a proportion of cases seek medical attention and only a subset of these have a sample submitted for analysis.[7] Not all of these will have a pathogen

Viral pathogens

Virus	Incubation period	Duration of symptoms	Common clinical features*	Mode of transmission*
Adenovirus	7-8 days	9-12 days	D, V	X
Astrovirus	3-4 days	2-3 days	V, D, Fe	X, F, Aer
Calicivirus	1-3 days	1-2 days	V, D, Fe	X, F, Aer
Hepatitis A	15-50 days	Variable	J	F, W, X
Rotavirus	1-2 days	4-6 days	D, V	X, F
SRSV	1-3 days	1-3 days	V, D, Fe	X, F, Aer

* See key in table on page 97
Adapted from Department of Health[4]



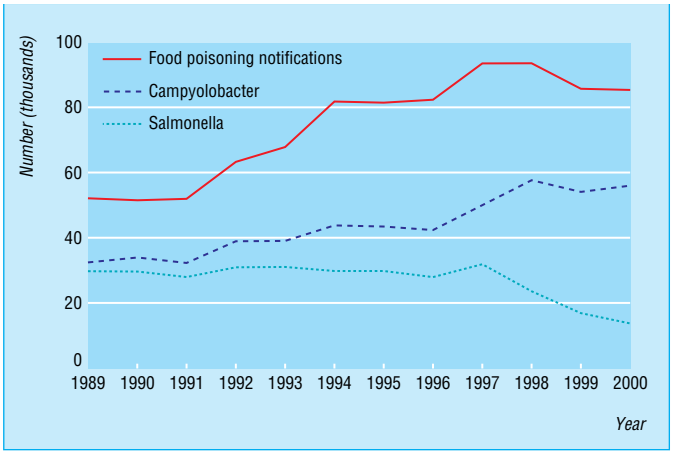
The biological, chemical, and physical hazards in food (adapted from[1])

Food poisoning biological toxins associated with seafood

Agent	Incubation period	Duration of symptoms	Common clinical features*	Mode of trans-mission
Paralytic shellfish poisoning	Minutes to hours	Several days	Tingling, numbness, incoordination, respiratory distress	F
Scombrotoxin	Minutes to hours	<12 hours	Allergic type; facial flushing V, N	F
Ciguatera fish poisoning	1-24 hours	Weeks	V, D + neurological	F

*See key in table on page 97

identified (particularly if the agent is a virus) and not all pathogens identified are reported centrally. The exact proportion in each category is unclear and may vary for different organisms. The National Infectious Intestinal Diseases Study attempts to estimate the true prevalence of the different pathogens and the degree of under-diagnosis and under-reporting.[4] Laboratory reports most likely represent patients at the severe end of the spectrum of diarrhoeal disease and are therefore a biased sample. In addition, laboratory reports of potentially foodborne microbes may not all have arisen by this route, with animal contact and person-to-person transmission accounting for a proportion. It is often only when cases are investigated epidemiologically that the exact route of transmission is identified. In England and Wales reports of laboratory confirmed infections are collated and published by the Communicable Disease Surveillance Centre (CDSC), the Epidemiology Unit of the Public Health Laboratory Service (PHLS). Similar national surveillance centres exist in most countries. Campylobacter and salmonella are by far the most commonly reported enteric pathogens and the most likely aetiological agents in a lot of the food poisoning notifications.



Food poisoning notifications and laboratory reports of campylobacter and salmonella: England and Wales 1992-2000 (Source: PHLS CDSC)

Diagnosis

A microbiological diagnosis is often not necessary for successful treatment of an individual suffering from food-related gastroenteritis (see box opposite for details when it is necessary).

Faecal culture remains the gold standard for the identification of most foodborne pathogens. Electron microscopy is used for the identification of Norwalk and other viruses. Newer more rapid molecular methods are coming on stream. Clinicians should consult their local laboratory for information on the exact diagnostic tests on offer so that they can take the appropriate specimens and request the optimum investigations.

Microbiological diagnosis and food-related gastroenteritis

Microbiological diagnosis is necessary for individuals suffering from food-related gastroenteritis:

- in severely ill, septicaemic or immunosuppressed patients;
- if clusters of similar cases are to be recognised and outbreaks controlled early;
- if the priority pathogens are to be identified to permit tailored public health interventions;
- to monitor ongoing trends and evaluate interventions;
- in nursery children, special needs children and the elderly in institutions where personal hygiene may be poor and there is a risk of secondary spread;
- for food handlers and healthcare workers to exclude excretors of typhoid and verocytotoxic *E. coli* (VTEC);
- when a patient wishes to take legal action against a restaurant or food retailer.

Treatment

In uncomplicated cases of acute gastroenteritis, symptomatic therapy alone is recommended; that is oral rehydration combined with antipyretics. Antimotility drugs are not advised for acute attacks as they have a very limited role as adjuncts to fluid and electrolyte replacement. Antibiotic therapy is only recommended in high-risk groups, for example, patients with underlying disease such as sickle cell anaemia, immunocompromised patients, cases who are more likely to develop complications, and those with prolonged fever and extraintestinal infections. Metronidazole is indicated if symptoms are caused by giardia.

Economic consequences of foodborne illness

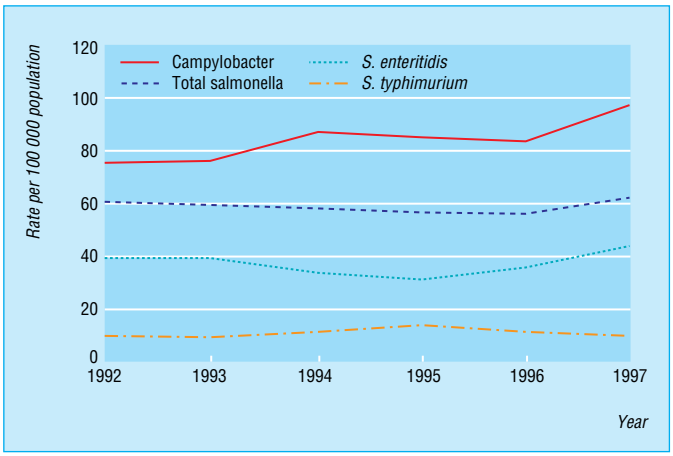
As well as morbidity and mortality associated with foodborne diseases, there are direct economic costs to society in treating sick people. There are also economic losses due to food being rejected in export markets, and loss of consumer confidence in brand names on the domestic market. The US government has estimated the cost of human illnesses of seven foodborne pathogens to be between US$ 5.6 to 9.4 billion annually. The cost of salmonellosis in England and Wales in 1992 was estimated at between US$ 560 and 800 million.

Economic consequences contributing to food safety concerns[1]

- Economic costs to society in treating sick people
- Economic losses due to rejection of food in export markets
- Loss of consumer confidence in brand names

Over 70% of costs were directly associated with treatment and investigation of cases, and costs to the economy of sickness related to absence from work. Estimates in 1999 showed that foodborne diseases annually cause approximately 76 million illnesses, 325 000 hospitalisations and 5000 deaths in the United States each year.[8] The BSE crisis in Britain is predicted to cost over US$ 5.5 billion. These costs are associated with the loss of export sales, compensating farmers, paying slaughterhouses and renderers, and research and control measures. The dioxin crisis in Belgium in 1999 saw the fall of the national government and economic costs in excess of US$ 2 billion to the Belgium economy.[1]

The true estimates of foodborne disease and the likely economic costs are unknown. In industrialised countries only a small proportion of cases of foodborne diseases come to the notice of health services and even fewer are investigated. Very few non-industrialised countries have established foodborne disease reporting systems, and in those that have, only a small fraction of cases are reported. The incidence of foodborne disease is likely to be far higher than in industrialised countries because of such factors as poor nutritional status, inadequate sanitation, low standards of hygiene and reduced likelihood of immediate medical attention. There are many hidden costs associated with foodborne illnesses. Tourists will shun regions with a high incidence of foodborne illness as will multinational firms choosing the location of foreign investment. Repeated bouts of diarrhoea damage children's mental and physical development and play havoc with their schooling.[1]



Principal faecal pathogens identified in humans: England and Wales 1992-97 (Source: PHLS CDSC)

Principal foodborne pathogens

Bacteria

Campylobacter

The importance of campylobacter as a human pathogen was recognised in the 1970s and it is the commonest gastrointestinal pathogen isolated from humans in the United Kingdom.[9] These bacteria do not usually multiply on food as they will not grow at temperatures below 28°C and grow best at 42 °C. Infection in humans usually causes an acute self-limiting enterocolitis. Bacteraemia and reactive arthritis are rare complications and an association with Guillain–Barré syndrome has been observed.[10] Large outbreaks are rarely recognised and reported to the Communicable Disease Surveillance Centre in England and Wales: only 240 (0.2%) of 122 250 cases of campylobacter infections reported between 1992 and 1994 were associated with general outbreaks.[11] There are two principal species, *C. jejuni* and *C. coli*, although most cases are reported without a species. *C. jejuni* is the most common cause of bacterial food poisoning in many industrialised countries. These are fragile organisms which require micro-aerophilic conditions for growth and do not survive or multiply very well on foods. However, the low infectious dose means that a small level of contamination may result in illness. Campylobacter may be transmitted by food, particularly poultry,[12] unpasteurised milk,[13] and contaminated water.[14] *Campylobacter* spp are especially common in birds but have also been isolated from a wide range of animals. Wild birds are considered to be an important reservoir of infection for domestic and food animals. Cases have also been associated with the consumption of pasteurised milk where the milk bottle tops had been pecked by birds, who presumably transferred the bacteria on their beaks.[15] Although many foods may be contaminated with campylobacter, it is considered by many authorities that poultry and poultry products are of particular importance as a source of human infection. *Campylobacter* spp do not multiply very effectively in most foods, however, *Campylobacter* spp may

Recommendations to control transmission of *Campylobacter* spp in the food chain[16]

Recommendations to control transmission of *Campylobacter* spp (and other foodborne pathogens) from farm to fork include:

- poultry industry complying with the highest possible degree of biosecurity measures
- industry development and implementation of evidence-based standard operating procedures, in order to prevent or minimise product contamination with *Campylobacter* spp
- food businesses should implement and document a food safety management system based on the principles of the Hazard Analysis Critical Control Point (HACCP)
- staff involved in food production should be appropriately trained in food safety to a level commensurate with their work activities
- the provision of a supply of potable water for food preparation
- consumers should practise basic good hygiene when handling food and should cook high-risk raw foods thoroughly[16]

survive through the food distribution system and because consumption of a small number of organisms (500 or less) may be associated with illness, proliferation in food is not a prerequisite for infection.[16]

In the United Kingdom there is a characteristic seasonal distribution of reports, with a peak in the early summer, in contrast to reports of salmonella which reach their maximum in the late summer and autumn. There are also regional variations in the reporting of campylobacter infections which may reflect a higher incidence in rural than urban populations.

Salmonella (non-typhoid)

Salmonella is a genus made up of over 2200 serotypes. They are widely dispersed in nature, being found in the gastrointestinal tracts of domesticated and wild mammals, reptiles, birds, and insects (apart from *S. typhi* and *S. paratyphi* which only colonise humans). Although there is a vast range of serotypes, a small number predominate. Salmonella is the second most commonly reported bacterial cause of gastroenteritis after campylobacter. In England and Wales in 1997 there were 32 169 human cases of salmonellosis but four serotypes accounted for 90% of cases (*S. enteritidis* 22 806, *S. typhimurium* 4695, *S. hadar* 692, and *S. virchow* 650). Although the clinical management of most cases infected with the different serotypes is similar, for epidemiological purposes, it is important that isolates are sent to a reference laboratory for definitive typing. Salmonella can be phenotyped using serotyping, phage typing and antibiograms, and genotyped using molecular methods including plasmid profile analysis, pulsed field gel electrophoresis, and DNA sequencing. The National Reference Laboratory in England and Wales is the PHLS Laboratory of Enteric Pathogens. Most countries have such a facility and in the European Union all such labs are collaborating in the Enter-net Project to standardise methodology and share surveillance and outbreak data. Detailed typing can be used to link strains identified in humans, food samples, environment samples and livestock in the course of outbreak investigations. The standardisation of phage typing techniques, the electronic transfer of molecular typing patterns, formal and informal links established through international networks including Enter-net and the notification of national outbreaks to these networks are extremely important, particularly in the case of an international outbreak.

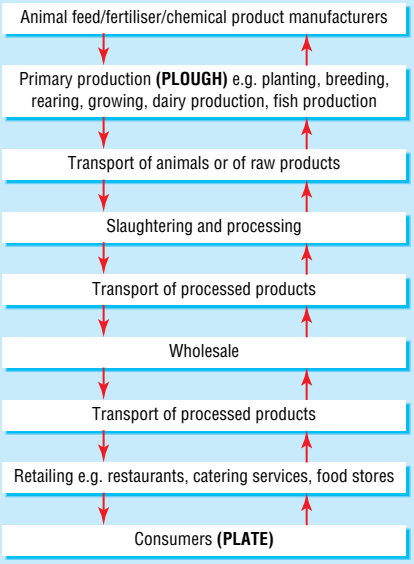
Incidence rates of infection are highest in children under 10. This may reflect increased susceptibility but is also likely to be due to these young children being brought to medical care and having faecal specimens submitted for examination. Most cases suffer from gastroenteritis with a subset (<1%) suffering from invasive disease.

Undercooked (or improperly heat treated) food from infected food animals is most commonly implicated. There has been a dramatic rise in human salmonellosis in the United Kingdom since the mid 1980s due to an unprecedented increase in one subtype, *S. enteritidis* phage type (PT) 4. The reservoir for this subtype is infected poultry. Humans become infected from consuming under-cooked poultry, raw eggs, dishes made from the same, or foods cross-contaminated from these.[17-19] Egg-associated salmonellosis is caused when *S. enteritidis* infects the ovaries of apparently healthy hens and these hens lay contaminated eggs. If the eggs are eaten raw or undercooked, salmonellosis may result. The increase in

Traceability from plough to plate

The plough to plate (or farm to fork) approach to food safety embraces all elements which have an impact on the safety of food, at every level of the food chain. The communication of information throughout the food chain is a critical element of this approach. This approach encompasses the production of all foods and can be applied not only to meat but also to milk, eggs, fish and other products from product from aquaculture, as well as fruit and vegetables.

The Stages of Food Production in the traceability from farm to fork approach:



S. enteritidis PT4 in the United Kingdom has been seen in many countries and this pathogen is now a global problem.

Although *S. enteritidis* has declined by 50% since its peak year of 1997 and *S. typhimurium* by 40% in England and Wales since 1995, these serotypes continue to dominate. The decline in *S. enteritidis* phage type 4 (PT4) is the result of vaccination of poultry flocks but in its place new types are emerging.

In 2001 in England and Wales there was a significant increase in salmonellosis which is associated with a range of salmonellas that have links with foreign travel and the importation of food stuffs (for example, contaminated desiccated coconut).[20] Another emerging global problem is *S. typhimurium* definitive type 104 (DT 104). It is the second most prevalent salmonella found in humans in England and Wales, increasing from fewer than 250 in 1990 to 4006 cases in 1996. Over 90% of isolates are multi-antibiotic resistant to ampicillin, chloramphenicol, streptomycin, sulphonamides and tetracyclines, with a subset of the strains having reduced susceptibility to trimethoprim and ciprofloxacin.[21] This subtype is prevalent in cattle, sheep, pigs, and poultry, and therefore a wide array of foods can be contaminated. Reduced susceptibility to fluoroquinolone drugs has also emerged in the poultry-associated serotypes *S. hadar* and *S. virchow*. The development of these strains is a consequence of the use of antibiotics in animal husbandry.[22] This acquisition of resistance has resulted in a progressive reduction in options for the management of invasive salmonellosis in humans.

Enterovirulent E. coli *including* E. coli *O157*

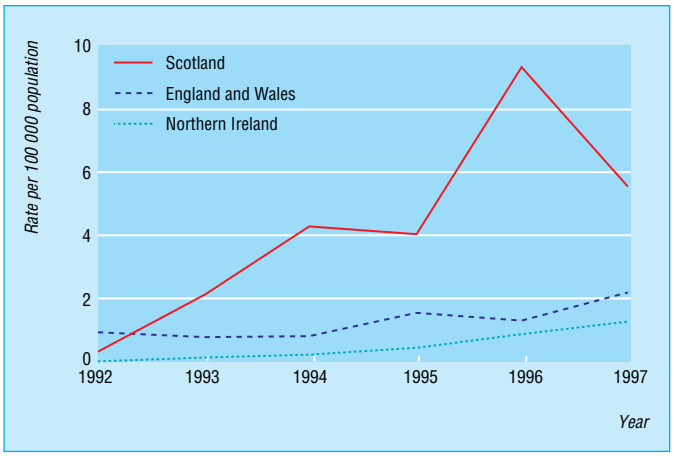
The enterovirulent *E. coli* include all those *E. coli* believed to be associated with diarrhoea. Enterotoxigenic *E. coli* are an important cause of travellers' diarrhoea, enteropathogenic *E. coli* are an important cause of childhood diarrhoea, enteroinvasive *E. coli* cause a disease that is similar to shigella-like dysentery, enteroaggregative *E. coli* have been associated with diarrhoeal disease in infants and travellers, and enterohaemorrhagic *E. coli* are associated with haemorrhagic colitis and haemolytic uraemic syndrome.

Of the latter, currently the most important is verocytoxin (VT) producing *E. coli* O157. It is now a major foodborne public health problem, with one outbreak in central Scotland in 1996 resulting in 500 cases and 20 deaths.[23] Diarrhoea caused by VTEC O157 is often accompanied by bleeding or haemorrhagic colitis and may be complicated by haemolytic uraemic syndrome (HUS) or thrombotic thrombocytopenic purpura.[25] HUS is most likely in children under 5 years and is a major cause of renal failure in children.

E. coli O157 is part of the normal gut flora of cattle primarily, and to a lesser extent sheep. Raw or undercooked minced beef or cooked meats cross-contaminated from raw meat are the high risk foods.[25–27] However, any food that could possibly be contaminated with ruminant faeces poses a risk and infections have been associated with vegetables, cheese, milk, fruit juice, and water. The organism is not particularly heat-resistant and is easily killed by cooking; however the low infective dose means that foods must be cooked thoroughly and be protected post-cooking from cross-contamination with raw foods. *E. coli* O157 first came to prominence as a human pathogen in 1982.[28] In addition to the foodborne route, infections are transmitted by the person to person route, particularly in nurseries and old people's homes, and by direct contact with farm animals.[26] Nobody knows where this strain of *E. coli* came from and, more worryingly, nobody knows where it is going to, as the incidence is increasing in many countries including the United Kingdom where rates of infection vary geographically. In 1996 the highest rate of infection was seen in

Low infectious dose of verocytotoxigenic *E. coli* (VTEC)

The number of VTEC required to cause illness is very low. The precise infectious dose is not known but it has been reported to be as low as 10 cells. VTEC can survive the high acidity of the stomach and with such small numbers, infection can occur without any growth of the bacteria in food.[24,25]



Human *E. coli* O157 infections: UK 1992-97 (Source: Department of Health)

Scotland with a rate of 9.25 per 100 000 population compared with 1.27 in England and Wales and 0.85 in Northern Ireland.

Listeria monocytogenes

L. monocytogenes causes serious human illness such as perinatal infections, septicaemia, and meningitis. In pregnant women it can cause a flu-like illness, which can result in miscarriage, stillbirth or birth of a severely ill infant. More recently, it has been recognised that *L. monocytogenes* can cause mild gastrointestinal symptoms. Listeriosis is a comparatively rare but very serious human illness with a case fatality ratio of ~20%. Highly susceptible individuals include pregnant women, neonates, elderly people and immunocompromised individuals. *L. monocytogenes* is ubiquitous in the environment and so can be transferred to foods from a wide variety of sources. Foods associated with transmission are characteristically highly processed, have extended shelf lives at refrigeration temperatures, are capable of supporting the growth of *L. monocytogenes* and are consumed without further cooking. *L. monocytogenes* is psychrotrophic, —that is, potentially capable of growing, albeit slowly, at refrigeration temperatures as low as 0°C.[1]

Individuals highly susceptible to listeriosis include

- Pregnant women
- Neonates
- Elderly people
- Immunocompromised individuals

Clostridium botulinum

C. botulinum produces a neurotoxin which is one of the most poisonous substances known to man. Foodborne botulism occurs when food becomes contaminated with spores from the environment, which are not destroyed by initial cooking or processing. If the food provides a suitable environment for growth the spores will germinate leading to toxin production. The toxin itself is heat sensitive and so a further heat treatment of the food would prevent illness. Botulism can be prevented by using food preservation methods that are designed to inhibit the growth of *C. botulinum*. For example, low acid food, (pH > 4.4) canned foods are heat treated to 121°C for three minutes (known as the "botulism cook") or equivalent.[1]

Parasites

Foodborne parasitic diseases are a major public health problem that affects millions of people, predominantly in non-industrialised countries. The incidence of parasitic disease associated with the consumption of foods of animal origin has declined in industrialised countries in recent years, where improvements in animal husbandry and meat inspection have led to considerable safety and quality gains. The situation in non-industrialised countries is very different in that these diseases are associated with poor standards of sanitation and hygiene, low educational standards, and extreme poverty.

Foodborne parasitic diseases

Foodborne parasitic diseases are a major public health problem that affect millions of people, predominately in non-industrialised countries. The incidence of parasitic disease associated with the consumption of foods of animal origin has declined in industrialised countries in recent years, where improvements in animal husbandry and meat inspection have led to considerable safety and quality gains. The situation in non-industrialised countries is very different in that these diseases are associated with poor standards of sanitation and hygiene, low educational standards and extreme poverty.

Parasites are organisms that live on other living organisms known as hosts. They may be transmitted from animals to humans, from humans to humans, or from humans to animals. Foodborne parasitic disease occurs when the infective stages of parasites are eaten in raw or partially cooked protein foods, or in raw vegetables and fruits that are inadequately washed prior to consumption. These organisms then live and reproduce within the tissues and organs of infected human and animal hosts, and are often excreted in faeces. The parasites involved in foodborne disease usually have complex life cycles involving one or two intermediate hosts. The foodborne parasites known to cause disease in man are broadly classified as helminths (multicellular worms) and protozoa (single-celled microscopic organisms). These include the major helminthic groups of trematodes, nematodes, and cestodes and some of the emerging protozoan pathogens, such as cryptosporidia and cyclospora. The illnesses they can cause range from mild discomfort to debilitating illness and possibly death.[1,27,29]

Principal foodborne parasites[1,27,29]

Parasite	Food involved in transmission to man	Pathogenesis
(a) Helminths		
(i) Trematodes		
Opisthorchis (liver fluke)	Many species of freshwater fish including the family *Cyprinidae* (for example whitefish, carp, tench) Ingestion of infective larvae (metacercariae) in fish muscle and subcutaneous tissue	The liver flukes, *Opisthorchis viverrini*, *O. felineus* and *Clonorchis sinensis*, are biologically similar, foodborne trematodes which chronically infect the bile ducts and, more rarely, the pancreatic duct and gallbladder of human beings and other mammals
Clonorchis (liver fluke)	Many species (ca. 110) of freshwater fish, mainly *Cyprinidae* (carp, roach, dace). Ingestion of infective larvae (metacercariae) in fish muscle	
Paragonimus (lung fluke)	Raw, salted or partially cooked flesh of fresh and brackish-water crabs, crayfish and shrimps. (Wild boar meat suspected as a source of infection)	Paragonimiasis—symptoms include cough, hemoptysis and pleuritic chest pain
(ii) Cestodes (tape worm)		Symptoms commonly are trivial or absent. Where worms attach to the jejunum patients develop B-12 deficiency anaemia. Massive infections may be associated with diarrhoea, obstruction of the bile duct or intestine and toxic symptoms
Diphyllobothrium	Humans acquire infection by eating raw or inadequately cooked fish	
(iii) Nematodes (round worm)		
Anisakis	Humans acquire infection by eating raw or lightly cooked fish	Usually manifests as cramping, abdominal pain and vomiting
(b) Protozoa		
Giardia intestinalis	Transmission is foodborne, waterborne (faecally contaminated food or water) or person to person spread	Chronic diarrhoea, malabsorption, weight loss
Toxoplasma gondii	Transmission is foodborne (from raw or inadequately cooked infected meat) waterborne or by faecal-oral route (infected cats)	Mostly asymptomatic. In severe cases: hepatitis, pneumonia, blindness, severe neurological disorders. Can also be transmitted transplacentally resulting in a spontaneous abortion, a stillborn, mental/physical retardation
Cryptosporidium parvum	Transmission is foodborne, waterborne, animal-person or by faecal-oral route	Often asymptomatic Abdominal cramps, vomiting, weight loss, diarrhoea

Cryptosporidium

Cryptosporidium parvum is a coccidian protozoan parasite which was identified as an important human pathogen in 1976. Infection rates in the United Kingdom show variation in both age and seasonal distribution. The age distribution reflects endemicity; few infections occur under the age of 1 year, probably due to passive maternal immunity and protection from exposure to the environment. An almost logarithmic increase is observed in toddlers and young children, moderately high rates in young adults, and few cases after the age of 40 years. Seasonal peaks occur in the spring and late autumn. Cryptosporidiosis is a zoonoses with its primary reservoir in calves and lambs. The oocysts are resistant to chlorination and outbreaks associated with drinking water and swimming pools are common.[30,31] Cryptosporidium is highlighted here because it can contaminate the public water supply and extensive outbreaks of gastroenteritis can result.[30] Transmission can also occur by direct contact with animals and by the person to person route. Before the advent of effective antiretroviral therapy, cryptosporidiosis in patients infected with HIV was life threatening.

Viral gastroenteritis **(see box on page 96)**

Although primarily transmitted by the person to person route, transmission by the food or waterborne route has been documented for astrovirus and calicivirus (human calicivirus

and Norwalk-like virus).[32] Norwalk-like virus, also known as Small Round Structured Virus (SRSV), is the most commonly identified foodborne virus.[33] Humans are the reservoir for SRSV and shellfish, particularly oysters grown in sewage-contaminated water, are a source of infection. Infected foodhandlers are another source and can contaminate food during preparation. Aerosolisation of vomitus-containing virus particles has been proposed as a mode of transmission of the virus and may also be a source of food contamination. Clinical and epidemiological features, including stool culture negative for bacteria, duration of illness 12-60 hours, incubation period 24-48 hours and vomiting in >50% of cases, can assist in making the diagnosis when no laboratory results are available.[34] Sensitive detection assays have now revealed that shedding of the virus in faeces may continue for up to a week after the illness subsides.

Principal chemicals affecting food safety

Pesticide residues

Pesticides are chemicals or biological products used to control harmful or undesired organisms and plants, or to regulate the growth of plants as crop protection agents.

Pesticides can also be toxic to humans since certain biochemical pathways are relatively conserved across species as are some enzymes and hormones. In the context of food safety, exposure to pesticides is classified as acute or chronic. An acute intoxication usually has an immediate effect on the body whilst a chronic effect may reveal itself over the life span. The severity depends on the dose and the toxicity of the pesticide compound or breakdown product. Toxic effects that have been identified can include enzyme inhibition, endocrine disruption, and carcinogenic action, depending on the compound in question.

In Europe the control of pesticides is based on Council Directive No 91/414/EEC. Under this regulation, pesticides must be evaluated for safety based on dossiers prepared by their manufacturers. If a pesticide is accepted it is placed on a positive list with a maximum residue limit (MRL) assigned to it.

Pesticide residues in food[1]

- Insecticides
- Fungicides
- Herbicides
- Rodenticides
- Molluscicides
- Plant growth regulators

Veterinary drug residues

Veterinary drugs include antibacterial compounds, hormones and non-steroidal anti-inflammatory preparations. As animal husbandry practices have intensified over the past few decades, antibacterial substances have been increasingly used as growth promoters to increase feed conversion efficiency and for prophylaxis and therapy to prevent outbreaks and treat disease. Similarly, hormones are administered to increase growth rate and meat yield.

The excessive use of antibacterial compounds in animal husbandry has raised concerns about the development of resistant bacteria and the effect that this may have on the usefulness of antibiotics in human medicine. There have also been concerns about the risk of allergic reactions in humans to antibacterial residues in food of animal origin.[1]

Main types of veterinary drugs[1]

Antibacterial compounds

- Aminoglycosides
- β-lactams
- Fluoroquinolones
- Macrolides
- Sulphonamides
- Tetracyclines
- Quinolones

Hormones

- β-agonists
- Resorcylic lactones
- Steroids
- Stilbenes
- Thyrostat

Environmental/industrial contaminants

This group of contaminants are of environmental origin or are by-products of industrial processes. Polyhalogenated hydrocarbons (PHH) are a category of environmental contaminants that includes toxophene, dioxins, and polychlorinated biphenyls (PCBs). Certain PHH are manufactured for use in plastics, paints, transformers, and herbicides although their use is now either banned or severely

restricted. In most industrialised nations the compounds have become ubiquitous in the environment. Hence contamination of the food chain is inevitable and it has been estimated that in Western industrialised countries, 90% of human exposure is through ingestion of contaminated foods like fish and milk.

Foods that are rich sources of fats and oils tend to accumulate PHH because the compounds are lipophilic and bioaccumulate in lipid rich tissues and fluids. Oily fish from areas such as the Baltic Sea, where levels of PHH in the water are high, may contain elevated levels of these contaminants. Similarly, cows that graze on polluted pasture can accumulate unacceptable concentrations of PHH in their milk. A recent incident in Belgium introduced PCBs and dioxins into the food chain via contaminated animal feed resulting from the accidental incorporation of industrial oil into the feed ration.[1]

Prevention and control of foodborne infectious disease

If the incidence of food poisoning is to be reduced, everyone from primary producers, workers throughout the industry to consumers has a role to play. Farms are the reservoir for many foodborne pathogens and these harmful microbes need to be controlled in livestock. Best husbandry practices will assist farmers to deliver livestock and produce free of pathogens and residues to the next link in the chain. Processors must have rigid process controls to ensure that the food is produced with safety paramount. The distribution, wholesale and retail networks must adhere to the best hygienic practices so that food is delivered to the point of sale in the best possible condition.

Modern food control programmes take consideration of the role of the food producers and food processors in managing the safety of their products. The various sectors of food production, processing and distribution possess knowledge about the safety and shelf-life of their products and have an important role in assisting regulators in achieving national food safety goals. One of the most effective ways for the food sector to protect the health of consumers is to base their food safety management programmes on the seven principles of the Hazard Analysis Critical Control Point (HACCP) system. European food hygiene law requires that food businesses identify the steps in their activities which are critical to ensuring food safety and make certain that adequate safety procedures are identified, implemented, maintained and reviewed.

Even with the best control measures in place a proportion of raw meat and poultry may contain harmful microbes. Food handlers, whether working in the commercial setting or in a domestic kitchen, must be aware of this and take appropriate hygienic precautions.

Raising awareness through continuous innovative education programmes for everyone from consumers including school children to workers throughout the food industry is the mainstay for reducing the incidence of foodborne infectious diseases. Foodhandlers must be aware of their responsibility to handle food hygienically. Infected humans are the reservoir for typhoidal salmonellas and SRSV but can potentially spread any infection by contaminating food during preparation if their personal hygiene is poor. Symptomatic individuals should be excluded from preparing food until they have been at least 48 hours symptom free for most microbes, and until microbiological clearance has been confirmed if infected with enteric salmonellas and VTEC.[35] Any individual incapable of

The seven basic principles of a HACCP system*

1. Conduct a hazard analysis.
2. Determine the Critical Control Points (CCPs) in the process.
3. Establish critical limit(s) for each CCP.
4. Establish a monitoring programme to ensure control of the CCP.
5. Establish corrective action to be taken when monitoring indicates that a CCP is not under control.
6. Establish verification procedures to demonstrate the effectiveness of the HACCP system.
7. Establish documentation concerning all procedures and records appropriate to these principles and their application.

*HACCP means Hazard Analysis and Critical Control Points. It is a systematic approach to identifying and controlling hazards (microbiological, chemical, and physical) that could pose a threat to the production of safe food—in simple terms, it involves identifying what could go wrong and planning to prevent it.

Prevention of food poisoning outbreaks[19]

The same four readily preventable faults repeatedly contribute to outbreaks in food poisoning:

- grossly contaminated raw product
- inadequate refrigeration allowing harmful microbes to multiply at warm temperatures to dangerous levels
- cross contamination from raw to cooked food
- insufficient heat treatment when cooking[19]

practising basic personal hygiene should be excluded from working with food.

The table on page 106 provides a list of useful websites.

New variant Creutzfeldt–Jakob disease (nvCJD): is it a foodborne disease?

The magnitude of the epidemic of bovine spongiform encephalopathy (BSE) in the United Kingdom is without precedent in the recorded history of the transmissible spongiform encephalopathies. The public health and the scientific issues have become entwined in the political and commercial response, to increasing national and international concern over the possibility of transmission of spongiform encephalopathy from cattle to man. Medical practitioners and other public health professionals are having to respond to the intense public anxiety regarding the risk of exposure to BSE contaminated material. Documented transmission of BSE to non-human primates indicates that such anxiety is not without some scientific foundation.[36] We are left facing possible present consequences of past events over which we now have no control.

In March 1996 the UK CJD Surveillance Unit described a distinct variant of CJD in 10 people aged under 42 (average age 27 years) with dates of illness since January 1995.[37] This variant had not been previously recognised and is characterised by behavioural change, ataxia, progressive cognitive impairment, and a prolonged duration of illness (up to 23 months) compared to classical CJD. In addition the EEG is not typical of classical CJD and the brain pathology, although showing marked spongiform change and extensive amyloid plaques, is also not typical. The most striking features of these cases are the extensive plaque formation and a pattern of prion protein immunostaining which is unique and remarkably consistent between cases. After reviewing the data on these variant CJD cases the UK Spongiform Encephalopathy Advisory Committee (SEAC) advised the British Government "that in the absence of any credible alternative, the most likely explanation at present is that these cases are linked to exposure to BSE material before the Specified Bovine Offal (SBO) ban was introduced in 1989".[38] The SBO ban prohibits the use of those tissues (brain, spinal cord, thymus, tonsils, spleen, and intestine) most likely to contain the infective agent of BSE in products for human consumption.

Ten nvCJD cases were announced in March 1996. By the end of June 2002, a total of 124 cases of nvCJD (number of deaths from definite or probable nvCJD: 115, number of probable nvCJD cases still alive: 9) had been reported in the United Kingdom. The overall median age at death was 28 (range 14-74 years). The median number of days from onset to diagnosis was 334 days and from onset to death 411 days. Of the 124 cases, 68 (55%) were male. The recent analyses showed that the underlying incidence is increasing by an estimated 18% per year based on date of symptom onset, or 20% per year based on date of death.[41] In relation to the number of nvCJD cases in countries other than the United Kingdom, as in March 2002 there were five definite and probable cases of nvCJD in France, one in Italy,[42] one in the Republic of Ireland and one in Hong Kong (reported to be associated with the United Kingdom).[43] With a possible incubation period of up to 25 years, nobody knows how many people in the United Kingdom and elsewhere worldwide may be currently incubating this new variant CJD. If further cases continue to appear, we could be witnessing the unfolding of one of the greatest man-made public health catastrophes ever.

Possible vehicles associated with principal aetiological agents of food poisoning

Aetiology	Principal possible vehicles
Bacillus species	Reheated rice
Campylobacter	Poultry, raw milk, water
Clostridium perfringens	Cooked red meat and poultry
Clostridium botulinum	Canned products, honey
Cryptosporidium	Water
E. coli (non-VTEC)	Salads, raw milk, vegetables, cheese, water
E. coli (VTEC)	Undercooked meat, unpasteurised milk and cheese, vegetables, water
Hepatitis A	Seafood, water, (any food contaminated by infected food handler)
Listeria monocytogenes	Cheese, paté
Salmonella (non-typhoid)	Poultry, eggs, red meat, dairy products
Salmonella (typhoid)	Water, (any food contaminated by sewage or infected food handler)
Scombrotoxin	Mackerel, tuna, and other scromboid fish
Staph. aureus	Food contaminated by infected food handler
SRSV	Oysters and other seafood (any food contaminated by infected food handler)

Adapted from Department of Health (2000)[5]

Factors reported to BSE agent support the responsibility for nvCJD in humans

There are several factors supporting the hypothesis that the BSE agent is responsible for the emergence of the new form of CJD in humans. These include:

- the temporal association—BSE came first followed by nvCJD
- the geographical association—the UK has had the bulk of the BSE cases and the majority of the nvCJD cases
- a report describing similar clinical, molecular, and neuropathological features in three BSE experimentally infected macaque monkeys[39]
- molecular typing of prion proteins from cases of BSE and nvCJD has revealed that they are indistinguishable from each other and different from sporadic CJD cases, supporting the hypothesis of a causal link[40]

Useful web addresses

Bulletins of foodborne diseases and communicable diseases

Insititute de Veille Sanitaire	http://www.invs.sante.fr/
EPI-Insight, National Disease Surveillance Centre, Ireland	http://www.ndsc.ie/Publications/EPI-Insight/
Canada Communicable Disease Report	http://www.hc-sc.gc.ca/pphb-dgspsp/publicat/ccdr-rmtc/
Communicable Disease Report, Public Health Laboratory Service, UK	http://www.phls.co.uk/publications/cdr/index.html
Communicable Diseases Intelligence, Australia	http://www.health.gov.au/pubs/cdi/cdihtml.htm
EuroSurveillance, European Union	http://www.eurosurveillance.org/eurosurv/index.htm
Infectious Agents Surveillance Report, NIH, Japan	http://www.nih.go.jp/
Jaargang nummer Infectieziekten bulletin, the Netherlands	http://www.isis.rivm.nl/inf_bul/home_bul.html
Kansanterveyslaitos, Finland	http://www.ktl.fi/
NSW Public Health Bulletin, NSW, Australia	http://www.health.nsw.gov.au/public-health/phb/phb.html
EPI-News, Statens Serum Institut, Denmark	http://www.ssi.dk/en/index.html
Weekly epidemiological report, WHO	http://www.who.int/wer

Food science, technology, food safety and public health

Institute of Food Science and Technology	http://www.ifst.org/
FoodNet, Canada	http://foodnet.fic.ca/
International Food Information Council Foundation, USA	http://www.ific.org
National Food Safety Database, University of Florida, USA	http://www.foodsafety.org/
Institute of Food Science and Technology	http://www.ifst.org.uk
Campden and Chorleywood Food Research Association	http://www.campden.co.uk
Leatherhead Food Research Association	http://www.lfra.co.uk
Food Communications Information Service, UCC	http://www.ucc.ie/fcis
Institute of Food Research	http://www.ifr.bbsrc.ac.uk

National and international food safety agencies

Food and Agriculture Organization of United Nations (FAO)	http://www.fao.org/
Organization for Economic Co-operation and Development (OECD)	http://www.oecd.org/EN/home/0,,EN-home-0-nodirectorate-no-no-no-0,FF.html
The Joint FAO/WHO Food Standards Program (Codex Alimentarius Commission)	http://www.codexalimentarius.net/
World Health Organization, Food Safety Program	http://www.who.int/fsf/
Food Standards Australia New Zealand (FSANZ)	http://www.foodstandards.gov.au http://www.foodstandards.govt.nz
The Danish Veterinary and Food Administration	http://www.lst.min.dk/java_enab/f_uk.html
Communicable Disease Surveillance Centre (CDSC)	http://www.phls.co.uk/
Department of Health, UK	http://www.doh.gov.uk/
Food Standards Agency, UK	http://www.foodstandards.gov.uk/
INED "Mortalité, santé, épidémiologie", France	http://matisse.ined.fr/~mesle/
Institut Pasteur, France	http://www.pasteur.fr/externe
Agence française de sécurité sanitaire des aliments (AFSSA), France	http://www.afssa.fr/
Department for Environment, Food and Rural Affairs, UK	http://www.defra.gov.uk/
Directorate-General, Health and Consumer Protection, European Union	http://europa.eu.int/comm/dgs/health_consumer/index_en.htm
Food Safety Authority of Ireland	http://www.fsai.ie
National Institute of Public Health and the Environment (RIVM), the Netherlands	http://www.rivm.nl/
Canadian Food Inspection Agency (CFIA), Canada	http://www.inspection.gc.ca/english/toce.shtml
Center for Disease Control and Prevention (CDC), USA	http://www.cdc.gov/
Food and Drug Administration (FDA), USA	http://vm.cfsan.fda.gov/
Food Safety and Inspection Service (FSIS), USA	http://www.fsis.usda.gov/
National Institutes of Health (NIH), USA	http://www.nih.gov/
United States Department of Agriculture (USDA), USA	http://www.usda.gov/

Conclusion

As society changes, so do the bacteria involved in foodborne disease and this trend is likely to continue. The globalisation of our food supply poses greater risks to consumer health from the mass production and distribution of foods and increased risk for

food contamination. If the regulators and all stakeholders in the food production chain are to ensure that consumers' health is adequately protected and that they are provided with the reassurances that they are seeking, then issues such as BSE, dioxin contamination, antibiotic resistant bacteria, *Salmonella enteritidis* in poultry, *E. coli* O157, and the increasing incidence of campylobacter infection will have to be dealt with. The solution is sequential incremental risk reduction along the food chain with all stakeholders playing their part; also, the communication of any unavoidable residual risk to consumers with clear instructions on how to manage such risks. While there are specific control programmes required, and more research and surveillance needed to understand the epidemiology of different agents, increased hygiene standards across the food chain will have the effect of reducing foodborne disease.

The authors acknowledge the work of Dr David Thomkins, Consultant Microbiologist, Public Health Laboratory Leeds and his colleagues on the National Infectious Intestinal Diseases Study[5] team, in preparing the tables: Bacterial pathogens, protozoal pathogens, and viral pathogens.

References

1 Reilly A, Tlustos C, Anderson W, O'Connor L, Foley B, Wall PG. Food safety: A public health issue of growing importance. In: Gibney MJ, Vorster HH, Kok FJ (eds). *Introduction to Human Nutrition, 2002*. USA: Blackwell Science Publishing Company, 2002.

2 Killalea D, Ward LR, Roberts D *et al*. International epidemiological and microbiological study of outbreak of *Salmonella agona* infection from ready to eat savoury snack. 1 England and Wales and United States. *BMJ* 1996; **313**: 1093-4.

3 US Bureau of the Census, International Population Reports, *An Aging World II*. US Government Printing Office, Washington DC, USA. 1992; **25**: 92-3.

4 Department of Health. *A report of the study of infectious intestinal disease in England*. London: Stationery Office, 2000.

5 Department of Health. *Management of outbreaks of foodborne illness*. London: Department of Health, 1994.

6 Wall PG, de Louvois J, Gilbert RJ, Rowe B. Food poisoning: notifications, laboratory reports and outbreaks: where do the statistics come from and what do they mean? CDR. *Communicable Dis Rep* **4**: R130-5.

7 Feldman RA, Banatvala N. The frequency of culturing stools from adults with diarrhoea in Great Britain. *Epidemiol Inf* 1994; **113**: 41-4.

8 Mead PS, Slutsker L, Dietz V *et al*. Food-related illness and death in the United States. *Emerg Infect Dis* 1999; **5**: 607-25.

9 Skirrow MB. Campylobacter enteritis: a new disease. *BMJ* 1997; **2**: 9-11.

10 Healing TD, Greenwood M, Pearson AD. Campylobacters and enteritis. *Rev Med Microbiol* 1992; **3**: 159-67.

11 Peabody RG, Ryan MJ, Wall PG. Outbreaks of campylobacter infection: rare events for a common pathogen. *Communicable Dis Rep* 1997; **7**: R33-7.

12 Humphrey TJ, Henley A, Lanning DG. The colonisation of broiler chickens with *Campylobacter jejuni*: some epidemiological investigations. *Epidemiol Infect* 1993; **110**: 601-7.

13 Djuretic T, Wall PG, Nichols G. General outbreaks of infectious intestinal disease associated with milk and dairy products in England and Wales: 1992-1996. *Communicable Dis Rep* 1997; **7**: R41-5.

14 Palmer SR, Gully PR, White JM *et al*. Water-borne outbreak of Campylobacter gastro-enteritis. *Lancet* 1983; **i**: 287-90.

15 Hudson SJ, Lightfoot NF, Coulson JC, Russell K, Sisson PR, Sobo AO. Jackdaws and magpies as vectors of milk borne human Campylobacter infection. *Epidemiol Infect* 1991; **107**: 363-72.

16 Food Safety Authority of Ireland. *Control of Campylobacter species in the food chain*. Report, Food Safety Authority of Ireland, Dublin, 2002.

17 Coyle E, Palmer S, Ribeiro C *et al*. *Salmonella enteritidis* phage type 4 infection: association with hens' eggs. *Lancet* 1988; **ii**: 1295-6.

18 Humphrey TJ. Contamination of egg shells and contents with *Salmonella enteritidis*: a review. *Int J Food Microbiol* 1994; **21**: 31-40.

19 Cowden JM, Wall PG, Adak G, Evans H, Le Baigue S, Ross D. Outbreak of foodborne infectious intestinal disease in England and Wales: 1992-1993. *Communicable Dis Rep* 1995; **5**: R109-17.

20 Ward LR, Threlfall EJ, O'Brien SJ. *The changing trends of salmonellosis in England and Wales*. Proceedings of the International Symposium on Salmonella and Salmonellosis, St. Brieuc, France, May 2002; pp. 393-6.

21 Threlfall EJ, Ward LR, Rowe B. Increasing incidence of resistance to trimethoprim and ciprofloxacin in epidemic *Salmonella typhimurium* DT104 in England and Wales. *Eurosurveillance* 1997; **2**: 81-4.

22 World Health Organization. *Medical impact of the use of antimicrobial drugs in food animals*. Report. Geneva: WHO, 1997.

23 The Pennington Group. *Report on the circumstances leading to the 1996 outbreak of infection with* E. coli *O157 in Central Scotland, the implications for food safety and the lessons to be learned*. Edinburgh: Stationery Office, 1997.

24 Doyle MP, Zhao T, Meng J, Zhao S. *Escherichia coli* O157:H7. In: Doyle MP, Beuchat LR, Montville TJ (eds). *Food Microbiology: Fundamentals and Frontiers*. Washington DC: ASM Press 1997, 171–91.

25 Food Safety Authority of Ireland. *The Prevention of E. coli O157:H7 Infection—A Shared Responsibility*. Report, Dublin: Food Safety Authority of Ireland, 1999.

26 Wall PG, Mcdonnell RJ, Adak GK, Cheasty T, Smith HR, Rowe B. General outbreaks of verocytotoxin producing *Escherichia coli* O157 in England and Wales from 1992-1994. *Communicable Dis Rep CDR Rev* 1996; **6**: R26-R33.

27 Reilly A, Käferstein F. Food safety and products from aquaculture. *J Appl Microbiol, Symposium Supplement* 1999; **85**: S249-57.

28 Riley LW, Remis RS, Helgerson SD *et al*. Haemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N Engl J Med* 1983; **308**: 681-5.

29 Benenson AS. Control of communicable disease manual, *An Official Report of the American Public Health Association* 1995; American Public Health Association, Washington DC, USA.

30 Richardson AJ, Frankenburg RA, Buck AC *et al*. An outbreak of waterborne cryptosporidiosis in Swindon and Oxfordshire. *Epidemiol Infect* 1991; **107**: 485-95.

31 Smith HV. Environmental aspects of *Cryptosporidium* species: a review. *J R Soc Med* 1990; **83**: 629-31.

32 Hedberg CW, Osterholm MT. Outbreaks of foodborne and waterborne viral gastroenteritis. *Clin Microbiol Rev* 1993; **6**: 199-210.

33 Luthi TM, Wall PG, Evans HS, Caul EO. Outbreaks of foodborne and waterborne viral gastroenteritis in England and Wales: 1992-1994. *Communicable Dis Rep CDR Rev* 1996; **6**: R131-6.

34 Kaplan JE, Feldman R, Campbell DS, Lookabaugh C, William CG. The frequency of a Norwalk like pattern of illness in outbreaks of acute gastro-enteritis. *Am J Public Health* 1982; **72**: 1329-32.

35 Anon. The prevention of human transmission of gastrointestinal infections, infestations and bacterial intoxications. *Communicable Dis Rep* 1995; **5**: R158-72.

36 Baker HF, Ridley RM, Wells GAH. Experimental transmission of BSE and scrapie to the common marmoset. *Vet Rec* 1993; **132**: 406.

37 Will RG, Ironside JW, Zeidler M *et al*. A new variant of Creutzfeldt–Jakob disease in the UK. *Lancet* 1996; **347**: 921-5.

38 Calman K. New variant of Creutzfeldt–Jakob Disease (CJD). London: Department of Health 1996. PL/CMO (96)5.

39 Lasmezas CI, Deslys JP, Demaimay R *et al*. BSE transmission to macaques. *Nature* 1996; 381: 743-4.

40 Collinge J, Sidle KCL, Meads J, Ironside J, Hill AF. Molecular analysis of prion strain variation and the aetiology of new variant CJD. *Nature* 1996; **383**: 685-90.

41 Andrews N. Incidence trends and short term predictions for variant Creutzfeldt-Jakob disease in the United Kingdom—update. *Eurosurveillance Weekly* 2002; **6**: 020725.

42 Salmaso S. First case of vCJD reported in Italy. *Eurosurveillance Weekly* 2002; **6** (6), 020207.

43 Food Standards Agency. Minutes of the UK Food Standards Agency Board Meeting, 14th of March 2002, Exeter, UK. FSA paper number 02/05/01.

15 Food sensitivity

One man's meat is another's poison

Lucretius 96-55 BC

In affluent countries the idea is now widespread that a variety of symptoms (not just those of classical allergy) are caused by individual (hyper)sensitivity to certain foods or substances in them; that such sensitivity has become more common; and that food processing may have something to do with it. The media, various unorthodox practitioners, and some groups of lay people have spread the "news". Medical practitioners meanwhile are equipped with little information, most of it confusing, and no reliable diagnostic test to answer their patients' needs.

This is one of the most polarised topics in human nutrition. On one hand many lay people are concerned about sensitivity to food and believe that they suffer from it. On the other hand the medical and food science establishment has been fairly dismissive and declares that most food sensitivity (except adult lactase insufficiency) is uncommon.

The subject is at the interface between scientific immunology, food technology, and quackery. Good clinical research has been lacking, but recently a few academic departments have started to apply the methods of clinical science to unravel this confusing area. At present it is impossible to give estimates of the objectively confirmed prevalence of most types of sensitivity to food.

A little epidemiology

Young *et al.*[1] sent questionnaires to 15 000 households, half across the UK, half in their local (High Wycombe) area. 20% of individuals who replied said they had some form of food intolerance. Those who did and were living locally were asked further standard questions. About half had relevant symptoms and some of them agreed to take eight, double-blind, placebo-controlled challenges with cows' milk, egg, wheat, soya, orange, prawns, nuts, and chocolate (disguised in cans of cornflour and suitable flavouring). Each challenge was taken for three days, followed by three days of placebo in random order. Nearly a quarter reacted. From this painstaking work the authors estimated that 2% of British adults have objectively verifiable sensitivity to the eight foods, and a few more may be sensitive to other foods less commonly reported in their questionnaire.

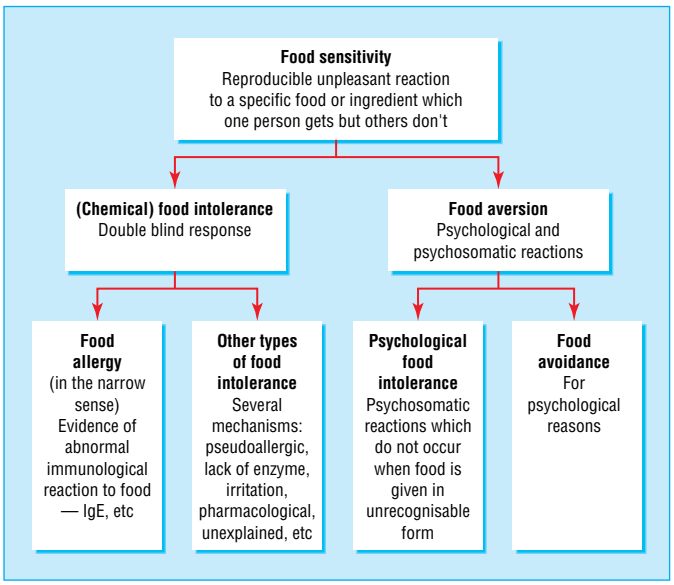
In children 5-7%[2] and 4-10%[3] are estimated to have allergy to one or more foods. This is most prevalent in the first three years of life and fades as children grow older for most foods but not usually for peanuts.

Terminology

The words describing food sensitivity are imprecise and often used to mean different things.

- **Food allergy** is commonly used by lay people (and by doctors talking to patients) as the broad term, including non-immunological (and sometimes even psychosomatic) reactions. In technical communication the term 'allergy' should be confined to immunological reactions.[4-6]
- **Food sensitivity or hypersensitivity** is sometimes used in the narrow sense to mean only immunological reactions.
- **Adverse reaction to food** is not used in this chapter because it conveys no meaning of individual susceptibility and includes food poisoning (dealt with in chapter 14).
- **Pseudoallergic** and **anaphylactoid** reactions are used for, for example, asthma or angio-oedema after food with no immunological abnormalities detectable in the patient.
- **Food idiosyncrasy** is used in some classifications for non-allergic food intolerance.

The classification shown opposite is developed from and compatible with the definitions in the report of the Royal College of Physicians.[5]



Classification of types of food sensitivity

Diagnosis

Diagnosis of food sensitivity is easy when there is a characteristic early response to a food that is eaten at least occasionally. The patient often notices the association and its reproducibility and tells the doctor the diagnosis.

Diagnosis is more difficult, however, if the clinical reaction is delayed or varies or does not always happen, and if the food

Four ways of presentation

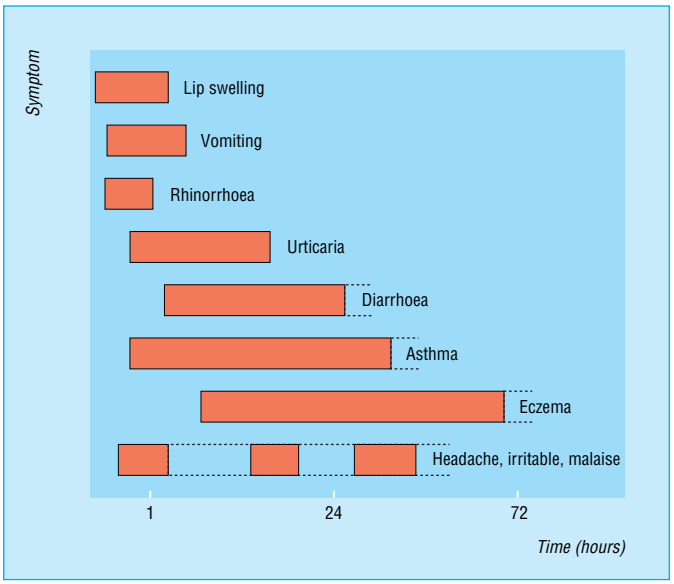
- "Whenever I eat peanuts I get swollen lips, then itchy spots and I sometimes vomit"
- "I can't eat peanuts" (reason why, vague or based on a single episode long ago)
- "I wonder if this rash could be caused by something in my diet?"
- "I've given up eating peanuts because the lady in the health food shop (or the lady next door) says I must be allergic to them"

is eaten most days. Such reactions are also made more difficult to judge if someone else has already incriminated a food on circumstantial evidence or because of prejudice. There are no straightforward diagnostic tests for food sensitivity comparable with the electrocardiogram for coronary disease or the blood urea for renal failure.

Skin tests—Drops of extracts of one or more suspected food antigens are dropped on to the skin and the skin is then pricked or scratched through the drop. A positive response is a wheal and flare within 20 minutes. This indicates the ability of skin mast cells to degranulate in response to the antigen, because they have on their surface IgE specific to the food antigen. Reliability of the test result depends on the quality (specificity) and concentration of the antigen.

The **radioallergosorbent test** (RAST) is a radioimmunoassay performed on serum to show the presence of IgE specific for the food antigen. It is positive in association with IgE mediated food sensitivity.

The skin test and RAST are both sensitive and specific methods for detecting specific IgE, but they do need to be interpreted with care. The **presence** of IgE antibodies to a food does not mean that the patient is **clinically** allergic to it. Many atopic people have IgE antibodies but no symptoms, so positive results must be interpreted in conjunction with a careful history. Foods should not be restricted on the basis of a positive result if no clinical symptoms exist. On the other hand, a negative skin test (or RAST) result is good evidence that a reaction to food is **not** mediated by IgE. A negative skin test result does not exclude food sensitivity mediated by a mechanism other than the IgE immunological reaction.



Time course of symptoms of intolerance[7]

Dietary manipulation

The more general diagnostic procedure to indicate food sensitivity is dietary manipulation and recording of symptoms. Such procedures give diagnostic information in food sensitivity of all types. There are several strategies.

Diet diary

The patient or parent keeps a list of all foods eaten and notes any symptoms. This is simple and cheap and can be done at home for several weeks. It is liable to subjective bias, not suitable if the reaction was serious, and difficult to interpret if the responsible agent is present in several foods.

Temporary exclusion of one or a few suspected foods

Avoidance can be for about a week each time. This is an open trial, liable to subjective bias, but it causes little inconvenience. The method is not suitable if the reaction was serious. It is relatively easy to carry out for foods not eaten every day, such as strawberries and shellfish, but more difficult for wheat and milk, which are widely distributed in the diet.

Few foods (oligoantigenic) diet followed by reintroduction of foods one by one

All the foods that commonly provoke sensitivity reactions are eliminated from the diet for two or three weeks. One food is then added back every three to seven days. Elimination diets carry a risk of nutritional deficiency if taken for long or not properly managed.

Patient-blind or double-blind challenges

After the patient has been stabilised on a standard or elimination diet, foods or ingredients are given in capsules or incorporated into oligoantigenic masking foods. The patient is "blind" to what the foods are. Injection of adrenaline should be available if the reaction was of allergic type.

Responses expected with different types of food sensitivity

Type of food sensitivity	Open trial	Blind trial	Immune response
Food allergy	+	+	+
Food intolerance	+	+	−
Psychosomatic	+	−	−
Food aversion	−	−	−

Few foods diets

- The meat least likely to cause reactions is *lamb*
- The least antigenic cereal is *rice*
- Vegetables: *peeled potatoes* and *lettuce*
- Fruits: *pears* (peeled)
- Fat: a refined vegetable *seed oil*, for example, (cold pressed) *sunflower* or *safflower*
- Drink: *water* and *sugar*
- If the diet continues for more than a week an uncoloured multivitamin supplement should be taken
- Other foods are included in some elimination diets, depending on the type of reaction and the suspected ingredients

Clinical reactions

Urticaria, angio-oedema and anaphylaxis

True food protein allergy, IgE mediated, is potentially the most serious sensitivity reaction to food, and the best understood. A small amount of ingested protein crosses the intestinal wall, and attaches to receptors on T lymphocytes. In an allergic response these secrete interleukin 4 which influences B lymphocytes to secrete IgE. Antibodies of this class bind to mast cells in the tissues and sensitise them. Then when the allergen contacts the sensitised mast cells they react by degranulation and liberation of histamine and other kinins. Reactions occur very quickly, starting with the oral allergy syndrome—lips, cheeks, tongue or throat swell and itch within minutes of contact with the food. Soon reactions may occur in the larynx (oedema, the most dangerous reaction), gastrointestinal tract (vomiting or diarrhoea), the skin (urticaria), and the bronchi (wheezing). Or several systems can be affected, with an acute drop of blood pressure and respiratory difficulty—**anaphylaxis**. Treatment of anaphylaxis is with prompt injection of adrenaline, followed by hydrocortisone.[8] For urticaria and other localised reaction the main line of treatment is with a non-sedative antihistamine.

Urticarial reactions can also be mediated without IgE, for example, be caused by eating large amounts of foods which contain histamine-releasing agents, such as, strawberries, shellfish, papaya, or those that contain histamine itself or other amines, such as some wines, fermented cheeses, and sausages.

Chronic urticaria is not usually associated with IgE-mediated food allergy and there is not usually a history linking it with a particular food. A minority of cases have been found sensitive to certain food additives—benzoic acid and compounds (E210, 211, 212, 213, 216, 218) or the azo dyes tartrazine (E102) or sunset yellow (E110).[10] This can be demonstrated by placebo-controlled challenge testing in a specialist clinic. Physical factors, for example, exercise and warmth, tend to cause urticaria on their own and may exacerbate a reaction to food.

Asthma and rhinitis

Foods can precipitate some attacks of asthma in infancy but come well behind infections. The role of foods in asthma diminishes during childhood and it is an uncommon precipitant in adults. Inhalants, irritants, infections, pollens and moulds, changes in the weather, and exercise are then all more important. Food sensitivity asthma in adults is largely confined to those exposed to dusty grain, flour, coffee, etc, in their work.

Eggs, seafood, and nuts are among the foods most likely to provoke asthma in children. Skin tests are usually positive, indicating that IgE is implicated, but if the skin test is positive for a food in an asthmatic child, asthma may not follow a double blind challenge. In addition, the response to foods is sometimes psychosomatic.

The food preservatives sulphur dioxide (SO_2) and sodium metabisulphite can aggravate bronchospasm in established asthmatics. These patients are very sensitive to the irritant effect of SO_2 gas, which is liberated from sodium metabisulphite in acid foods and inhaled in low concentration as the food is swallowed.

Eczema

Infantile eczema and flexural eczema in adults are associated with high serum titres of IgE and often with multiple positive skin tests. In infantile eczema the response to food taking or elimination is slower and less clear cut than in urticaria.

Foods that can cause allergic reactions including anaphylaxis

- Cows' milk (esp. young children)
- Hens' eggs (esp. young children)
- Peanuts
- Tree nuts: brazil, almond, hazel
- Fish
- Shellfish
- *Also* wheat, legumes (peas, soya), fruits (citrus, strawberries, apples), vegetables (tomato, celery), sesame, etc

Peanuts top the list of foods responsible for anaphylaxis. Most fatal reactions have occurred from foods eaten outside the sufferer's home. Waiters, hosts, and food-pack labels cannot always be relied on. People with peanut allergy must be constantly vigilant with any unfamiliar food and consider carrying self-injectable adrenaline. In atopic families the lactating mother should not eat peanuts (or products containing peanuts) herself or give any to the child for at least three years.[9]

Foods likely to contain tartrazine (food colour E102) include:

- fruit squash and cordial
- coloured fizzy drinks
- pickles
- bottled sauces
- salad cream
- cakes (shop bought)
- cake mix
- soups (packets and tins)
- custard
- instant puddings
- coloured sweets
- filled chocolates
- jelly
- ice cream and lollies
- jam
- marmalade
- curry powder
- mustard
- yoghurt

Tartrazine is water soluble and gives a pleasant lemon yellow colour to foods. It is also used in some medicine capsules. Incidence of sensitivity appears to be between 1 in 10 000 and 1 in 1000.

Tartrazine is one of several permitted azo colours. It has been the most tested and incriminated in reactions, but some of the other colours (if eaten in comparable "dosage") might possibly cause reactions in the rare people who are sensitive to tartrazine.

Foods likely to contain sulphur dioxide

- Wines, chilled fruit juices
- Pickled onions, dried fruits
- Commercial precut chips
- Salads in salad bars*
- Fresh fruit salad in hotels*

*From "stay fresh" spray

Statistically significant responses to skin tests have been reported in infants apparently sensitive to a food—for example, exacerbation after milk or improvement after withdrawing egg. A controlled trial showed improvement in 14 out of 20 children with infantile eczema when egg and milk were removed. Elemental diets (glucose, oil, amino acid mixture, vitamins, and minerals), though expensive, have been helpful in severe infantile eczema. Breast feeding reduces the chance of eczema but only partly in babies with a strong atopic family history. In such cases the mother should avoid eating common allergenic foods.

In adults with eczema a response to few foods diets is less likely.

Migraine

Tension, relaxation, menstruation, bright lights, and hypoglycaemia are among the major precipitants of migraine. Foods can also precipitate attacks. The different factors can be cumulative; several may be needed before an attack occurs. Attacks may come on many hours after eating a provoking food. Suggestibility and placebo effect have been well established in the responses of migraine sufferers.

Chocolate, cheese, and citrus fruits are reported to precipitate migraine. They contain pressor amines—tyramine, phenylethylamine, and synephrine, respectively. Alcoholic drinks are another precipitant, notably related to cluster headaches: red wines and some other drinks contain histamine. Other attacks may come about from foods that produce nausea, such as fatty foods. Nitrates, found in some sausages, occasionally cause headaches ("hot dog headache").

True food allergy via IgE is not the usual mechanism in migraine. In a trial at Great Ormond Street, in children with severe recurrent migraine most recovered on an exclusion diet. Reintroduction of foods, first "open", later disguised, showed that cows' milk, eggs, chocolate, orange, and wheat were most likely to provoke an attack and tartrazine less commonly.

Foods reported to precipitate migraine

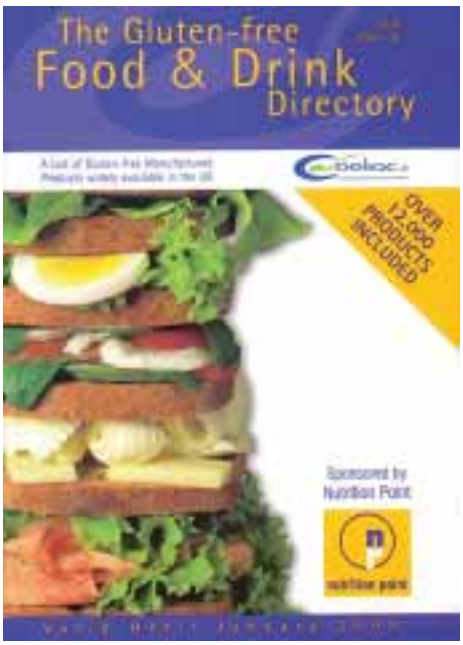
Gastrointestinal reactions

Many different gastrointestinal sensitivity reactions to food are known and they can act through several different mechanisms. Early symptoms are lip swelling, tingling in the mouth or throat, and vomiting. Later symptoms include diarrhoea, bloating, or even steatorrhoea. Remote symptoms—urticaria, asthma, headache, joint pains—can be associated.

In young children immediate intolerance is not uncommon to cows' milk or egg white, nuts, seafood, and some fruits. Tolerance increases with age. Cows' milk allergy can produce a variety of effects, including gastrointestinal bleeding or protein losing enteropathy; eosinophilia may be present.

Coeliac disease

Sensitivity to wheat gluten is the cause of coeliac disease with jejunal atrophy. It took from 1888, when coeliac disease was classically described, to 1953 before it was found that a fraction of wheat was responsible. A few other foods have been linked with occasional mucosal damage in children, such as cows' milk and soya. (See diets for gluten-sensitive enteropathy in chapter 13.)



The Gluten-free Food & Drink directory (reproduced with permission from Coeliac UK)

Infantile colic

When it occurs, this lasts for about the first three months. It occurs at least as often in breast fed as in bottle fed infants. In the former there is no consistent evidence that the colic is related to anything the mother eats.

Irritable bowel syndrome

Evidence for food sensitivity in the irritable bowel syndrome is conflicting. One study, which performed exclusion and double-

blind challenge, indicated that food sensitivity was common: wheat, dairy foods, maize, some fruits, tea, and coffee were mostly implicated. Another study could not show food sensitivity in most patients.

Intestinal lactase insufficiency
This is the rule in most adults of Asian and African origin and seen in a minority of white people. There is diarrhoea, abdominal distension, discomfort, and flatus after milk, but this usually has to be a cupful or more.

Attention-deficit/hyperactivity disorder
Feingold in the USA suggested that children (usually boys) with overactivity, short attention span, and impulsive behaviour might improve on a diet which omitted foods containing artificial colours or natural salicylates, or both. Although organisations of parents of difficult children have had faith in this hypothesis, objective confirmation is sparse. Most double-blind tests with food colours reported significant effects in either none or only one or two of the hyperactive children tested, and much of the information Feingold used on the salicylate content of foods was wrong. The salicylate content of most diets, by modern specific analysis, is only 1-2 mg/day.[11] Nevertheless, there are some children with a combination of overactivity and physical symptoms (rashes, rhinitis, headaches) suggestive of food sensitivity that have improved on an elimination diet, at Great Ormond Street. They appeared on challenge to be sensitive to different foods, but tartrazine and benzoic acid were top of the list.[12]

Arthritis
Gouty arthritis is aggravated by alcohol and by high purine and high protein diets. Although most patients with rheumatoid arthritis do not respond to food exclusions or challenges, a patient whose arthritis was clearly shown to be aggravated by milk and cheese was reported from the Hammersmith Hospital. A small minority of similarly reacting rheumatoid patients was found in a subsequent study in Florida.[13]

Pseudo food allergy[14]
Apparent reactions to food are quite often psychological rather than organic in origin. Patients may be convinced, but on circumstantial or no real evidence, that they are sensitive to certain foods. These beliefs are encouraged by some popular books and unorthodox practitioners, and are not easy for the busy general practitioner to change. Avoiding a broad range of foods carries a risk of malnutrition. Parents have also been reported as having inflicted supposed allergies on their children—a variant of Meadow's syndrome.

Specialist food "allergy" (sensitivity) clinics are available at some of the teaching hospitals, such as Addenbrooks, Cambridge; Great Ormond Street, London; Guy's, London, etc.

Food intolerance databanks have been set up in several European countries. In the United Kingdom the Leatherhead Food Research Association manages the databank.

Some less common food sensitivities

- *Favism*—haemolytic anaemia after eating broad beans. The basic defect is red blood cell glucose-6-phosphate dehydrogenase deficiency.
- *Bitter lemon purpura*—"bitter lemon" contains quinine, which may rarely precipitate thrombocytopenic purpura.
- *Chinese restaurant syndrome*—as first described, the symptoms were facial pressure, burning sensation in upper trunk and shoulders, and chest pain soon after eating foods rich in monosodium glutamate, particularly Chinese wonton soup. But delayed headache and nausea have subsequently been found to be more common, and asthma has sometimes been precipitated.
- *Sensitivity to tea or coffee*, or both—these common social beverages can cause a variety of pharmacological effects, for example, vomiting, headaches, tachycardia, which have sometimes been proved by objective tests.[6]

References

1 Young E, Stoneham MD, Petruckevitch A, Barton J, Rona R. A population study of food intolerance. *Lancet* 1994; **343:** 1127-30.
2 Moneret-Vautrin DA. Food allergy and intolerance in infancy. In: Walker AF, Rolls BA (eds). *Infant Nutrition*. London: Chapman & Hall, 1994.
3 Bindslev-Jensen C. ABC of allergies: food allergy. *BMJ* 1998; **316:** 1299-302.
4 Lessof MH. *Food allergy and other adverse reactions to food*. Brussels: ILSI Europe, 1994.
5 Royal College of Physicians and the British Nutrition Foundation. Food intolerance and food aversion. *J R Coll Phys Lond* 1984; **18:** 2-41.
6 American Academy of Allergy and Immunology Committee on Adverse Reactions to Foods. National Institute of Allergy and Infectious Diseases. *Adverse reactions to foods*. (NIH Publication No 84-2442.) Bethesda, MD: National Institutes of Health, 1984.
7 Lessof MH. Reactions to food in adults. In: Lessof MH (ed). *Clinical reactions to food*. Chichester: John Wiley, 1983.
8 Ewan PW. ABC of allergies: anaphylaxis. *BMJ* 1998; **316:** 1442-5.
9 Sampson HA. Managing peanut allergy. *BMJ* 1996; **312:** 1050-1.
10 Greaves MW. Current concepts: chronic urticaria. *New Engl J Med* 1995; **332:** 1767-72.
11 Janssen PLTMK, Hollman PCH, Reichman E, Venema DP, van Staveren WA, Katan MB. Urinary salicylate excretion in subjects eating a variety of diets shows that amounts of bioavailable salicylates in foods are low. *Am J Clin Nutr* 1996; **64:** 743-7.
12 Egger J. Food allergy and the central nervous system. In: Reinhardt D, Schmidt E (eds). *Food Allergy, Nestle Nutrition Workshop Series*, vol 17. New York: Raven Press, 1988.
13 Panush RS. Nutritional therapy for rheumatic diseases. *Ann Intern Med* 1987; **106:** 619-21.
14 Pearson DJ. Pseudo food allergy. *BMJ* 1986; **292:** 221-2.

Further reading

Buttriss J. (ed) *Adverse Reactions to Food. The report of a British Nutrition foundation Task Force*. Oxford: Blackwell Science, 2002.

16 Processing food

In the guessing game, "Animal, vegetable, or mineral?" nearly all foods are animal or vegetable or the two combined. What we eat is other organisms, some of them—fruits and cereals—still viable. The basic material of our foods has an anatomy and a histology. It is made of cells which contain numerous enzymes and biochemical compounds that are the same as or similar to those in the human body. Foods are also inhabited by micro-organisms. These multiply after slaughter or harvesting and nearly all our basic foodstuffs deteriorate rapidly because of autolytic decay and microbial activity. Foods can become dangerous because of these unless we treat them in some way. Cereal grains are the major exception. They have a lower water content and normally keep for years when dry at ambient temperatures. The major early civilisations were built on this property, and even today the only staple foods that are kept in savings banks are the cereals. Insect activity also leads to appreciable losses of food.

Doctors need to know about food

- "Doctors need to know as much about foods as they do about drugs."
- "Each medical school should have an expert on the biochemistry of foods among their preclinical lecturers."

Two recommendations from the International Union of Nutritional Sciences Workshop on Nutrition Education for Medical Practitioners, held at the Royal College of Physicians, London, March 1984.[1]

Why foods have to be processed or prepared

(1) Foods are preserved so that they can be kept edible for longer. Preservation reduces wastage and hence cost. It enables people in affluent communities to eat their favourite foods all the year round and also enables us to benefit from economies of scale by growing large quantities of a food on large areas of suitable land. Preservation helps to feed the growing world population.

(2) Processing or preparation makes foods safe to eat by destroying or retarding the growth of pathogenic micro-organisms (such as salmonella, clostridia, or fungi) or inactivating natural toxins (such as trypsin inhibitors and goitrogens).

(3) Processing or preparation improves the attractiveness of food, its flavour, and its appearance. This is not a frivolous function. Poorly prepared food is often the chief complaint of patients in hospitals and nursing homes.

(4) Processing also provides convenience. In Mrs Beaton's time most women spent their lives working for hours each day in the kitchen. Modern convenience foods have liberated women in countries like Britain. Much of the work that used to be done by hand in the Victorian kitchen is now done by machines, some controlled by computer, in the food processing plant or factory.

Food technology

- **Handling after harvest**—Foods must be kept at the best temperature, atmosphere, etc, between the farm and market or factory.
- **Food processing** is treatment of food in a plant or factory before it is sold in the shop.
- **Food preparation** is treatment of food in the kitchen, at home, in a catering establishment or take-away shop. It is a wider term than "cooking", which implies the use of heat.
- Between each treatment stage food has to be **stored**; the conditions affect how long it keeps and its quality.

Methods of food preservation and processing

Drying

Traditionally this was sun drying or smoking, but nowadays tunnel drying, spray drying, and freeze drying produce concentrated forms of the foods—for example, milk powder or instant coffee powder. Bacteria, which require water, cannot grow and autolytic enzymes are inhibited.

Freezing

This prevents bacterial growth because bacterial enzyme activity slows then stops as temperature is lowered. In addition, water in the food is not in an available form.

Refrigeration
Does not destroy micro-organisms but ensures that those present cannot multiply or do so only slowly; it also slows autolysis by enzymes in the food.

Addition of salt or sugar
This is a third way of lowering the "water activity" by increasing osmotic pressure and preventing bacterial growth, just as sugar has been shown to reduce infection in wounds.

Acidification
Acids also prevents microbial growth, for example, foods pickled in vinegar.

Heat
This is used in several ways. Blanching (1-8 min at 100°C depending on the food) before freezing and canning inactivates enzymes, that would otherwise continue autolysis of the food. Pasteurisation of milk (72°C for 15 seconds) destroys pathogenic organisms but not others. Cooking destroys all or nearly all organisms (except spore formers, depending on conditions). Sterilisation of canned and other sealed foods is performed by subjecting the food to a high temperature with or without pressure.

Food irradiation
A new technique, food irradiation, is useful for replacing ethylene oxide in sterilising spices (whose flavour would change on heating). It can also extend the shelf life of strawberries and mushrooms, inhibit sprouting of potatoes and onions, and destroy pathogens on poultry.[2]

Fermentation
This produces acid or ethanol, or both, which inhibit pathogenic and spoilage organisms.

Chemical preservatives
Benzoic acid, propionic acid, and sorbic acid, naturally present in cranberries, Gruyère cheese, and rowanberries, respectively, are added to certain specified foods in controlled amounts to prevent microbial growth. Sodium metabisulphite, which liberates sulphur dioxide in the food, is used for the same purpose.

Packaging
Once a food has been heat sterilised reinfection is prevented by sealing it in a can or airtight plastic bag or multilayer paper or plastic carton. Packaging of unsterilised food, though not preventing bacterial growth, reduces contamination and prevents loss of water through evaporation, etc.

Separation methods
Unlike the methods above these are not used to preserve foods. Milling produces different fractions of cereal flours. Pressing produces oils from oil seeds and juice from fruit.

Fresh or processed?
Most "fresh" fruit in the greengrocers comes from overseas or has been stored for weeks or months. "Fresh" vegetables too, though mostly home grown, are often stored. Bananas are picked in the tropics before they are ripe, shipped and stored at a controlled even temperature (>13°C), and ripened by exposure to ethylene gas (which is given off naturally by ripening fruit). Oranges are picked ripe, shipped, and stored at a lower temperature (3°C) in dry air with a raised carbon dioxide level. The skins have to be protected from mould

From the earliest civilisations

Some of the most important processes used for our foods go back to the earliest civilisations.

- The Sumerians had a simple dairy industry. Butter is mentioned several times in the Bible (for example, Isaiah vii, 15).
- The ancient Egyptians brewed beer and discovered how to make raised bread. Beer has been called "liquid bread".

Many processes are very old

- Wine was part of the way of life of the ancient Greeks. Wine and raisins (another very old food) are two different ways of preserving the energy of grapes in palatable forms.
- Salted pork—forerunner of our ham and bacon—was a common meat for the ancient Greeks.
- Cheese is mentioned in the Bible (for example, II Samuel xvii, 29). It is a remarkable method for preserving the energy, calcium, and other nutrients of milk.
- Jam, another way of preserving fruit, was being made long before the Industrial Revolution.

Nutrients per 100 g in wholemeal (100%) and white [breadmaking] (72%) wheat flour[3]

Nutrients	Wholemeal	White
Energy kJ (kcal)	1318 (314)	1451 (346)
Protein (g)	12.7	11.5
Fat (g)	2.2	1.4
Starch (g)	62	74
Sugars (g)	2.1	1.4
Dietary fibre (g)	8.6	3.7
Thiamin (mg)	0.47	0.32*
Riboflavin (mg)	0.09	0.03
Niacin (mg)	5.7	2.0*
Vitamin B-6 (mg)	0.5	0.15
Total folate (μg)	57	31
Vitamin E (mg)	1.4	0.3
Iron (mg)	3.9	2.1*
Zinc (mg)	2.9	0.9
Calcium (mg)	38	140*
Total phosphorus (mg)	320	120
Phytate phosphorus (mg)	240	30

*As fortified in the UK. *Note*: These are wheat **flours**. In bread nutrient contents are lower because over a third of the fresh weight is water

infection, for example, by treating with fungicides mixed with waxes after they have been washed. Although these fruits have been in artificial environments, they are intact and alive and their cells are absorbing oxygen and producing carbon dioxide.

Likewise, it is worth money to understand what factors determine tenderness and flavour in meat. There is a speciality of meat science, which shares some of the histological and biochemical knowledge used by clinicians who specialise in muscular diseases.[4] Before slaughter, animals should have adequate muscle glycogen. This is converted to lactic acid, which acts as a weak preservative. Fresh meat is tough because of rigor mortis. This disperses during hanging in a controlled chilled temperature of −1.4 °C and the meat is tenderised.

Food additives

Salt, vinegar, nitrates, and sugar have been used for centuries and are still among the most used food preservatives today. Hops are a traditional preservative for beer. Many of history's great voyages of exploration were made in search of food additives. Marco Polo journeyed to the East to obtain exotic spices. Cortez brought back vanilla from the Aztecs.

Food additive controls in Europe are now harmonised throughout the EU. The use of food additives is subject to strict legislative controls and they are only permitted if considered safe following scrutiny by independent experts, particularly the EC Scientific Committee for Food. To ensure that safe intake levels are not exceeded, many of the additives are only permitted in certain foods and up to specified maximum levels.

- **Preservatives** such as nitrates in bacon and ham and sulphur dioxide in dried fruit prolong the shelf life of foods by protecting them against deterioration caused by micro-organisms.
- **Antioxidants** are used to prevent the slow oxidation of oils and fats by atmospheric oxygen and development of rancidity.
- **Emulsifiers** keep oil and aqueous phase together in sauces. Lecithin is an example.
- **Humectants** prevent foods from drying out. Glycerol is an example.
- **Food acids** are acids that occur naturally. They are used for flavour or for technical reasons—for example, to adjust the pH in certain jams so that the pectin sets—and can also act as preservatives.
- **Anti-caking agents** stop lumps forming in powdery foods.
- **Thickeners** may be vegetable gums, cellulose derivatives, or starch derivatives.
- **Added nutrients** are now given by name on the label. A number of foods are fortified or enriched with nutrients—for example, margarine includes vitamins A and D, flour and bread and some breakfast cereals contain some B vitamins, and iron, and textured vegetable proteins contain vitamin B-12.
- **Miscellaneous**—Other additives include: flour treatment agents, firming agents, stabilisers, flavour enhancers, propellants, and glazing agents.
- **Colours**—Forty-three colours are permitted in the European community. More than half of them are natural, for example, beetroot red, chlorophyll and various carotenoids. Twenty of the permitted colours, such as some azo dyes and erythrosine (an iodine compound), are synthetic.
- **Flavourings**—Flavourings are added to foods in minute quantities and without hazard to public health. They are either natural flavour preparations (for example, essential oils) or chemically defined flavouring substances. These may be natural (isolated from products of plant or animal origin,

EC code number for food additives on labels (some examples)

Preservatives	
Benzoic acid	E210
Propionic acid	E280
Sorbic acid	E200
Sodium metabisulphite	E223
Antioxidants	
Butylated hydroxyanisole (BHA)	E320
Butylated hydroxytoluene (BHT)	E321
Propyl gallate	E310
Tocopherols	E306-309
Emulsifiers	
Monoglycerides and diglycerides of fatty acids	E471
Lecithins	E322
Humectants	
Glycerol	E422
Sorbitol	E420
Acids	
Acetic acid	E260
Citric acid	E330
Malic acid	E296
Anti-caking agents	
Calcium phosphate	E341
Magnesium carbonate	E504
Thickeners	
Guar gum	E412
Locust bean gum	E410
Pectin	E440
Carboxymethylcellulose	E466
Colours—natural	
Beetroot red	E162
Chlorophyll	E140
β-carotene	E160(a)
Colours—synthetic	
Tartrazine	E102
Brown FK	E154
Erythrosine	E127
Flavour enhancers	
Monosodium glutamate	E621
Sodium inosinate	E631

for example, menthol from peppermint oil) or nature-identical (made synthetically but chemically identical to a substance that occurs naturally, for example, citral) or artificial (not yet found in nature, for example, ethyl vanillin). There is no EU-wide flavouring legislation at present, but the EU is working towards creation of a list of permitted flavouring substances. The names of individual flavours are not declared on labels and they are not in the E code, but if flavourings are used this is mentioned (non-specifically) in the ingredients list. Flavour recipes in food are regarded as trade secrets.

- **Artificial sweeteners** permitted under EU law include saccharin, aspartame, and acesulfame K.

Contaminants

Unintentional additives can get into foods somewhere along the chain from farm to plate. Examples are pesticides, other farm chemicals, veterinary residues, drugs, heavy metals (lead, cadmium, mercury, and arsenic), industrial chemicals—for example, polycyclic aromatic hydrocarbons and polychlorinated biphenyls (PCBs), atmospheric and water pollutants, radionuclides, and phthalate plasticisers. Foods should be monitored for most of these by government agencies, such as, in Britain, the Food Standards Agency.

Foods and ingredients derived from gene technology[5-8]

Direct genetic manipulation of our food production is new and controversial. For centuries desirable genetic evolution was achieved by selective breeding of plants and animals used for food, mainly for external characters, recently for chemical composition. Rape plants were selectively bred in Canada '*to not*' express the characteristic (but potentially toxic) erucic acid in their seed oil—hence canola oil. With the traditional selective breeding many genes change one way or the other, along with the desired modification. With the new GM (genetically modified) technology a single gene from another species or an extra copy of an indigenous gene is inserted into the plant or animal, on a virus or bacterium, or is shot in on gold particles. The source of the new gene may be from a very different species; genes isolated from fish have been inserted into strawberries to give resistance to cold.

Gene technology is being developed—or already used—for several purposes:

- to make it easier to grow the food, for example, to introduce resistance to an endemic plant virus or to a particular herbicide that keeps the weeds down ("Roundup Ready" soyabeans)
- to make processing easier, for example, to reduce a spoilage enzyme and so prolong shelf life
- to make a food more palatable, for example, "Flavr Savr" tomatoes
- to produce a desired change in nutritional composition, for example, rice which contains (pro-vitamin A) β-carotene in the grain.

A primary aim of regulatory authorities is to determine if there can be any difference between, say, GM maize and conventional maize as food that could affect health—loss of nutrient(s) or appearance of a potential toxin or allergen. This is the concept of **substantial equivalence** and is the responsibility of food standards agencies. But the question of approving GM crops also has environmental and legal aspects. Not all GM crops are used for food (for example cotton, carnations).

Foods can be altered with gene technology in five ways:

- a chemically defined substance, obtained by gene technology, is used in production of the foods, for example, porcine somatotrophin (to improve the growth of pigs) or chymosin (replacing rennet in cheese making)
- less well defined ingredients in the food, for example, starch from insect-resistant maize, mycoprotein from genetically modified yeast
- foods/drinks produced using genetically modified organisms (GMOs), for example, wine or beer from yeast modified to result in an altered flavour profile
- transgenic plants or animals—that is, containing new or altered DNA, for example, soyabeans containing a gene that makes them resistant to a particular herbicide ("Roundup Ready" soyabeans)
- foods in which the genetically modified micro-organism is still present, for example, a "live" yoghurt with new properties.

Why all the fuss about GM foods?

- The process of GM technology is not natural. People object on religious grounds.
- Multinationals are controlling agribusiness vertically and globally. Though the inserted genes are natural, GM seeds are patented. Farmers have to contract with the seed company not to re-plant the seeds they harvest.
- GM crops will not help peasant farmers in developing countries who cannot afford to pay the premium and then not re-use seeds after harvest. Yields may not necessarily be better. The market price may be lower if the demand is for non-GM.
- The first generation of foods from GM crops has no advantage for consumers. Maize that is resistant to a herbicide does not taste any better, is not more nutritious and could possibly have some deleterious effect long term.
- Antibiotic resistant genes have often been used as markers for successful implantation of the new gene in seedling plants.
- There may be environmental damage, for example spread by pollen to weeds of the same botanical family, resulting in "superweeds", or "superbugs".
- GM food crops are based on a concept of fighting, rather than adapting to the ecology. We would be placing living things in an environment where there is no evolutionary history of how to accommodate them.

The growing of GM crops has to be considered and approved by a country's Department of the Environment. Here the **precautionary principle** should be used, and approval to grow a GM crop should start with experimental plots, well separated from conventional crops of the same botanical family.

The United States has been more permissive and the European Union has been and still is very cautious about approving GM foods. In the United States about 30% of the soyabeans now grown, and a lot of the maize, is from plants genetically modified to be resistant to the herbicide glyphosate ("Roundup"). These foods are not segregated or labelled in the wholesale markets; there is no way they can be identified by simple chemical analysis and they are likely to be present of course in soya and maize products being exported from north America. But these soyabeans and maize plants are not permitted to be grown in the European Union at present.

In the European Union multiple authorities, both national and in Brussels, both environmental and food authorities, have to scrutinise applications to grow GM plants. Consumers want any GM foods eventually approved, to be clearly labelled. Major supermarkets have a policy of not stocking GM foods and are giving increased shelf space to "organic" foods, a term which now covers non-GM.

In mid 2002 food aid maize shipped from the America to Zambia was impounded and not distributed to hungry people. The government feared that some might be be sown with the result that next year's maize could not be exported to Europe.

Scientific developments in plant breeding have made a huge leap. Regulators, politicians and consumers are in disarray. Meanwhile there is no present likelihood that doctors will be able to diagnose any disease caused by eating a GM food.

Potentially toxic substances in foods

Natural	
Inherent, naturally occurring	Usually present in the food and affect everyone if they eat enough, for example, solanine in potatoes
Toxin resulting from abnormal conditions of animal or plant used for food	For example, neurotoxic mussel poisoning; honey from bees feeding on rhododendron nectar
Consumer abnormally sensitive	For example, coeliac disease from wheat gluten; allergy to particular food; or drug induced, for example, cheese reaction
Contamination by pathogenic bacteria	Acute illness, usually gastrointestinal, for example, infection with *Salmonella* spp, or campylobacters or toxins produced by *Staphylococcus aureus* or *Clostridium botulinum* (food may not appear spoiled)
Mycotoxins	Food mouldy or spoiled, for example, aflatoxin B_1, from *Aspergillus flavus*, is a liver carcinogen
Manmade	
Unintentional additives: manmade chemicals used in agriculture and animal husbandry	For example, fungicides on grain, insecticides on fruit, antibiotics or hormones given to animals
Environmental pollution	For example, organic mercury, cadmium, polychlorinated biphenyls, and radioactive fallout can affect any stage of food chain
Intentional food additives: preservatives, emulsifiers, flavours, colours, etc	The most thoroughly tested and monitored of all chemicals in food

Are our foods safe?

Britain had the first food safety legislation in the world, the Sale of Food and Drugs Act 1875; this and subsequent acts have been replaced by the Food Act 1984. Food safety became the responsibility of the Food Standards Agency in 2000. Its expert committees are in regular informal communication with the EC Scientific Committee for food, the Joint FAO/WHO Expert Committee on Food Additives, the US Food and Drug Administration, and food toxicologists round the world.

Deliberate food additives are not intrinsically toxic substances. They have been tested in several animal species and are kept under review continuously by food toxicologists. Amounts permitted in foods are such that the maximum intake does not exceed the acceptable daily intake (usually 1/100 the highest level that has no effect in animal tests). Toxicological tests have not so far systematically examined the chances of hypersensitivity reactions in man, and such occasional reactions to tartrazine, sulphur dioxide, and monosodium glutamate are described in chapter 15.

Whereas some natural components in the diet are risk factors for certain types of cancer (chapter 3), the World Cancer Research Fund's expert panel[9] agrees with previous reviewers that intentional food additives do not cause cancer (when used in quantities prescribed by the regulations).

Testing food additives

- Acute toxicity is tested in male and female animals of at least three species.
- Distribution of the compound in the body and its metabolism are studied.
- Short-term feeding trials are done on at least two species of animal (one non-rodent).
- Long-term toxicity is assessed in at least one metabolically appropriate animal species.
- Reproduction studies involve giving the compound to experimental animals over at least two generations.
- Testing for mutagenicity (in bacteria) and carcinogenicity (in tissue culture) is undertaken.
- Observations in man are reviewed.

Losses of nutrients

Some losses of nutrients occur during food processing but they are qualitatively and quantitatively similar to the losses that happen in domestic cooking. Most processes in the food factory are scaled up versions of one or another home recipe. Factory processes are standardised and controlled, but home

cooking varies from excellent to bad. Nutrient losses are roughly predictable and can easily be measured by analysis at different stages.

Two vitamins are more unstable than the others when heated, vitamin C and folate, but whereas vitamin C lasts better in acid medium, folate does not. Thiamin (vitamin B_1) is moderately unstable when heated. Riboflavin decomposes in ultraviolet light. Water soluble vitamins dissolve into the cooking water and the more water used the more vitamins are likely to be wasted. Mineral nutrients are stable but can also be washed out if large amounts of cooking water are used. Lysine, the limiting amino acid in cereals, is the most unstable of the essential amino acids. The golden crust of bread is coloured by a complex of sugars and lysine which becomes biologically unavailable. There is some loss of linoleic acid in oils when they are reused for frying, especially at high temperatures.

Losses of **vitamin C** are worth considering in detail. The factors that cause the oxidative breakdown of vitamin C are tissue damage (which liberates ascorbic acid oxidase) by bruising or freezing of leafy vegetables, heating in alkaline water—for example, with sodium bicarbonate added—contact with copper, and leaching into the processing or cooking water. Moderate losses occur between harvesting and cooking fresh vegetables and when a bottle of fruit juice is opened and kept at room temperature. There is little difference in losses of vitamin C between these three methods of cooking: boiling, microwave, and pressure cooking, but the less water used the less vitamin is thrown away in the water. There are substantial losses of vitamin C when cooked vegetables are kept warm before they are served, or refrigerated until the next day.

Percentage retention of vitamin C in peas after different stages of preparation (after Mapson)[10]

Fresh	Frozen	Canned	Air dried
—	Blanching 75	Blanching 70	Blanching 75
—	Freezing 75	Canning 63	Drying 45
—	Thawing 71	Diffusion 40	—
Cooking 44	Cooking 39	Heating 36	Cooking 25

Effect of different conditions on stability of nutrients in foods (based on Harris and Karmas[11]**)**

	Effect of solutions			Effect of exposure to			Cooking losses
Nutrients	**Acid**	**Neutral**	**Alkaline**	**Oxygen**	**Light**	**Heat**	**(% range)**
Vitamins							
Vitamin A	U	S	S	U	U	U	0-40
Vitamin D		S	U	U	U	U	0-40
Vitamin E	S	S	S	U	U	U	0-55
Thiamin	S	U	U	U	S	U	0-80
Riboflavin	S	S	U	S	U	U	0-60
Niacin	S	S	S	S	S	S	0-50
Folate	U	U	S	U	U	U	0-80
Vitamin C	S	U	U	U	U	U	0-100
Amino acids							
Leucine, isoleucine, methionine, valine and phenylalanine	S	S	S	S	S	S	0-10
Lysine	S	S	S	S	S	U	0-40
Tryptophan	U	S	S	S	U	S	0-15
Threonine	U	S	U	S	S	U	0-20
Mineral salts	S	S	S	S	S	S	0-3

U = unstable, S = stable

On average, losses of vitamin C in cooking may be taken as 70%—that is, 30% retention—in leafy vegetables and 40% in root vegetables. Food tables usually give values for cooked vegetables, as well as for the raw food (see table opposite).

Vitamin C content in raw and cooked peas and mangetout[3]

	mg/100 g
Peas	
raw	24
boiled	16
frozen, boiled	12
canned, reheated	1
Mangetout peas	
raw	54
boiled	28
stir fried in oil	51

Perspective

- Some loss of nutrients is inevitable in food processing, but for most nutrients losses are small.
- Manufacturing losses, when they occur, are often in place of similar losses through cooking at home.

- The importance of the losses in a particular food has to be considered in relation to the whole diet. If a food makes only a small contribution to the intake of nutrients, processing losses are not of practical importance. On the other hand, changes in any food that makes a major contribution to nutrient supply—for example, milk for babies and cereals in some adults—need continued vigilance.
- There are some beneficial effects of processing or preparation: destruction of trypsin inhibitor in legumes and liberation of bound niacin in cereals. Nutrient enrichment is possible.
- Other advantages of food processing are protection from pathogenic organisms, better flavour, and cheaper price. Often the ultimate choice is between dried, canned, or frozen peas (say) in late winter or no peas at all.

Geography of food processing

Last century cows were kept in towns because there was no way of preventing milk from souring.

To provide the milk, cheese, yoghurt, and cream for the people of London today about 0.5 million cows are needed. They in turn each need about 7 acres of farm land to feed them through the year. From these 3.5 million acres scattered across the south of England fresh milk is pooled, transported, pasteurised, bottled, and distributed or processed in other ways. Other foods—fruits, meat, fish in Britain—may come from half way around the world. This complex movement and distribution of foods would be impossible without food processing.

References

1 Truswell AS (rapporteur). Nutrition education for medical students and practitioners: report of a workshop. *UN University Food Nutr Bull* 1984; **6**: 75-81.
2 *Food irradiation. A technique for preserving and improving the safety of food*. Published by WHO in collaboration with FAO. Geneva: WHO, 1988.
3 Holland B, Welch AA, Unwin ID, Buss DH, Paul AA, Southgate DAT. *McCance & Widdowson's the composition of foods*, 5th edn. Cambridge: Royal Society of Chemistry, 1991.
4 Lawrie RA. *Meat science*, 6th edn. Cambridge: Woodhead, 1998.
5 Dixon B. The paradoxes of genetically modified foods. *BMJ* 1999; **318**: 547-8.
6 Horton R. Genetically modified foods: "absurd" concern or welcome dialogue? *Lancet* 1999; **354**: 1314-15.
7 Gaskell G, Bauer MW, Durant J, Allum NC. Worlds apart? The reception of genetically modified foods in Europe and the US. *Science* 1999; **285**: 384-7.
8 The grim reaper. *The Economist* 2002; Aug 24: 42-43.
9 World Cancer Research Fund in association with American Institute for Cancer Research. *Food, nutrition and the prevention of cancer: a global perspective*. Washington DC: American Institute for Cancer Research, 1997.
10 Mapson LW. Effect of processing on the vitamin content of foods. *Br Med Bull* 1956; **12**: 73-7.
11 Harris RS, Karmas E. *Nutritional evaluation of food processing*, 2nd edn. Westport, CT: Avi, 1975.

17 Nutritional support

Nigel Reynolds, Christopher R Pennington

Nutritional support is required for the prevention of starvation and the treatment of malnutrition in patients who are unable to ingest or absorb sufficient nutrients. There is controversial evidence that the provision of specific nutrient substrates may modify the response to disease. This chapter will review current concepts with respect to the need for nutritional support, the route of nutrient delivery, the prescription of nutrient substances, and the quality of nutritional management.

The problem of malnutrition

Studies have demonstrated that malnutrition is common in hospital patients although often it goes unrecognised. Unless nutritional management is initiated nutritional status declines in most patients during hospital stay, particularly in those who are malnourished on admission. Furthermore, nutritional depletion has been shown to continue for two months after discharge in malnourished postoperative patients.

Malnutrition is defined in terms of tissue wasting which may arise through starvation and the metabolic action of cytokines generated in response to tissue injury. There is evidence that nutrient deprivation is of major importance in most patients through a reluctance or inability to eat, the lack of available food for prolonged periods in hospital, and impaired intestinal function. Common causes of malnutrition are given alongside.

Some causes of malnutrition in hospital patients

Anorexia
- Depression
- Chronic disease

Inability to eat
- Neurological disorders
- Swallowing disorders

Intestinal disease
- Inflammatory bowel disease
- Radiation enteritis
- Gluten enteropathy
- Short bowel syndrome
- Hollow visceral myopathy

Inflammatory response to infection

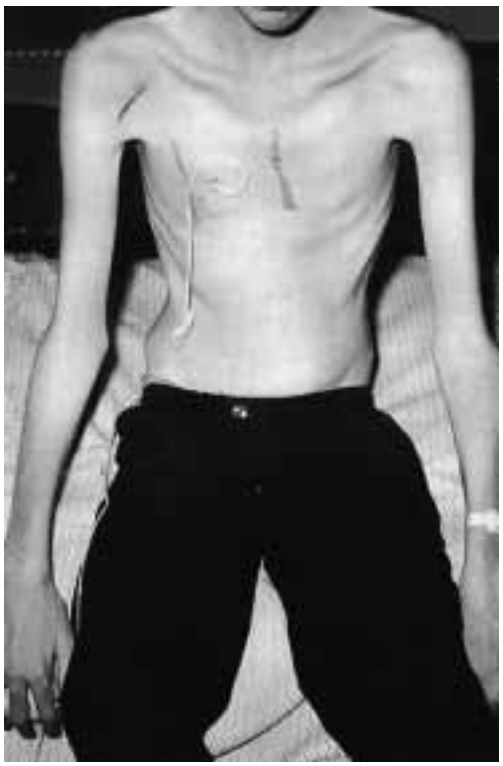
Studies of starvation in normal healthy subjects reveal that death is likely from 60 days when weight loss in excess of 30% has been sustained. The time before tissue loss becomes irreversible is significantly reduced in hospital patients who have already lost weight before admission and in whom the process is accelerated in the presence of inflammatory mediators. Cytokines lead to proteolysis and lipolysis combined with the suppression of appetite.

There are other reasons for the early recognition of malnutrition and where appropriate the provision of nutritional support. When patients are metabolically stressed through infection or trauma, tissue wasting can only be retarded not reversed. Data from the Minnesota experiment and from re-feeding patients with anorexia nervosa have demonstrated that a very long time is required for tissue

Malnutrition in hospital patients

Study*	Patients (no)	Type of patients	% Malnourished
Prevalence of malnutrition in hospital patients			
Bistrian *et al.* 1974	131	General surgical	50.0
Bistrian *et al.* 1976	251	General medical	44.0
Hill *et al.* 1977	105	General surgical	50.0
Incidence of malnutrition on admission to hospital			
Willard *et al.* 1980	200	General medical General surgical	31.5
Bastow *et al.* 1983	744	Orthopaedic surgical	52.8
Zador and Truswell 1989	84	General surgical	14.0
Larsson *et al.* 1990	501	Care of elderly	28.5
Cederholm *et al.* 1993	200	General medical	20.0
McWhirter and Pennington 1994	500	General medical General surgical Orthopaedic surgical Respiratory medicine Care of elderly	40.0
Giner *et al.* 1996	129	Intensive care	43

* (For details of the studies see Pennington 1998)[1]



A malnourished patient

repletion. Of more importance is the fact that organ function is impaired by starvation, long before current indices of nutritional status become abnormal. Thus obese subjects who starved for two weeks demonstrated loss of muscle power and increased muscle fatiguability of comparable magnitude to malnourished hospital patients. This has implications for the mobilisation of patients after surgery and illness. Furthermore respiratory muscles are also affected, adding to the risk of clinical complications of malnutrition. Other clinically important consequences of nutritional depletion include impairment of wound healing, and of immune response and digestive function.

Some of the effects of malnutrition

Impaired mental function
- Apathy, fatigue, poor cognition

Impaired muscle function
- Respiratory failure
- Delayed mobilisation

Impaired immune function
- Increased incidence of infection

Miscellaneous
- Impaired thermogenic response
- Impaired wound healing

The recognition of malnutrition

There is no single clinical method that will reliably diagnose malnutrition. The measurement of weight and height is used to calculate the Body Mass Index (BMI) by the formula: weight (kg) divided by the height^2 (m). The normal range is 20-25. Adult patients with values of 19 or below are malnourished. Information from the BMI may be supplemented by measuring the mid-arm muscle circumference and triceps skinfold thickness which respectively correlate with protein and fat stores. Values are compared with standard reference ranges for the patient population (see page 72).

There are significant limitations in the interpretation of these measurements. They are affected by changes in the hydration status and inter-observer error, and the patient may suffer from the effects of starvation due to impaired organ function long before such measurements become abnormal. Some clinicians employ hand grip dynamometry as a method of detecting nutritional influences on muscle power. All these measurements are time consuming. Nutritional screening of all patients admitted to hospital has been recommended as a simple method of detecting patients at nutritional risk who may merit further investigation. Such a scheme is summarised in the box opposite.

Nutritional screening*

Questions
- Reduced food consumption
- Unintentional weight loss

Measurements
- Weight
- Height

*Reproduced from Lennard-Jones *et al.* 1995[2]

The role of nutritional support

Nutritional support is clearly indicated in patients who are unable to eat or who have prolonged intestinal failure. Under these circumstances it is required to prevent death from starvation. When used to treat malnutrition, nutritional support will reduce morbidity in many patient groups and some of the studies which demonstrated benefit are summarised in the table opposite. There are other studies in which no benefit or increased morbidity has been observed with nutritional treatment. Many of these studies were characterised by the inappropriate use of nutritional support and in particular parenteral nutrition with excessive substrate administration.

Nutritional support may be administered in the form of oral supplements, enteral tube feeding or parenteral (intravenous) nutrition. Estimates suggest that 3-4% of hospital beds are occupied by patients who are receiving nutritional support by parenteral or enteral tube feeding. Approximately four times the number of patients are tube fed compared to those who receive parenteral nutrition.

Randomised controlled studies which examine the influence of nutritional support on the length of hospital stay (LOS) (Adapted from Booth and Morgan 1995[3])

Study*	Patient group	Nutritional management	Reduction of LOS days
Bastow *et al.* 1983	122 elderly females with fracture of neck of femur	Nocturnal nasogastric supplementary EN	9 (in very thin group)
Askanazi *et al.* 1986	35 radical cystectomy	Postoperative PN	7
Delmi *et al.* 1990	59 elderly patients with fracture of neck of femur	Oral supplements postoperatively	16
Rana *et al.* 1992	40 patients undergoing moderate or major abdominal surgery	Oral supplements postoperatively	3.3
Eisenberg *et al.* 1993	459, 86% general surgical, 14% general medical	Preoperative PN	0
Mac Burney *et al.* 1994	43 bone-marrow transplant patients	Glutamine-supplemented PN	7
Bower *et al.* 1995	368 intensive care patients	Early EN with formula supplemented with argenine, nucleosides, and fish oil	8 (in patients who tolerated at least 821 ml per day)
Keele *et al.* 1997	100 patients following moderate or major abdominal surgery	Postoperative supplements	0

*For details of the studies see Booth and Morgan 1995[3] and Pennington[1]
EN enteral nutrition; PN parenteral nutrition

Enteral nutrition

Oral supplements

Oral supplements containing the recommended provision of micronutrients (vitamins and trace elements) should be used when it is anticipated that they will provide a large part of the diet. These preparations include a range of flavours and when

used between meals they do not significantly reduce the consumption of food. Oral supplements augment dietary intake in patients who experience difficulty in taking an adequate diet. They are unsuitable when patients are profoundly anorexic or suffer from swallowing disorders.

Enteral tube feeding

Enteral nutrition delivered by tube is cheaper, safer, and more physiological than parenteral nutrition. In particular, in common with oral nutrition, it stimulates intestinal and biliary motility and provides a greater range of nutrients. These include glutamine and short chain fatty acids, important substrates for the enterocyte and colonocyte respectively, yet which are not part of routine parenteral nutrition prescriptions because of potential problems with the stability of the solution. Enteral nutrition may have a role in the protection of the mucosal barrier function in the ill patient, although studies on the prevention of intestinal translocation of micro-organisms by enteral nutrition in the animal model have not been replicated in the human. Some of the indications for enteral tube feeding are summarised in the box opposite. The ability to infuse nutrients over prolonged periods can fully exploit residual intestinal function in patients with intestinal impairment.

The methods of delivering enteral tube feeding are given in table opposite. Post-pyloric placement is required for patients with gastric stasis, notably in the postoperative period in the critically ill patient. Percutaneous tubes are needed when prolonged treatment is envisaged, notably in patients with swallowing disorders due to chronic neurological disease. This approach is also useful in the younger patient with cystic fibrosis. These younger patients may prefer to have a conventional gastrostomy replaced with a button gastrostomy giving a more acceptable cosmetic appearance.

Enteral feeding solutions most commonly used are based on whole protein substrates and are termed polymeric. Occasionally feeds with a low sodium content are useful in patients with liver cirrhosis and some patients with intestinal failure may benefit from a peptide feed in which much of the lipid is in the form of medium chain triglycerides. Feeds supplemented with glutamine, arginine, nucleotides, and fish oils have been formulated for nutritional support in the critically ill patient. So far there is no convincing clinical evidence to support the use of these more expensive products. However, polymeric feeds containing fibre are useful for the regulation of bowel function in patients who are dependent on artificial nutritional support, particularly elderly patients with chronic neurological disease.

The complications of enteral feeding may be considered in three groups:

(1) Metabolic complications include disorders of glucose and electrolyte balance. Very malnourished patients are prone to a condition termed the **re-feeding syndrome**. During starvation the body adapts to use less carbohydrate and more fat metabolism such that metabolic tolerance of carbohydrates can be impaired. With the introduction of artificial nutritional support there is rapid intracellular passage of phosphate, magnesium, and potassium resulting in low serum concentrations. Thiamin depletion can be dangerous in this situation, so the vitamin should be routinely provided. There are potentially serious life-threatening effects on cardiac, bone marrow, brain, and respiratory function. Intravenous replacement of electrolytes may be required.
(2) Gastrointestinal symptoms are common. Diarrhoea is often associated with the use of antibiotics which suppress the

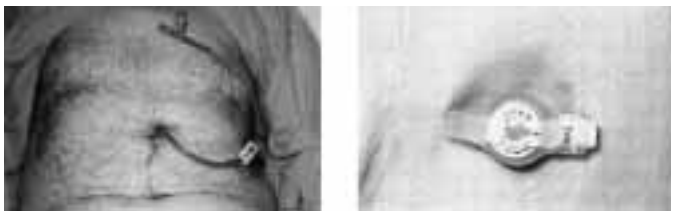
Some indications for enteral tube feeding

Problem	Examples
Anorexia	Cirrhosis, Crohn's disease, some forms of malignancy
Swallowing disorders	Cerebrovascular disease, motor neurone disease, oesophageal stricture
Gastric stasis	Postoperative patient, intensive care patient
Intestinal malfunction	Crohn's disease, cystic fibrosis

Examples of common methods of enteral tube feeding

Route	Placement	Comment
Nasogastric	Nurse or patient	Easy access Commonly displaced Suitable for short-term or intermittent feeding
Nasojejunal	Surgeon at operation Endoscopist	Useful in patients with gastric stasis, postoperative or ITU Readily displaced
Percutaneous gastrostomy	Endoscopist Radiologist using fluoroscopy Surgeon laparoscopically	More suitable for prolonged feeding Relatively safe Can be converted to a button gastrostomy
Percutaneous jejunostomy	Surgeon at operation Endoscopist	Alternative to parenteral nutrition for some postoperative patients Significant short- and long-term morbidity

(Reproduced from Lennard-Jones *et al.* 1995[2])



(*a*) (*b*)

(*a*) A gastrostomy tube, (*b*) a button gastrostomy. (Reproduced from Pennington CR, 1998[4] by kind permission of The Medicine Publishing Company)

activity of colonic bacteria, thus reducing the availability of short-chain fatty acids from fibre. Short-chain fatty acids are an important fuel for the colonocyte; they stimulate sodium and water transport across the colon. The problem of diarrhoea can be reduced with fibre-containing feeds and possibly by post-pyloric feeding.

(3) Complications of nutrient delivery include pneumonia as a result of aspiration or displacement of the gastric tube. Stomal infection is common in patients with percutaneous feeding tubes; peritonitis can occur when the stomach or jejunum are not opposed to the abdominal wall at the time of percutaneous tube insertion.

Parenteral nutrition

Parenteral nutrition is needed when the intestinal tract is unavailable or intestinal function is inadequate. Examples of some potential indications for short-term and prolonged parenteral nutrition are given in the box opposite.

The nutrient solution is compounded in a large multilayer bag under sterile conditions in the pharmacy, or provided as a standard solution available commercially. The solution contains glucose, lipid, amino acids, electrolytes, minerals, vitamins, and trace elements. Typically the volume ranges from 2 to 3 litres, the non-protein energy provision is 20-40 kcal per kg and the nitrogen provision 0.2-0.3 g per kg. These solutions contain all the essential amino acids, but not all the non-essential amino acids are included because of potential problems with the stability of the lipid solution. There is evidence that under some circumstances some of these "non-essential" amino acids are required. Thus in the stressed patient glutamine is considered to be conditionally essential. There is debate about the potential benefit of adding glutamine, in the form of dipeptides, to the parenteral nutrition prescription in critically ill patients to improve immune and gut barrier function. Furthermore as conventional fat solutions contain n−6 fatty acids that promote the formation of pro-inflammatory cytokines, there may be theoretical merit in deploying structured lipids which contain n−3 fatty acids. More clinical evidence is needed to support the use of these newer substrates.

In hospital parenteral nutrition is commonly needed for only two weeks, when it is conveniently administered by finebore peripheral cannulae. Peripheral cannulae have the advantage of avoiding the risks of central vein cannulation but have the drawback that it may be difficult to meet predicted energy and protein needs using tolerable volumes of fluid and lipid. The hypertonic nature of nutrient solutions can lead to phlebitis and loss of venous access. Critically ill patients and patients who need prolonged parenteral nutrition require central venous access. Venous access for parenteral nutrition is summarised in the box opposite.

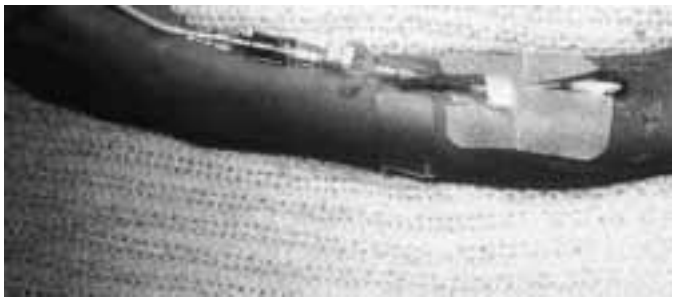
Parenteral nutrition is infused continuously in the stressed unstable patient. Cyclical feeding, with overnight infusion and a heparin lock during the day, has metabolic advantages in the majority of patients as mobilisation can be facilitated. The complications of parenteral nutrition are summarised in the box on page 124. Catheter-related infection can be prevented with appropriate protocols dictating catheter care and guiding the training of staff and patients. The hepatobiliary complications which accompany prolonged treatment in some patients may reflect the lack of enteral nutrition, with the formation of biliary sludge or the excess administration of glucose or other nutrients. There is good evidence that nutritional care of a high standard can be most cost-effectively established by using multidisciplinary nutritional support teams.

Some indications for parenteral nutrition

Short-term
- Severe inflammatory bowel disease
- Mucositis following chemotherapy
- Patients with multiorgan failure
- Major surgery
- Severe acute pancreatitis

Long-term
- Inflammatory disease
  Crohn's disease, radiation enteritis
- Motility disorders
  Hollow visceral myopathy, scleroderma
- Short bowel syndrome
  Mesenteric infarction, Crohn's disease

Parenteral nutrition solution compounded in the pharmacy



Peripheral parenteral nutrition catheter

Venous access for parenteral nutrition

• Peripheral parenteral nutrition	Venflon Ultrafine 15 cm cannula
• Peripherally inserted central catheter	PICC Line
• Central parenteral nutrition	Non-cuffed catheter with detachable hub Cuffed Broviac type of catheter Subcutaneous vascular port

Some complications of parenteral nutrition

Nutritional and metabolic
- Problems of glucose homoeostasis
- Electrolyte imbalance
- Micronutrient deficiencies
- Micronutrient excess, for example manganese

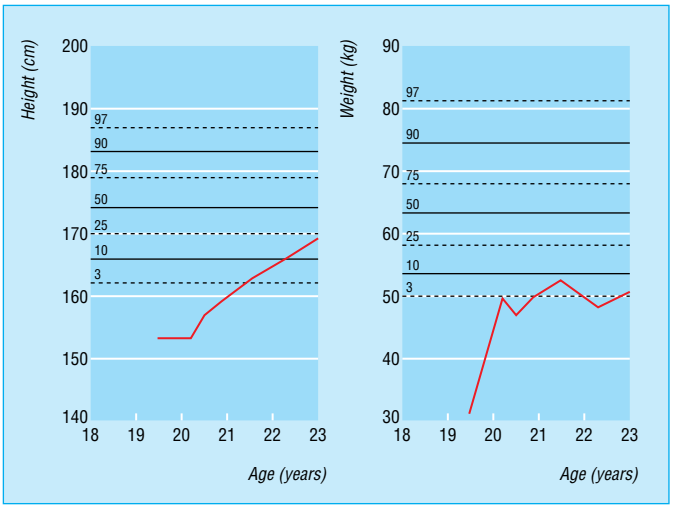
Catheter-related
- Infection: exit site, tunnel, lumen
- Occlusion: lipid, fibrin
- Central vein thrombosis
- Fracture of catheter

Effect on other organ systems
- Liver disease
- Biliary disease
- Osteoporosis

References

1 Pennington CR. Disease associated with malnutrition in the year 2000. *Postgrad Med J* 1998; **74**: 65-71.
2 Lennard-Jones JE, Arrowsmith H, Davison C, Denham AF, Micklewright A. Screening by nurses and junior doctors to detect malnutrition when patients are first assessed in hospital. *Clin Nutr* 1995; **14**: 336-40.
3 Booth K, Morgan S. *Financial issues for clinical nutrition in NHS hospitals.* Lancaster: Nutricia Clinical Care, Lancaster University 1995, pp 9-10.
4 Pennington CR. Artificial nutrition. *Medicine* 1998: **26**; 22–6.

The author, Christopher Pennington, sadly passed away in 2002.



Growth charts demonstrating belated growth between the ages of 20 and 23 years. (Reproduced from Pennington CR, 1998[4] by kind permission of The Medicine Publishing Company)

Further reading

Elia M. Changing concepts of nutrient requirements in disease: implications for artificial nutritional support. *Lancet* 1995; **345**: 1279-84.
McWhirter JP, Pennington CR. The incidence and recognition of malnutrition in hospital. *BMJ* 1994; **308**: 945-8.

18 Some principles

There are two questions affecting health about any food.

(1) Is it safe, or will it harm me (*a*) immediately or (*b*) later if I eat it repeatedly?
(2) Is it good for me?

Is it "food"?

For a food one has not eaten before question 1(*a*) predominates. If it has not been contaminated or infected the answer depends ultimately on folklore. In every culture there are parts of plants and animals that the group recognises as food but other cultures do not. Only a minority of plants can be expected to be freely edible. For most plants it would be an evolutionary advantage to possess a toxin that discourages animals from eating it. Our folklore about which plants are edible comes down from unknown ancestors who took the risk of eating an unfamiliar plant, sometimes with unfortunate results.

Is this food good for me?

Simple trial and error by people with primitive technology cannot answer questions 1(*b*) or 2. One of the difficulties for professionals who give advice about healthy diets is that there is no immediate symptom of well-being corresponding to the surge of amino acids or vitamins that blood samples can show. The feelings of satiety and of inner warmth after a meal are much the same after a good nutritious one as after a meal that contains only "empty calories". One rare exception is the gratifying faecal results that occur within hours of eating wheat bran in people inclined to constipation. This is probably why the fibre hypothesis was accepted by lay people years before it was well supported by scientific human experiments. The only reliable way to answer questions 1(*b*) and 2 is by the methods of nutritional science.

Origins of our scientific knowledge about human nutrition

Comparative and evolutionary

Homo sapiens and their predecessors have been on the earth one million years or more. Ninety-nine per cent of this time our ancestors lived as hunter-gatherers. Agriculture started only 10 000 years ago. There has not been enough time for our species to evolve new metabolic mechanisms required by the recent food supply. Natural selection, which must work chiefly via reproductive success, has been distorted by inequality of wealth and lately by technology. It is difficult to see how it could modify diseases that start in middle age. But presumably our bodies have evolved well-adapted for doing what hunter-gatherers did and eating what they ate. We have information from archaeological records and from studies of the few, fast disappearing groups of contemporary hunter-gatherers.[1,2]

Hunter-gatherers were lean. Some groups ate more plant than animal foods; others (especially in the cold northern winters) ate mostly meat (not only the muscle) or seafoods. They ate a large variety of foods, depending on the season but had no salt or alcohol and concentrated sugar only rarely (as wild honey) and only occasionally cereal. The photograph shows hunters about to set out, !Kung bushmen in the northern Kalahari, Botswana (taken by the author in 1968[1])

Experiments of nature and travellers' tales

From people who eat different foods from us, under stable conditions or during a disaster, we can form hypotheses about the physiological effects of different food patterns that we could not easily persuade our fellow countrymen to adopt.

We have, for example, learnt about the physiological role of ω-3 polyunsaturated fatty acids from the Eskimos,[3] and about deficiency diseases from nutritional experiences of prisoners of war.[4]

Epidemiological studies
These studies range in the power of their design. Associations and correlations of disease characteristics and dietary variables do not prove cause and effect, but prospective studies, especially if repeated in different groups, give valuable information on the relation between usual diets and chronic diseases.[5]

Animal experiments
Animal experiments were the principal technique for working out the vitamins.[6] The right animal model has to be used. Understanding of scurvy was static and controversial until Norwegian workers found (in 1910) that guinea pigs are susceptible like man because, unlike most animals, they cannot synthesise ascorbic acid from glucose.

Clinical records
Clinical records have been informative about the role of diet in disease, including inborn errors of metabolism. Information about requirements for trace elements has come from experiences with total parenteral nutrition.[7]

Food analysis
The independent variables in nutritional epidemiology and in dietetic treatment of disease are food constituents. Food analysis is work that is never finished; foods keep changing and demand develops for constituents not measured before, such as different types of fatty acids and potentially protective phytochemicals. To facilitate international sharing of what food composition data there is INFOODS (the International Network of Food Data Systems) set up in 1983.

Human experiments and trials
These last from hours to years and many different variables can be measured.

Evidence-based nutrition advice[8]
Official dietary guidelines and (if permitted) health claims on foods should be judged on the best available evidence. For judging the efficacy of drugs the best evidence is a meta-analysis or systematic review of all randomised controlled trials (RCTs) of the effect of drug versus placebo on disease outcome. These are paid for by pharmaceutical companies as part of the cost of developing new drugs. For nutrition RCTs with disease outcome are scarce. Available evidence may be epidemiological—cohort/prospective studies are more reliable than case-control or ecological studies. Or they may be short-term controlled trials with a physiological variable as outcome, for example plasma lipids or blood pressure. The evidence, say about vegetables and health, will never consist mostly of RCTs. Emphasis instead has to be on **all** the evidence, including animal studies and molecular biology and critical interpretation of the observational epidemiology.

Some examples of human experiments and trials

- Intervention trial of low saturated fat diet in half of 850 middle-aged male veterans in Los Angeles over five years
- Trials of vitamin C against placebo for preventing colds during winter
- Experimental depletion of a single nutrient in human volunteers
- Long-term testing of the value of novel protein foods
- Experiments measuring energy expenditure
- Metabolic studies—for example, to assess the effect of diet on plasma cholesterol
- Absorption and uptake studies—for example, glycaemic index after different foods containing carbohydrates

The three groups of substances in the edible portion of foods

Energy and nutrients
Water and packing
Other substances
Colour, flavouring, etc.
Natural non-nutritive substances, some of which appear to be protective, some of which are potentially toxic

The three groups of substances in foods

Energy and nutrients
Man needs oxygen, water and enough food energy (calories), 9 or more indispensable amino acids in proteins, essential fatty acids (ω-6 and ω-3 polyunsaturated) a small amount of

carbohydrate, 13 vitamins, and 17 elements scattered across the upper half of the periodic table (in addition to hydrogen, carbon, nitrogen, and oxygen: see figure on page 126).

Together they add up to over 40 nutrients, many of which are normally taken for granted; the minor nutrients are present in sufficient amounts in a diet of mixed foods. But for long-term total parenteral nutrition all the minor vitamins and trace elements must be included in the required postabsorptive amounts.

For some of the nutrients **you can have too much of a good thing**. Generous intakes of saturated fat raise the plasma cholesterol concentration and contribute to coronary heart disease. People with high salt intakes have more hypertension. Too much food energy leads to obesity.

Water and packing

All foods contain water. In many it is more than half the weight. The percentage of water is higher in some fruits and vegetables than in milk. The more water a food contains, the fewer calories. But this water has to be counted in the diet of patients with anuria. The "packing" of plant foods—that is, dietary fibre—is not all inert. Some fractions have physiological effects: arabinoxylans (hemicelluloses) of wheat increase faecal bulk and speed colonic transit; pectins slow absorption of lipids and glucose.

All the rest

There are many other substances in most foods. They include flavours and colours.

Potentially beneficial substances

It has long been noticed that higher intakes of vegetables and fruits are associated with lower rates of chronic degenerative diseases.[10] In the 1970s this was attributed to fibre or β-carotene or vitamin C. But in the 1990s it looked as if other bioactive substances that are not among the classical nutrients might also be protective. Some of these have antioxidant activity, and antioxidants in food and drink have attracted research interest since publication of the oxidised LDL hypothesis of atherogenesis. But phytochemicals may act by other mechanisms; one group are weak oestrogens, phytoestrogens. Evidence about these substances is indirect, mostly epidemiological association[11] or effects *in vitro* or in animals. Some promising possibilities are shown in the box opposite.

Potentially toxic substances

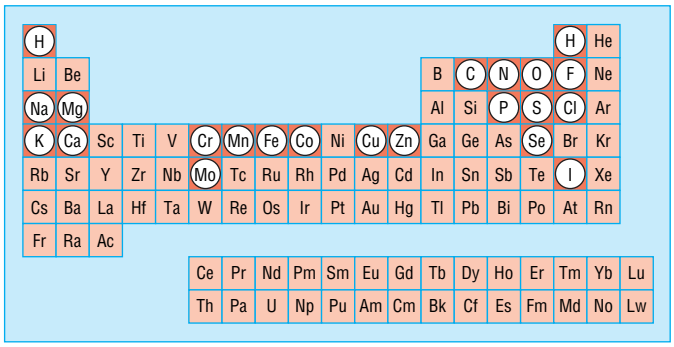
In most natural foods there are inherent substances that are potentially toxic but usually present in small amounts—for example, solanine in potatoes, nitrates and oxalates in spinach, thyroid antagonists in brassica vegetables, cyanogenetic glycosides in cassava and apricot stones, etc. Then there are substances that only some people are sensitive to—for example, in some people wheat causes gluten enteropathy, broad beans favism, and cheese a tyramine effect in patients taking monoamine oxidase inhibitors.

Other toxins get into foods when their environment is unusual—for example, toxic shellfish after a "red tide"—or if polluted with industrial contaminants, such as methyl mercury, polychlorinated biphenyls, etc. Microbiological infection can produce very potent toxins, such as botulism and aflatoxin. Deliberate food additives are not known to be toxic—if they were they would not be permitted by international or national food administrations. A few can cause sensitivity reactions in a minority of people (see chapter 15 on food sensitivity).

Amount of adult requirements for different nutrients

Adult daily requirements in foods	Essential nutrients for man
2-10 μg	Vitamin D, Vitamin B-12
c 50 μg	Vitamin K, Se, biotin, Cr
c 100 μg	Biotin, I, Mo
200-400 μg	Folate
1-2 mg	Vitamin A, thiamin, riboflavin, vitamin B-6, F, Cu
c 5 mg	Mn, pantothenate
c 15 mg	Niacin, vitamin E, Zn, Fe
c 50 mg	Vitamin C
300 mg	Mg
c 1 g	Ca, P
1-5 g	Na, Cl, K, essential fatty acids
c 50 g	Protein (10 or more essential amino acids)
50-100 g	Available carbohydrate
1 kg (litre)	Water

Figures are approximate and in places rounded. The range of requirements for different nutrients is about 10^9



Periodic table of the elements. Those essential for man are blocked in. In addition, boron, silicon, nickel, arsenic, and vanadium are still under consideration as ultra-trace nutrients[9]

Non-nutrient bioactive substances in food and drink that might help protect against chronic diseases

- **Carotenoids** other than β-carotene: *lycopene* (red pigment of tomatoes) and *lutein* (xanthophyll, in leafy vegetables). Though not pro-vitamin A they are antioxidants, are absorbed and seen in the plasma, and lutein is one of the pigments of the retinal macula lutea.
- **Polyphenols (flavonoids)**, antioxidants that occur in tea (especially green tea), wine (especially red)[12]: *catechins*; and in apples and onions: for example, *quercetin*.
- **Phytoestrogens**, especially isoflavones in soya: *genistein* and *daidzein*. Higher consumption of soya and soya products in East Asia might contribute to the lower incidences of breast and prostate cancers in that region.[13]

Patterns of nutrients in different foods

If animals are fed only one food sooner or later they will become ill and die. No single food contains all the essential nutrients. Wheat (wholemeal flour) lacks vitamins A, B-12, C, and D and is very low in iron and calcium (if unfortified); beef (muscle) contains little or no calcium, vitamins A, C, or dietary fibre. On the other hand, wheat is a good source of dietary fibre and beef of iron and vitamin B-12. The two together provide more nutrients than either alone but between them have no vitamin C or D and hardly any calcium. Addition of citrus fruit or salad brings vitamin C into the mixture, and milk or cheese adds the missing calcium and a little vitamin D.

This is the theory behind the "basic four" food groups for educating the public about nutrition. Each group has some deficiencies which the other three make up between them. You should aim to eat each day from each of: the bread and cereals group; the meat, poultry, and fish group; the vegetable and fruit group; and the milk group.

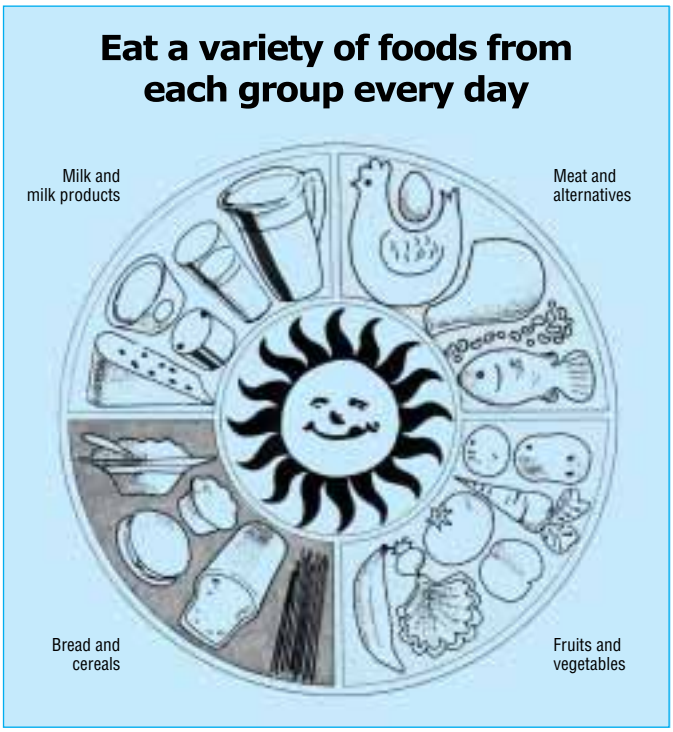
Variety

It is not enough to have daily servings of the same food from each group. One should also choose variety within food groups for two reasons. First, the characteristic nutrients in each group vary greatly for individual foods. Among fruits the vitamin C ranges from negligible (for dried fruits, grapes, and figs) up to 115-180 mg/100 g for stewed blackcurrants and canned guavas (this is in the British food tables; the international range goes up to about 3000 mg/100 g).[14]

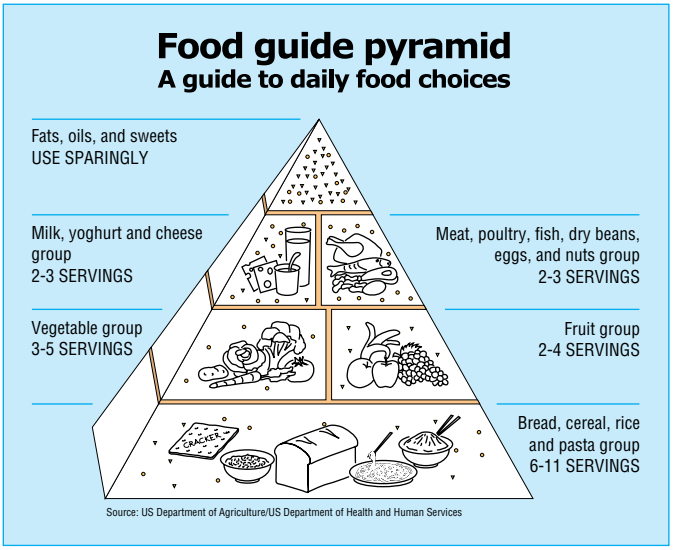
Second, natural toxins do not follow any of our arbitrary groupings of foods. The wider the variety of individual foods that people eat, the less their chance of acquiring harmful amounts of the toxins that are inevitable in foods but usually in small and subclinical amounts.

Blending dietary guidelines with food groups

The four groups are intended to minimise deficiency of traditional nutrients—protein, calcium, vitamin C, etc. In affluent countries, however, more disease is probably caused by too much fat, salt, and alcohol and not enough fibre. So we have to modify the older message. In the United States the Departments of Agriculture and of Health use a pyramid for nutrition education.[15] In the base (largest) layer ("eat most") is the cereal food group. The middle layer ("eat moderately") is for the vegetable group and the fruit group. The upper (smallest) layer ("eat least") is for the dairy group and the meat, etc, group. The areas allocated to each group convey broad quantitative recommendations and are accompanied by recommendations for numbers of servings. The divided plate on page 37 is based on the same principle.



A Canadian food guide



Possible modifications of four food groups to incorporate dietary guidelines

- **Bread**—Yes, but wholegrain and with lower salt. Prefer lower fat, low salt *cakes* and *biscuits*
- **Meat**—Lean cuts with the fat removed and not fried. Alternate with *fish* (grilled) and *legumes*
- **Vegetables** slightly cooked, not with salt
- **Fruit** fresh, not canned in syrup or dried
- **Milk** with half or all the cream removed

Junk foods and nutritious foods

Whether a food is nutritionally bad or good depends on the rest of the diet. As Hippocrates taught, "All things in nutriment are good or bad relatively". An extra portion of saturated fat is bad in Britain but would be good for starving children in north east Africa. An orange does nothing for someone who takes vitamin C tablets but is important for an elderly person who eats no vegetables. Value judgements about foods are being made all the time; they are nearly always subjective and often wrong.

A good objective method is to work out for a typical serving of the food its provisions of important nutrients, as a percentage of their recommended dietary intakes, compared with its content of energy (calories), also as a percentage of a standard daily intake. For each nutrient:

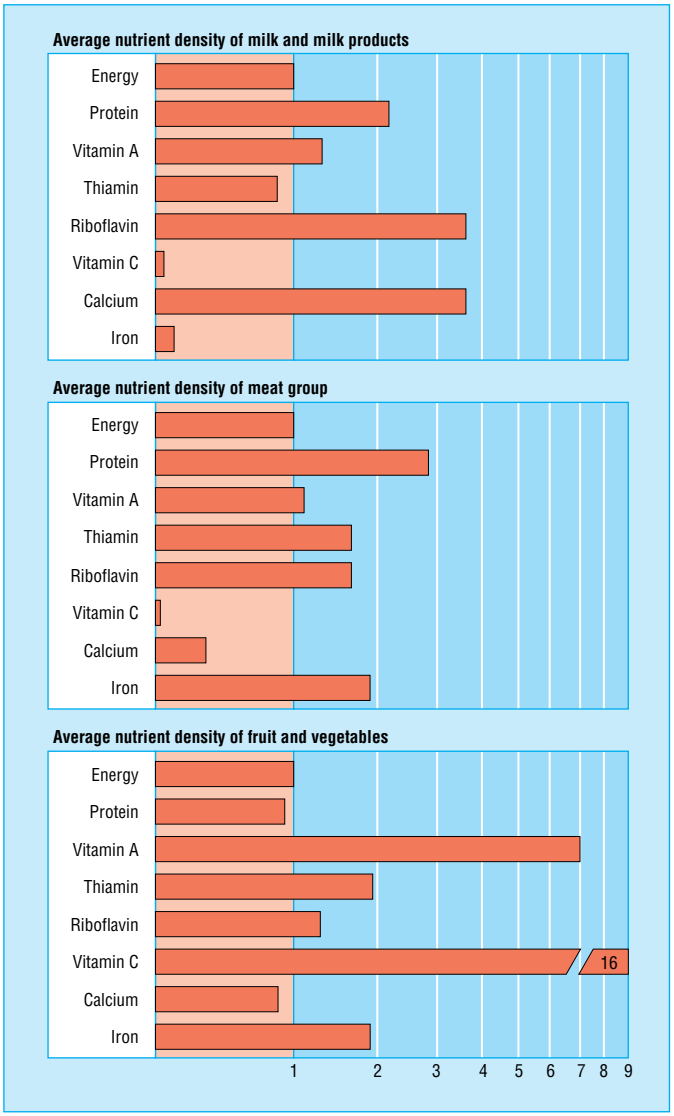
$$\text{the index of nutritional quantity} = \frac{\text{nutrient as \% standard}}{\text{energy as \% standard}}$$

"Nutrient dense" foods have high ratios of important nutrients to energy (calories).

The profile of indices for major nutrients can be put in an array. Other components in the food, like cholesterol, saturated fatty acids, and dietary fibre can be treated in a similar way by using a dietary goal as the standard.

The table below, modified from an American book,[17] shows that egg contains a smaller proportion of fat per energy (calories) than butter; the fat is less saturated and egg is also a good source of protein and some other nutrients. Egg and butter both contain some vitamin A but egg contains thiamin, riboflavin, iron, calcium, protein—not found in butter. However, an egg contains much more cholesterol than $\frac{1}{2}$ oz (14g) butter.

Calculations of this type should be made before authorities advise communities to eat more or less of a food. Applying them to the 1995 Department of Health recommendations[18] about diet to prevent cardiovascular disease means that the amount of butter eaten should be reduced more than the amount of egg, because reduced saturated fat is recommended but current cholesterol intake is not considered excessive. In the United States, however, a dietary guideline[15] advises against high levels of dietary cholesterol and so recommends the general public to moderate its consumption of egg yolks.



Nutrient density is the ratio of a nutrient (expressed as % of recommended daily intake) to energy (expressed as % of a standard energy intake). In the total diet of mixed foods the density for each nutrient should exceed 1.0. (From Hansen[16])

Indices of nutritional quality (INQ) for butter and egg

	Butter ($\frac{1}{2}$ oz; 14 g)			Egg (50 g), hard boiled		
	Amount	% of standard	INQ	Amount	% of standard	INQ
Energy (kcal)	100	5	1.0	80	4	1.0
Vitamin A (mg)	0.129	11	2.2	0.078	7	1.6
Thiamin (mg)	0	0	0	0.04	4	1.0
Riboflavin (mg)	0	0	0	0.14	12	2.9
Niacin (mg)	0	0	0	0.03	0	0
Vitamin C (mg)	0	0	0	0	0	0
Iron (mg)	0	0	0	1.0	6	1.5
Calcium (mg)	3	0	0.07	28.0	3	0.8
Potassium (mg)	4	0	0.02	65	1	0.3
Protein (g)	0	0	0	6	12	3.0
Carbohydrate (g)	0	0	0	1	0	0.1
Fat (g)	12	15	3.1	6	8	1.9
Oleic acid (g)*	2.9	12	2.4	2	8	2.0
Linoleic acid (g)	0.3	2	0.3	0.6	3	0.8
Saturated fatty acids (g)*	7.2	25	5.1	1.7	6	1.5
Cholesterol (mg)*	32	11	2.2	225	75	19

Based on Hansen RG *et al.*[17] [The standards they used are energy 2000 kcal (8.4 MJ), vitamin A 1.2 mg, thiamin 1 mg, vitamin C 60 mg, riboflavin 1.2 mg, niacin 14 mg, iron 16 mg, calcium 900 mg, potassium 5000 mg, protein 50 g, carbohydrate 275 g, fat 78 g, oleic acid 24.5 g, linoleic acid 20 g, saturated fatty acids 28.5 g.] I have taken 300 mg as standard for cholesterol. These are all intakes per day
*Not essential nutrients

Calories do count

The law of conservation of energy applies to human nutrition as in the rest of nature. Atwater established this around 1900. A little more heat may be produced after some foods or in some people but the more calories (or kilojoules) you eat the more you can expect to store as adipose tissue.

Foods differ in their calorie content from 32 kJ/100 g (7 kcal/100 g) for celery, up to 3.7 MJ/100 g (899 kcal/100 g) for vegetable oils—a 128-fold range. This great range depends on the different energy values of fat, alcohol, protein, and carbohydrate and how much these are diluted by water. It is useful for doctors to know the energy values of average servings of common foods (there is a short list in chapter 11 on obesity).

Energy values as metabolised in the body of the main energy-yielding groups of food components (Atwater factors)

	kcal/g	kJ/g
Fat	9	37
Alcohol	7	29
Protein	4	17
Carbohydrate*	3.75	16

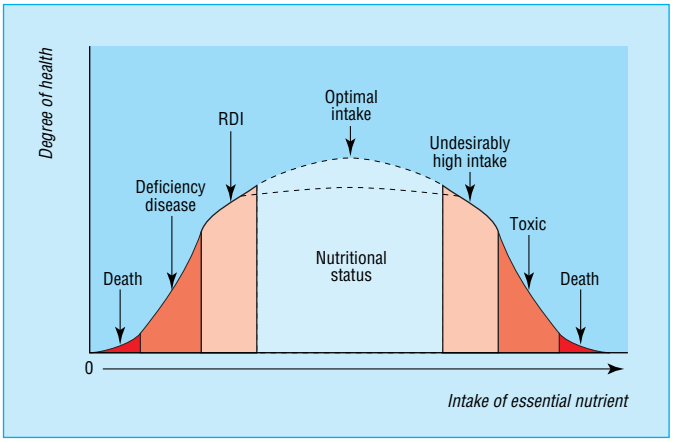
*This is for available carbohydrate. The energy provided by dietary fibre from its fermentation to volatile fatty acids in the large intestine is less than half this amount

No perfect diet

There are several diets that appear (in our present state of knowledge) to be good. We can advise on a better diet for Mr Smith or, as in a United States report, make recommendations "towards healthful diets", but there is no best diet. The reason is that man is an omnivore with enzyme systems that can adapt to ranges of intakes of many food components. There is, for example, an inducible enzyme, sucrase, in the small intestinal epithelium; if people eat sucrose this enzyme appears and digests it. There are several enzymes in the liver which oxidise amino acids; their activity increases when protein intake is high and falls in people on low protein diets.

The dose determines the effect

When the intake of one essential nutrient is varied, with the rest of the diet adequate in other nutrients and energy, the individual's state of health is likely to be very poor if intakes of the essential nutrient are inadequate and sustained. Health improves as the intake is increased, up to the nutritional requirement level. Above this, it has been thought that the state of health remains on a plateau up until the nutrient intake becomes undesirably high, beyond which toxicity may be seen. Recent experience with some nutrients suggests, however, that above the requirement level, which cures deficiency disease, there can be an optimal range of intake. The individual may not feel or function differently but has a reduced risk of degenerative disease or more favourable biochemical profile. Folate is a good example. Above the level that cures or prevents megaloblastic anaemia women have a reduced risk of a malformed baby, and older adults are less likely to have a raised plasma homocysteine.

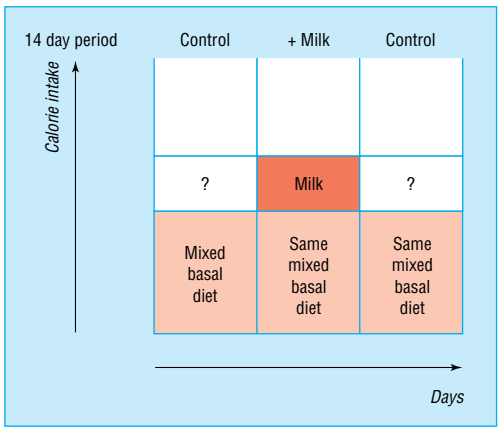


Ingestion of the RDI should guarantee no deficiency disease but beyond the RDI there may still be additional health benefits (for example, partial protection from a degenerative disease). The top of the dome beyond the RDI is then the optimal intake range. RDI corresponds to Reference Nutrient Intake. (Adapted from Truswell[20])

Replacement

For every food you remove from the diet another has to take its place. This principle is prominent in the design and interpretation of nutritional experiments. Does consumption of milk raise or lower the plasma cholesterol concentration? To test this an adequate but physiological amount of milk is to be given in a middle two or three week period. The plasma cholesterol value is measured at the end of this period and at the end of equal length control periods before and after.[19] But what should be given to replace the calories of the milk in the control periods? If nothing is given the periods will not be isocaloric.

To some extent the effect of milk on plasma cholesterol could be manipulated by the choice of the control food. We do not want to influence the experiment so might ask, "If people here stop drinking milk what would they drink (or eat) in its place: beer, water, fruit juice, fizzy drink, etc?" A similar situation applies in outpatients when the doctor or dietitian instructs a patient to cut out one food from his/her diet.



Unless he/she is to lose weight he/she will sooner or later choose other food(s) as replacement, which may affect the outcome.

Some concluding proverbs

People have been thinking about the safety and goodness of food, as well as its social roles and tastiness, ever since the Garden of Eden or its evolutionary counterpart. So it is perhaps not surprising that a number of proverbs about food and eating are being confirmed by nutritional science.

Moderation in all things
The recommendation of many expert committees on nutrition. Do not eat too much or too little of anything, and do not follow one of the extreme unorthodox regimens.

Man cannot live by bread alone
Though the original was about spiritual nourishment, it is also true that people have to eat more than one (type of) food.

Variety is the spice of life
You should eat a mixed and varied choice of foods.

Enough is as good as a feast
More leads to obesity. People's energy requirements differ. "Enough" is an individual amount.

You can have too much of a good thing
For example, saturated fat, salt, dietary cholesterol, vitamins A, D, and B-6, and alcohol.

One man's meat is another man's poison
The subject of chapter 15 on food sensitivity. For each of us there are foods we dislike and may well be foods that can make us ill.

There's no accounting for taste
Taste has to be considered in planning therapeutic diets.

A little of what you fancy does you good
Dietary prescriptions are sometimes more rigid than they need be. This proverb also speaks of the placebo effect; if someone believes a food is doing him good he may feel better for a time after eating it.

Old habits die hard
Food habits must be respected. Prescribed dietary changes are likely to be followed better if they are fitted into the least strongly held of an individual's food habits.

There's many a slip twixt cup and lip
People do not necessarily eat what they intend or say they eat. That patient you just put on a diabetic diet may not have understood you.

References

1 Truswell AS, Hansen JDL. Medical research among the !Kung. In: Lee RB, De Vore I (eds) *Kalahari hunter-gatherers*. Cambridge, MA: Harvard University Press, 1976.
2 Eaton SB, Eaton SB III, Konner MJ. Paleolithic nutrition revised: a twelve-year retrospective on its nature and implications. *Eur J Clin Nutr* 1997; **51**: 207-16.
3 Leaf A, Weber PC. Cardiovascular effects of n-3 fatty acids. *N Engl J Med* 1988; **318**: 549-57.
4 de Wardener HE, Lennox B. Cerebral beriberi (Wernicke's encephalopathy). *Lancet* 1947; **1**: 11-17.
5 Hu FB, Stampfer MJ, Manson JE *et al*. Dietary fat intake and the risk of coronary heart disease in women. *N Engl J Med* 1997; **337**: 1491-9.
6 Widdowson EM. Animals in the service of human nutrition. In: Taylor TG, Jenkins NG (eds) *Proceedings of the XVIII International Congress of Nutrition (Brighton, 1985)*. London: John Libbey, 1986:52–7.
7 Freund H, Atamian S, Fischer JE. Chromium deficiency during total parenteral nutrition. *JAMA* 1979; **241**: 496-8.
8 Truswell AS. Levels and kinds of evidence for public health nutrition. *Lancet* 2001; **357**: 1061-2.
9 Nielsen FH. Other trace elements. In: Ziegler EE, Filer LJ (eds) *Present knowledge in nutrition*, 7th edn, Washington DC: ILSI Press, 1996.
10 Steinmetz KA, Potter JD. Vegetables, fruit and cancer prevention: a review. *J Am Diet Assoc* 1996; **96**: 1027-39.
11 Hertog MGL, Feskens EJM, Hollman PCH, Katan MB, Kromhout D. Dietary antioxidant flavonoids and risk of coronary heart disease: the Zutphen Elderly Study. *Lancet* 1993; **342**: 1007-11.
12 Frankel EN, Waterhouse AL, Teissedre PL. Principal phenolic phytochemicals in selected Californian wines and their antioxidant activity in inhibiting oxidation of human low-density lipoproteins. *J Agric Food Chem* 1995; **43**: 890-4.
13 Adlerkreutz H, Markkanen H, Watanabe S. Plasma concentrations of phytoestrogens in Japanese men. *Lancet* 1993; **342**: 1209-10.
14 Brand JC, Cherikoff V, Lee A, Truswell AS. An outstanding food source of vitamin C. *Lancet* 1982; **ii**: 873.
15 US Department of Agriculture; US Department of Health & Human Services. *Dietary guidelines for Americans*, 5th edn. *Home and Garden Bulletin no 232*. Beltsville, MD: Food and Nutrition Information Center, USDA, 2000.
16 Hansen RG. An index of food quality. *Nutr Rev* 1973; **51**: 1-7.
17 Hansen RG, Wyse BW, Sorenson AW. *Nutritional quality index of foods*. Westport, CT: Avi Press, 1979.
18 Department of Health. *Nutritional aspects of cardiovascular disease. Report of the Cardiovascular Review Group, Committee on Medical Aspects of Food Policy*. London: HMSO, 1995.
19 Roberts DCK, Truswell AS, Sullivan DR. Milk, plasma cholesterol and controls in nutritional experiments. *Atherosclerosis* 1982; **42**: 323-5.
20 Truswell AS. New vitamin research: antioxidants and folate. Introduction. *Asia Pacific J Clin Nutr* 1993; **2** (Suppl 1): 1-3.

Index

Page numbers in bold refer to figures, those in italics refer to tables or boxed material


ABC
OF
OCCUPATIONAL AND
ENVIRONMENTAL
MEDICINE
SECOND EDITION
Edited by David Snashall and Dipti Patel
BMJ
Books

ABC OF OCCUPATIONAL AND ENVIRONMENTAL MEDICINE

Second Edition

ABC OF OCCUPATIONAL AND ENVIRONMENTAL MEDICINE

Second Edition

Edited by

DAVID SNASHALL

Head of Occupational Health Services, Guy's and St Thomas's Hospital NHS Trust, London
Chief Medical Adviser, Health and Safety Executive, London

DIPTI PATEL

Consultant Occupational Physician, British Broadcasting Corporation, London

BMJ Books

ЗНАНИЕ

First published in 1997 as *ABC of Work Related Disorders*
This edition published as *ABC of Occupational and Environmental Medicine—Second edition* 2003
by BMJ Publishing Group, BMA House, Tavistock Square,
London WC1H 9JR
www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

The cover shows an scanning electron micrograph of asbestos fibres. With permission from Manfred Kage/Science photo Library

ISBN 0 7279 1611 4

Typeset by Newgen Imaging Systems (P) Ltd, Chennai, India
Printed and bound in Malaysia by Times Offset

Contents

Contributors

Anil Adisesh
Consultant in Occupational Medicine, Trafford Healthcare and Salford Royal Hospitals NHS Trusts, and Honorary Clinical Lecturer in Occupational and Environmental Medicine, Centre for Occupational and Environmental Health, University of Manchester, Manchester

Nicola Cherry
Chair, Department of Public Health Sciences, University of Alberta, Canada

David Coggon
Professor of Occupational and Environmental Medicine, MRC Environmental Epidemiology Unit, Community Clinical Sciences, University of Southampton, Southampton

Tom Cox
Professor of Organisational Psychology, Institute of Work, Health and Organisation, University of Nottingham, Nottingham

Martyn Davidson
Chief Medical Adviser, John Lewis Partnership, London

William Davies
Consultant Occupational Physician, Occupational Health Unit, South Wales Fire and Local Authorities, Pontyclun, Wales

Tony Fletcher
Senior Lecturer, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London

Mats Hagberg
Professor, Chief Physician, and Director, Department of Occupational and Environmental Medicine, Sahlgrenska Academy at Gothenburg University and Sahlgrenska University Hospital, Gothenburg, Sweden

John Hobson
Consultant Occupational Physician, MPCG Ltd, Stoke on Trent

Malcolm IV Jayson
Emetrius Professor of Rheumatology and Professorial Fellow, University of Manchester, Manchester

Rachel Jenkins
Director of WHO collaborating Centre, Institute of Psychiatry, King's College, Denmark Hill, London

Paul Litchfield
BT Group Chief Medical Officer, London

Deborah Lucas
Head of Human Factors Team, Hazardous Installations Directorate Health and Safety Executive, Bootle, Merseyside

Ron McCaig
Head of Human Factors Unit, Better Working Environment Division Policy Group, Health and Safety Executive, Bootle, Merseyside

Ira Madan
Consultant Occupational Physician, East Kent Hospitals NHS Trust, Canterbury

Robert Maynard
Senior Medical Officer, Department of Health, Skipton House, London

Keith T Palmer
MRC Clinical Scientist and Consultant, Occupational Physician, MRC Environmental Epidemiology Unit, Community Clinical Sciences, University of Southampton, Southampton

Dipti Patel
Consultant Occupational Physician, British Broadcasting Corporation, London

Andy Slovak
Company Chief Medical Officer, British Nuclear Fuels plc, Warrington, Cheshire

David Snashall
Senior Lecturer in Occupational Medicine, Guy's, King's and St Thomas's School of Medicine, London; Honorary Consultant and Head of Service, Occupational Health Department, Guys' and St Thomas's Hospitals NHS Trust, London; Chief Medical Advisor, Health and Safety Executive, London

Ian R White
Consultant Dermatologist, St John's Institute of Dermatology, London

Preface

Although work is generally considered to be good for your health and a healthy working population is essential to a country's economic and social development, certain kinds of work can be damaging. Occupational health is the study of the effect—good and bad—of work on peoples' health and, conversely, the effect of peoples' health on their work: fitness for work in other words.

Work places are specialised environments, capable of being closely controlled. Generally, it is the lack of control imposed by employers that is the cause of ill health because of exposure to hazardous materials and agents at work, and of injury caused by workplace accidents.

Working life does not, however, begin and end at the factory gate or the revolving office door: many people walk, cycle, or drive to work—a journey that often constitutes the major hazard of the day. Others have to drive or travel by other means as part of their job, live away from home, be exposed to other food, other people, other parasites. Even work from home, increasing in some countries, can have its problems. Occupational health practitioners deal with all these aspects of working life.

A working population consists of people mainly between 15 and 70 years (disregarding for the moment the ongoing scandal that is child labour), who may be exposed for 8-12 hours a day to a relatively high concentration of toxic substances or agents, physical or psychological. At least that population is likely to be reasonably fit—unlike those who cannot work because of illness or disabilities, the young, and the very old, who are more vulnerable and spend a lifetime exposed to many of the same agents in the general environment at lower concentration. This enters the realm of environmental medicine of such concern to those who monitor the degradation of our planet, track pollution and climate change, and note the effect of natural disasters and man made ones, especially wars.

This book was first published in 1997 as the *ABC of Work Related Disorders*. It is a much expanded and updated version that attempts, in a compressed and easy to assimilate fashion, to describe those problems of health relating to work in its widest sense and to the environment.

The pattern of work is changing fast. There is pretty full employment in most economically developed countries now. Manufacturing industry is now mainly concentrated in developing countries where traditional occupational disease such as pesticide poisoning and asbestosis are still depressingly common. Occupational accidents are particularly common in places where industrialisation is occurring rapidly as was once the case during the industrial revolution in 19th century Britain. Work is also more varied, more intense, more service oriented, more regulated, and more spread around the clock in order to serve the 24 hour international economy. There are more women at work, more disabled people, and a range of new illnesses perhaps better described as symptom complexes which represent interactive states between peoples' attitudes and feeling towards their work, their domestic environment, and the way in which their illness behaviour is expressed.

All occupational disease is preventable—even the more "modern" conditions such as stress and upper limb disorders can be reduced to low level by good management and fair treatment of individuals who do develop these kinds of problems and who may need rehabilitation back into working life after a period of disability. These areas are covered in the chapters on musculoskeletal disorders, stress, and mental health at work. There are chapters also on the traditional concerns of the occupational health practitioner such as dermatoses, respiratory disorders and infections, and other chapters reflecting occupational health practice covering workplace surveys, fitness for work, sickness absence control issues, and, unfortunately increasing in prevalence, legal considerations. Genetics and its application to work and the effects of work on reproduction are described in chapter 17.

Concerns beyond the workplace are covered in the chapter on global issues and on pollution. The control of hazards in the general environment presents issues of problem solving at a different level. Ascertainment of exposure is more difficult than in workplaces, and to find solutions needs transnational political will and commitment as well as science to succeed. Many believe that the rash of "new" illnesses attributed to environmental causes are manifestations of a risk-averse public's response to poorly understood threats in the modern world and an unconscious wish to blame "industry," or some state institution—agencies that represent irresponsible emitters of toxins, inadvertent releasers of radiation, regardless sprayers of pesticides, or unwitting providers of vaccinations. Chapter 20 addresses this important subject.

In common with the previous edition, this new edition of *ABC of Occupational and Environmental Medicine* will still appeal to non-specialists who wish to practise some occupational medicine; but will also provide all that students of occupational and environmental medicine and nursing will need as a basis for their studies. Each chapter has an annotated further reading list. Most, but not all, of the book is written with an international audience in mind.

David Snashall

1 Hazards of work

David Snashall

Most readers of this book will consider themselves lucky to have a job, probably an interesting one. However tedious it might be, work defines a person, which is one reason why most people who lack the opportunity to work feel disenfranchised. As well as determining our standard of living, work takes up about a third of our waking time, widens our social networks, constrains where we can live, and conditions our personalities. "Good" work is life enhancing, but bad working conditions can damage your health.

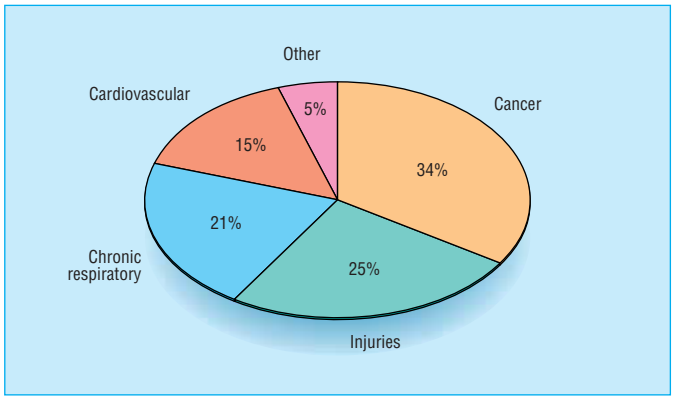
Global burden of occupational and environmental ill health

According to the International Labour Organisation (ILO), between 1.9 and 2.3 million people are killed by their work every year—including 12 000 children—and 25 million people have workplace injuries, causing them to take time off. Two million workplace associated deaths per year outnumber people killed in road accidents, war, violence, and through AIDS, and cost 4% of the world's gross domestic product in terms of absence from work, treatment, and disability and survivor benefits.

The burden is particularly heavy in developing countries where the death rate in construction—for example, is 10 times that in industrialised countries, and where workers are concentrated in the most dangerous industries—fishing, mining, logging, and agriculture.

In the United States some 60 000 deaths from occupational disease and 860 000 cases of work related injury occur each year.

Environmental disease is more difficult to quantify because the populations at risk are much larger than the working population. As an example, the US Centers for Disease Control and Prevention reckons that one million children in the world have lead poisoning.



Estimated global work related mortality (1.1 million every year, based on 1990-5 data). Other diseases include pneumoconioses, nervous system, and renal disorders

Reporting occupational ill health

Occupational diseases are reportable in most countries, but are usually grossly underreported. Even in countries like Finland (where reporting is assiduous), surveys have shown rates of occupational disease to be underestimated by three to five times.

Classifications of occupational diseases have been developed for two main purposes: for *notification*, usually to a health and safety agency to provide national statistics and subsequent preventive action, and for *compensation* paid to individuals affected by such diseases. There are no universally accepted diagnostic criteria, coding systems, or classifications worldwide. Modifications of ICD-10 (international classification of diseases, 10th revision) are used in many countries to classify occupational diseases, along with a system devised by the World Health Organization for classifying by exposure or industry. It is the association of these two sets of information that defines a disease as being probably occupational in origin. A number of reporting systems exist in the United Kingdom but these are not comprehensive, nor coordinated. After all, they arose at different times and for different purposes.

Children are more vulnerable to occupational disease—they are smaller, have the potential to be exposed for many years, and their tissues are more sensitive. They are also more likely to be exploited and, being less aware, more accident prone

Classification and notification of occupational diseases

The World Health Organization gives the following classification:

1. Diseases caused by agents
   - 1.1 Diseases caused by chemical agents
   - 1.2 Diseases caused by physical agents
   - 1.3 Diseases caused by biological agents
2. Diseases by target organ
   - 2.1 Occupational respiratory diseases
   - 2.2 Occupational skin diseases
   - 2.3 Occupational musculoskeletal diseases
3. Occupational cancer
4. Others

Notification

In addition to the diagnosis of occupational disease, additional information should be included in the notification. The ILO has defined the minimum information to be included:

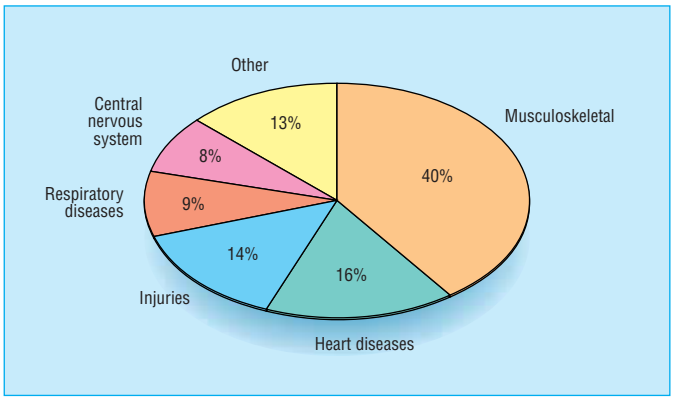
(a) Enterprise, establishment, and employer
   - (i) Name and address of employer
   - (ii) Name and address of enterprise
   - (iii) Name and address of the establishment
   - (iv) Economic activity of the establishment
   - (v) Number of workers (size of the establishment)

(b) Person affected by the occupational disease
   - (i) Name, address, sex, and date of birth
   - (ii) Employment status
   - (iii) Occupation at the time when the disease was diagnosed
   - (iv) Length of service with the present employer

Classification for labour statistics

- International Standard Classification of Occupations (ISCO)
- International Classification of Status in Employment (ICSE)
- International Standard Industrial Classification of all Economic Activities (ISIC)
- International Standard Classification of Education (a UNESCO classification) (ISCED)
- Classifications of occupational injuries

Occupational injuries are also reportable in Great Britain under the Reporting of Injuries, Diseases and Dangerous Occurrences Regulations 1995 and, for purposes of compensation, to the Department of Work and Pensions' Industrial Injuries Scheme. The recording of injuries is generally more reliable because the injuries are immediately obvious and occur at a definable point in time. By contrast, cause and effect in occupational disease can be far from obvious, and exposure to the hazardous material may have occurred many years beforehand. Given that, worldwide, industrial injuries and, in particular, occupational ill health are poorly recorded and reported, the economic losses to the countries concerned are massive.



Breakdown of costs for work related injuries and diseases. Other diseases include cancer, skin diseases, and mental disorders

Occupational or work related?

Some conditions, such as asbestosis in laggers, and lead poisoning in industrial painters, are hardly likely to be anything other than purely occupational in origin. (About 70 of these "prescribed" occupational diseases are listed by the Department for Work and Pensions.) However, mesothelioma can be the result of environmental exposure to fibrous minerals (as in the case of cave dwellers in Turkey), and lead poisoning can be a result of ingesting lead salts from—for example, low

The cost of disease and injury at work

- **1992: European countries**
  Direct costs for compensation of work related diseases and injuries:
  27 000 million ECUs
- **1995-1996: United Kingdom**
  Overall costs to society for workplace injuries and ill health (including net present value of costs in future years):
  £14-18 billion (2-2.5% of gross domestic product). Ratio of illnesses/injuries about 3:1
- **1992: United States**
  Total direct and indirect costs associated with work related injuries and diseases: US $171 000 million. This is more than AIDS and on a par with cancer and heart disease

United Kingdom occupational ill health statistics

No single source of information is available in the United Kingdom on the nature and full extent of occupational ill health. The statistics in the 2000-1 report by the Health and Safety Executive are based on the following sources:

- Household surveys of self reported work related illness (SWI): these have been held in 1990 and 1995, linked to the Labour Force Survey (LFS). Health and safety questions were also included in the Europewide LFS in 1999
- Voluntary reporting of occupational diseases by specialist doctors in The Health and Occupation Reporting (THOR) network (which succeeded the Occupational Disease Intelligence Network (ODIN) in 2002). THOR and ODIN comprise the Occupational Physicians Reporting Activities (OPRA) scheme, and six other schemes covering mental illness and stress, musculoskeletal disorders, skin diseases, respiratory disorders, hearing loss, and infectious diseases
- New cases of assessed disablement under the Department of Work and Pensions' Industrial Injuries Scheme (IIS): the most longstanding source, based on a list of prescribed diseases and associated occupations, again giving annual figures
- Statutory reports under the Reporting of Injuries, Diseases and Dangerous Occurrences Regulations (RIDDOR): these were expanded when RIDDOR 1995 replaced RIDDOR 1985 and are similar to the IIS list
- Deaths from occupational lung diseases recorded on death certificates (principally mesothelioma and other asbestos related diseases)

temperature, lead glazed ceramics used as drinking vessels, mainly in developing countries. In these situations the history and main occupation will differentiate the causes. The situation may be far less clear in conditions such as back pain in a construction worker or an upper limb disorder in a keyboard operator when activities outside work may contribute, as might psychological factors, symptom thresholds, etc. A lifetime working in a dusty atmosphere may not lead to chronic bronchitis and emphysema, but when it is combined with cigarette smoking this outcome is much more likely. Common conditions for which occupational exposures are important but are not the sole reason or the major cause can more reasonably be termed "work related disease" rather than occupational disease.

Some important prescribed diseases such as chronic bronchitis, emphysema, and lung cancer are work related in the individual case only on the "balance of probabilities." Certain occupations carry a substantial risk of premature death, whereas others are associated with the likelihood of living a long and healthy life. This is reflected in very different standardised (or proportional) mortality ratios for different jobs, but not all the differences are the result of the various hazards of different occupations. Selection factors are important, and social class has an effect (although in the United Kingdom this is defined by occupation). Non-occupational causes related to behaviour and lifestyle may also be important.

Proportional mortality ratios (PMR) in selected occupations

Occupation	High PMR	Low PMR
Teachers	Multiple sclerosis Leukaemia Aplastic anaemia Parkinson disease Bicycle accidents	Lung cancer Bronchitis Alcohol related disease
Doctors, dentists and nurses	Suicide Alcohol-related disease Hepatitis (doctors) Prostatic cancer (dentists)	Ischaemic heart disease
Farmers	Allergic pneumonitis Influenza Hernia Poisoning Accidents Epilepsy Suicide Haemolytic anaemia	Cancer Heart disease Alcohol related disorders
Construction workers	Cancer of pleura and peritoneum Asbestosis Nasal cancer Falls	Suicide

Presentation of work related illnesses

Diseases and conditions of occupational origin usually present in an identical form to the same diseases and conditions caused by other factors. Bronchial carcinoma—for example, has the same histological appearance and follows the same course whether it results from working with asbestos, uranium mining, or cigarette smoking.

The possibility that a condition is work induced may become apparent only when specific questions are asked, because the occupational origin of a disease is usually discovered (and it is discovered only if suspected) by the presence of an unusual pattern. For example, in occupational dermatitis, the distribution of the lesions may be characteristic. A particular history may be another clue: asthma of late onset is more commonly occupational in origin than asthma that starts early in life. Indeed, some 40% of adult onset asthma is probably occupational. Daytime drowsiness in a fit young factory worker may be caused not by late nights and heavy alcohol consumption but by unsuspected exposure to solvents at work.

The occupational connection with a condition may not be immediately obvious because patients may give vague answers when asked what their job is. Answers such as "driver," "fitter," or "model" are not very useful, and the closer a health professional can get to extracting a precise job description, the better. For example, an engineer may work directly with machinery and risk damage to limbs, skin, and hearing, or may spend all day working at a computer and risk back pain, upper limb disorders, and sedentary stress. Sometimes patients will have been told (or should have been told) their job is associated with specific hazards, or they may know that fellow workers have experienced similar symptoms.

Timing of events

The timing of symptoms is important because the symptoms may be related to exposure events during work. Asthma provides a good example of this: many people with occupational asthma develop symptoms only after a delay of

Exposure to solvents at work can be the cause of erratic behaviour at home

How to take an occupational history

Question 1
What is your job? or What do you do for a living?

Question 2
What do you work with? or What is a typical working day for you? or What do you actually do at work?

Question 3
How long have you been doing this kind of work? Have you done any different kind of work in the past?

Question 4
Have you been told that anything you use at work may make you ill? Has anybody at work had the same symptoms?

Question 5
Do you have any hobbies, like do-it-yourself or gardening, which may bring you into contact with chemicals?

Question 6
Is there an occupational health doctor or nurse at your workplace who I could speak to?

some hours and the condition may present as nocturnal wheeze. It is essential to ask whether symptoms occur during the performance of a specific task and if they occur solely on workdays, improving during weekends and holidays. Sometimes the only way to elucidate the pattern is for the person to keep a graphic diary of the time sequence of events.

Working conditions

Patients should be asked specifically about their working conditions. Common problems are dim lighting, noisy machinery, bad office layout, dusty atmosphere, draconian management, and bad morale. Such questioning not only investigates possibilities, but also gives the questioner a good idea of the general state of a working environment and how the patient reacts to it. A visit to the workplace may be a revelation, and just as valuable as a home visit if one wants to understand how a patient's health is conditioned by their environment and how it might be improved. Knowing about somebody's work can help to provide a context and to gain insight. Patients are often happy to talk about the details of their work: this may be less threatening than talking about details of their home life and can promote a better relationship between patients and health professionals.

The causes of occupational disease can extend beyond the workplace and can affect local populations through water or soil pollution. Overalls soiled with toxic materials such as lead or asbestos can affect members of workers' families when the overalls are taken home to be washed.

Trends in work related illnesses

Changes in working practices in the industrialised world are giving rise to work that is more demanding in a psychosocial sense but less so in terms of hard physical activity. Jobs are also safer (although this may not be true in those countries where extremely rapid industrialisation is occurring)—the result of a shift in many countries from agricultural and extractive industry via heavy factory industry to technology intensive manufacturing and services, which are inherently safer. Also, most countries have a labour inspectorate that can orchestrate a risk based strategy of hazard control with varying degrees of efficiency. Life outside work has also become safer, although rapid industrialisation and growing prosperity in some countries have meant huge increases in road traffic, with an accompanying increase in accidents. Traditional occupational diseases such as pneumoconiosis and noise induced deafness can be adequately controlled by the same strategies of hazard control used to limit accidental injury. However, the long latent period between exposure and appearance of occupational diseases makes attribution and control more problematic. Thus, the modern epidemic of musculoskeletal disorders and complaints of work induced stress may reflect a new kind of working population with different characteristics from its forebearers, as well as changes in the work environment itself.

Completely new jobs have appeared, with new hazards—for example, salad composers (dermatitis), aromatherapists (allergies), and semiconductor assemblers (exposure to multiple toxins).

Although working conditions are undoubtedly cleaner, safer, and in many ways better than before, work itself has changed. In the economically developed world there has been a shift from unskilled work to more highly skilled or multiskilled work in largely sedentary occupations. There is greater self employment and a remarkable shift towards employment in small and medium sized enterprises. The percentage of women in employment has been growing for

An example of the interface between occupational and environmental disease was the pollution of Minamata Bay in Japan by discharges of mercury from industrial sources and the severe neurological consequences on those who consumed the resulting contaminated fish

Annual death risks: some examples from the United Kingdom

Cause of death	Annual risk
Whole population	
Cancer	1 in 387
All forms of road accidents	1 in 16 800
Lung cancer caused by radon in dwellings	1 in 29 000
Lightning	1 in 1 870 000
Workers	
Fatalities to employees	1 in 125 000
Fatalities to the self employed	1 in 50 000
Construction	1 in 17 000
Agriculture, hunting, forestry, and fishing (not sea fishing)	1 in 17 200
Service industry	1 in 333 000
Activities	
Surgical anaesthesia	1 in 185 000 operations
Scuba diving	1 in 200 000 dives
Fairground rides	1 in 834 000 000 rides
Rock climbing	1 in 320 000 climbs
Rail travel accidents	1 in 43 000 000 passenger journeys
Aircraft accidents	1 in 125 000 000 passenger journeys

Antidiscrimination legislation in many countries has provided more working opportunities for disabled and older workers, and has provided their employers with some challenges. Occupational health professionals need to understand organisational development as well as occupational disease

Useful websites

WHO http://www.who.int/home-page
ILO http://www.ilo.org/public/English
ICOH http://www.icoh.org.sg/eng/index.html

Africa
http://www.sheafrica.info

Australia
http://www/nohsc.gov.au

Europe
http://europe.osha.eu.int

Finland
http://www.occuphealth.fi

Sweden
http://www.arbetslivsinstitutet.se

United Kingdom
http://www.hse.gov.uk
http://www.facoccmed.ac.uk

United States
http://www.cdc.gov/niosh/homepage.html
http://www.epa.gov/
http://www.acoem.org

decades. Not everyone can cope with the newer, more flexible, less stable, intensively managed work style demanded by modern clients and contractors.

Public perceptions and an expectation of good physical health and associated happiness, allied to improved sanitation and housing, availability of good food, and good medical services, have highlighted those non-fatal conditions which might hitherto have been regarded as trivial but which have large effects on social functioning (such as deafness), work (such as backache), and happiness (such as psychological illness), contributing in turn disproportionately and adversely to disability-free years of life. The public is also more environmentally aware and concerned that some of the determinants of ill health are rooted in modern life and working conditions, giving rise to allergies, fatigue states, and various forms of chemical sensitisation. The estimation, perception, and communication of risk—a social construct—may still, however, be quite primitive even in the most sophisticated of populations. The media definition of risk remains "hazard plus outrage," and life as a threat has become a reality for many.

The figures showing global work related mortality and the breakdown of costs for work related injuries and diseases use data from ILO, 1999 and ILO, 1995.

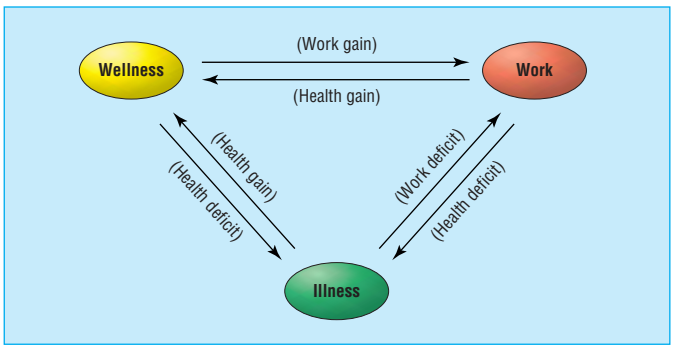
2 Occupational health practice

Anil Adisesh

Occupational health is a multidisciplinary activity that draws on a wide base of sciences for its implementation. The range of practitioners employed in any one organisation will tend to reflect the resources allocated, hazards identified, and the prevailing regulatory requirements of the host country. Occupational health is practised by physicians, nurses, safety and risk assessors, and occupational hygienists, sometimes with support from ergonomists, psychologists, toxicologists, and epidemiologists. The competent occupational health practitioner will have some understanding in all these fields but will have an area of special knowledge—for example, the physician will be primarily expert in occupational medicine.

The work of occupational health teams may contribute directly or indirectly to the intrinsic health values of the product (or service). For example, a cement manufacturer might add ferrous sulphate to reduce the likelihood of occupational allergic contact dermatitis in the product users; a hospital may screen and immunise healthcare workers for hepatitis B to prevent occupational acquisition but thereby also prevent iatrogenic disease in patients.

In the course of making recommendations to the UK government on improving access to occupational health support services, the Occupational Health Advisory Committee formed the view that occupational health embraced a range of functions.



The occupational health paradigm

The integration of safety and occupational health is common and many units are described as "occupational health and safety services." In private industry, in particular, environmental responsibilities have also been incorporated, to form a "safety, health, and environment" function, with management board level representation. The amalgamation of these activities can provide a global focus for occupational health, which then needs to engage with a wider public and political agenda.

The interaction between health and work has been a long held paradigm for occupational health that tends to emphasise the adverse effects of work on health and of ill health on capacity for work. It is perhaps time to add a third factor—"wellness"—to acknowledge that in favourable circumstances work contributes to good health (health gain), and healthier workers to better performance (work gain).

The Finnish concept of "maintenance of work ability" refers to a set of measures designed to assist workers to achieve a high level of work capacity in a changing job market over a working

Aims of occupational health services, formulated by the World Health Organization (WHO) and International Labour Organisation (ILO)

Occupational health should aim at: the promotion and maintenance of the highest degree of physical, mental, and social wellbeing of workers in all occupations; the prevention among workers of departures from health caused by their working conditions; the protection of workers in their employment from risks resulting from factors adverse to health; the placing and maintenance of the worker in an occupational environment adapted to his physiological and psychological capabilities; and, to summarise, the adaptation of work to man and of each man to his job

Joint ILO/WHO Committee on Occupational Health First Session (1950) and revised 12th Session (1995)

Mission statement for an occupational health department

"To support and ensure that the company health and safety community assists company management fulfil its responsibilities for employees' health and safety by promoting their physical, mental and social wellbeing, a safe, healthy environment and safe and healthy products"

Macdonald EB. Audit and quality in occupational health. *Occup Med* 1992;42:7-11

Occupational health functions

(a) Evaluating the effect of work on health, whether through sudden injury or through long term exposure to agents with latent effects on health, and the prevention of occupational disease through techniques that include health surveillance, ergonomics, and effective human resource management systems
(b) Assessing the effect of health on work, bearing in mind that good occupational health practice should address the fitness of the task for the worker, not the fitness of the worker for the task alone
(c) Rehabilitation and recovery programmes
(d) Helping the disabled to secure and retain work
(e) Managing work related aspects of illness with potentially multifactorial causes (for example, musculoskeletal disorders, coronary heart disease) and helping workers to make informed choices regarding lifestyle issues

Occupational Health Advisory Committee, 2000

Occupational health services can help an employer

- Comply with legal responsibilities
- Identify hazards and quantify health risks at work
- Implement controls for health risks at work
- Confirm the adequacy of controls through health surveillance
- Select and place workers according to health criteria for particular jobs
- Support employees with a disability
- Ensure fitness for work
- Manage work related disorders
- Control sickness absence and advise on ill-health retirement
- Develop policies relating to health and safety
- Promote health among workers
- Provide training and education in health aspects of employment
- Organise adequate first aid arrangements
- Reduce legal liability
- Design new work processes
- Provide travel health services for work related travel or postings overseas

lifetime. Its application has been found to reduce sickness absence rates and early ill-health retirement.

Occupational health services are in a good position to promote work ability maintenance by:

- Actions directed towards the improvement of employees' physical and mental health, and social wellbeing (by advising the client)
- Actions directed towards competence building, better control of work, encouragement, and motivation (by advising the managers)
- Actions directed to developing the work environment, work processes, and work community that are safe and healthy (by advising the employer).

A work ability index (WAI) has been developed consisting of a questionnaire (translated into many languages) which allows workers to rate their own work ability, track it over the years, and use the score distribution to act as an early warning of decline and guide interventions to improve matters.

The function of occupational health services is to minimise both work and health deficits while maximising health and work gains. To achieve these aims an occupational health service needs:

- to perform a health protection role
- to liaise with treating health professionals
- to undertake active work rehabilitation
- to engage in workplace health promotion and support national health screening programmes
- to provide advisory function to management and workers anticipating the benefits and losses that may arise from changes to work or work practices
- importantly, to monitor its own activities so that meaningful data accrue.

Provision of occupational health

Statutory provision of occupational health is the norm in many European countries, including France, Spain, the Netherlands, Belgium, Portugal, Germany, Denmark, and Greece. Italy provides occupational and environmental health services within the national health service. In the United Kingdom there is no regulation that requires the provision of occupational health services by employers, although all NHS employees should have access to an accredited specialist in occupational medicine.

The NHS is now being encouraged to make its occupational health services available on a commercial basis to small and medium sized enterprises that may not otherwise have access to occupational health. It is, however, more usual for firms to contract private occupational health provision, or to employ occupational health staff.

Employers can seek advice from the Employment Medical Advisory Service of the HSE, but they will usually be directed towards suitable sources of occupational health provision. To implement health and safety legislation effectively, an employer may need the support of a health professional—for example, to perform health surveillance under the Management of Health and Safety at Work Regulations 1999. It might also be prudent to take the advice of an independent health professional at several stages of the employment process, to ensure compliance with disability discrimination legislation and to support other risk management initiatives for the organisation.

The fact that so few private companies use occupational health services is perhaps indicative of their failure to manage certain risks. The HSE promotes a five step process of risk assessment for hazard identification and risk reduction.

Use of health professionals at work

- The following information is contained in research* carried out in 1992†
- In total, 8% of private sector establishments use health professionals to treat or advise about health problems at work. "Health professionals" includes physicians, nurses, and other professions allied to medicine (whether or not they have specialist occupational health (OH) qualification), occupational hygienists, health and safety consultants, and other practitioners with specific OH knowledge or qualifications
- The use of health professionals varies substantially by size of company, with over two thirds (68%) of large employers using professionals, compared with 5% of employers with less than 25 employees. In the private sector, use is highest in manufacturing (14%). The high level of use of health professionals in the public sector means that overall almost half the total workforce are employed by organisations using health professionals

*HSE Contract Research Report 57/1993 on Occupational Health Provision at Work
†There is no more recent data of a comparable nature available

A 2002 survey of UK occupational health provision commissioned by the Health and Safety Executive (HSE) found that among a sample of private companies:

- 15% received services comprising hazard indentification, risk management, and provsion of occupational health and safety information
- 3% received the above plus modification of workplace activities, training, measurement of workplace hazards, and monitoring of trends (mainly larger companies)
- Health was secondary to safety
- Services were mainly provided by private doctors and nurses
- There was rarely a health and safety budget
- Small and medium sized enterprises (SMEs) were generally happy with the occupational health and safety situation within their company

The Health and Safety Executive's five steps to risk assessment

Step 1: Look for the hazards
Step 2: Decide who might be harmed and how
Step 3: Evaluate the risks and decide whether the existing precautions are adequate or whether more should be done
Step 4: Record your findings
Step 5: Review your assessment and revise it if necessary

Access to Medical Reports Act 1988

The Act established a right of access by individuals to reports relating to themselves provided by medical practitioners for employment or insurance purposes and to make provision for related matters. The Act gives patients certain rights. The patient may:

- Refuse to allow a medical report from their treating doctor
- Allow the report to be sent unseen
- See the report during the six month period after it was written
- See the report before it is sent to the employer (a 21 day period is allowed)
- Ask their doctor to change any part of the report which they consider to be wrong or misleading before consenting to its release
- Append their own comments
- Refuse to let the doctor send the report

Health and safety risk management has tended to focus on accidents, yet the cost to employers of workplace injuries and work related illness is estimated (based on the UK Labour Force 1995-6 Survey) to be about £2.5 billion a year (at 1995-6 prices)—about £0.9 billion for injuries and £1.6 billion for illness. Figures from the US Bureau of Labor Statistics report 5 650 100 cases of non-fatal injury or illness in private industry in 2000, with 1 664 000 cases involving days away from work.

Communication

Before any important employment decisions are made, it is in the interests of all parties to gain a full understanding of the facts pertaining to an employee's medical situation. In these circumstances it may be necessary for the occupational health service to request information from the employee's treating doctor. Sometimes information that is not known to the treating doctor is available to the occupational health service. For example, screening procedures may have found a healthcare worker to be infected with hepatitis B virus; health surveillance may have found that a paint sprayer may develop occupational asthma. In such circumstances it is important for the treating doctor to be made aware of these diagnoses with the agreement of the employee.

Communication between an employer and an employee's treating doctor is usually initiated by the occupational health service requesting information from the doctor. Occasionally a request may come directly from a manager or personnel department. The request should be accompanied by appropriate authorisation to disclose medical details to an employer or their medical representative. In the United Kingdom this is under the provisions of the Access to Medical Reports Act 1988.

When asked to provide a report, the corresponding doctor must establish whether the report is intended to go to a doctor retained by the company or to a lay person, such as the employee's manager. A lay person may not fully understand medical jargon, and misinterpretation could give rise to unnecessary concern, to the detriment of the employee.

Reports received by an occupational health department are held in medical confidence, unless the employee has disclosed these or specifically requested that they are disclosed to the employer. The work related implications can be explained to management with advice based on a knowledge of the working environment. It is in everyone's interest (patients, family doctors, hospital doctors, employers, occupational physicians, society as a whole) to get patients back to work as safely and quickly as possible but to prevent their premature return. Rapid and accurate communication is the answer, but the biggest delay occurs when treating doctors fail to answer requests for information from the occupational health service. Delays often cause difficulty to patients, sometimes including financial loss resulting from the inability to work or perform overtime, pending decisions on fitness for work.

Opinions on the part of the treating doctor regarding fitness to work may be unhelpful when these have not been specifically asked for, particularly if the patient is aware of the opinion. For example, a family doctor may consider a "process worker" who is undergoing investigation for syncope as fit to work. The safety of the individual and others in the workplace may be at risk if the doctor is not aware of the duties entailed—for example, working alone in a control room, wearing breathing apparatus, and so on. Doctors may create legal liabilities for themselves in providing opinions when they are

Occupational health reports (The Association of National Health Occupational Physicians, 1996—see Further reading)

Occupational health reports to management must be in writing and include the following:

(a) Details (not clinical details, but information on functional limitations) of any disabilities which may temporarily or permanently affect the ability of the employee to undertake his or her full range of contracted work duties
(b) An estimate of the likely duration of absence or disability
(c) Fitness to undertake the full range of duties, or a limited range of his or her contracted work
(d) Whether and when any further review would be appropriate
(e) Whether an application for retirement on grounds of ill health could be supported (this requires an understanding of the criteria applicable to the scheme)

It is essential that an employee is fully aware of the advice that is being sent to management and the implications of this advice. The employee should be provided with a copy of the advice.

This letter to a manager from a doctor acting as medical adviser to the company contains too many medical terms

Dear Harry,
I saw Mr … . He was well until 19 …, when he had a coronary thrombosis. He made a good recovery from that until about 19 …, when he began to complain of constant ache in his legs, which was worse on exercise. He now has persistent ache in both legs and an exercise limitation of about 200 yards. He recently had an episode of right-sided hemianopia, in which the outside half of the vision in the right eye disappears due to vascular disease of the eye. This is related to his generalised vascular disease as instanced by his coronary thrombosis and by his leg pains. He also complained recently of some shortness of breath and when I examined him I noticed that his heart beat was irregular. This man has quite severe generalised vascular disease and his life expectancy is not good. However, the only problem affecting his ability to work presently is the difficulty in focusing, due to his recent eye problem. This will hopefully improve sufficiently for him to be able to undertake his work in the office, provided no further disaster occurs. I would hope that he can resume employment in three to four weeks. However, as I said previously, the prognosis here is extremely poor. I hope this is of some assistance to you in organising your plans.

Yours sincerely,

Letter is written by a specialist to support a patient's application for a job in a remote tropical location. Knowledge of the medical facilities and the risks of disease in an immunosuppressed person must be considered

Dear Dr …
I am writing in support of Ms J's application to work abroad. In 19 … Ms J had a right leg DVT which was treated with warfarin but one month later she had a pulmonary embolus. Eight months after this, in January, she had an acute illness with fever and a vasculitic rash. A diagnosis of SLE was made and she was treated with prednisolone. In June she had an epileptiform seizure due to cerebral SLE. Glomerulonephritis was diagnosed on renal biopsy in July. The changes were consistent with SLE. She was treated with azathioprine in addition to the prednisolone. She then developed hypertension.

The current situation is that she has heavy proteinuria, indicating active glomerulonephritis; however, she seems clinically well. Her treatment is prednisolone 10 mg daily, azathioprine 100 mg daily, bendrofluazide 5 mg daily, propranolol 320 mg daily, and prazosin, 10 mg twice daily. She will need to continue on long term immunosuppressants but the short term outlook is good, although her renal function is likely to deteriorate in the longer term. Given her fortitude with illness I am sure she would make an excellent field worker for the … project.

not aware of all relevant information and are without sufficient expertise.

The issue of payment for reports can also cause difficulties, and, ideally, fees should be agreed beforehand. Generally speaking, a higher fee is appropriate if the reporting doctor has been asked for an opinion on matters such as fitness for work; simply reporting on a previous diagnosis, and current and proposed treatment does not require exercising of specific judgement. As a matter of good practice and professional courtesy, payments to medical colleagues should be made promptly on the receipt of a report.

Having assessed the individual, the occupational physician may advise restriction of specific duties—for example, for a nursing care assistant with resolving back strain—that they can return to work under the restriction that no manual handling of patients nor of loads greater than 10 kg is undertaken. This still allows the nursing care assistant to perform a wide range of useful functions: assisting with food preparation and feeding, personal care tasks, checking supplies, and social interaction with clients. It is the skill of the manager to accommodate such advice.

A telephone conversation between the treating doctor and the occupational health department may help clarify the options in managing a return to work. Also, under disability discrimination legislation, there may be a duty on the employer to make a "reasonable accommodation" to facilitate work.

In the rare event of a complete disagreement between the occupational physician and the family doctor or specialist on an individual's fitness for work, legal authorities tend towards the occupational physician's opinion. They regard the occupational physician as being in fuller possession of all the facts, both clinical and relating to the actual work to be done, and therefore in a better position to make a balanced and independent judgement.

Ethics and confidentiality

Some doctors are wary of releasing medical details to occupational health professionals, believing that medical confidentiality may be compromised and information given to the employer. This should never happen. All communication between occupational health services and other doctors is held in strict medical confidence. Communication by occupational health services to managers is generally made in broad terms without revealing specific medical details. From a medical report indicating that an employee has angina on exertion, the occupational physician may inform the manager that "Mr. X has a medical condition that prevents him from working in the loading bay and performing other heavy manual work. He should be fit for his other duties as a senior storesman and will be kept under regular review."

It is unnecessary for a manager to be aware of specific medical details, but sometimes it is helpful, with the patient's agreement, for fellow workers to be aware of a medical condition such as epilepsy so that appropriate help can be given (or unhelpful actions avoided).

Some doctors also believe that occupational health services usually act in the interests of the employer, rather than the employee/patient. To behave in such a way is contrary to the ethics of occupational health practice, but this misconception still inhibits useful communication between the specialties. In fact, occupational health physicians and nurses act as independent and objective advisors to the individual *and* to the organisation, hopefully to their mutual benefit.

A "Mushroom worker"—without specific details the circumstances of work may not be obvious from the job title alone

General practitioners and hospital specialists may not be aware of the hazards associated with certain jobs: "blowing down" equipment with an air line, a poor practice that creates airborne dust and its attendant hazards

Cramped working conditions with ergonomic difficulties

Positive patch tests to acrylates in a worker who glued lead flashing onto window units. She had developed an allergic contact dermatitis affecting the hands. In such a situation, two way communication can be beneficial to the patient—a patient may see their general practitioner for hand dermatitis, and liaison with the occupational health department may help identify the cause

Ill-health retirement

Sometimes medical conditions will preclude a return to work because of permanent incapacity for a particular job. Information will often be requested in order to support ill-health retirement, or it may be necessary to explain why an employee's job is to be terminated because of incapacity (where a person has not attended work for an excessive period because of sickness absence, but recovery of fitness is envisaged), the latter being a managerial decision. The pension fund's grounds for ill-health retirement may be explicit and leave little room for clinical opinion, or may be quite open. There is potential for disagreement between the occupational physician and other medical advisers, particularly if restricted duties or redeployment are viable propositions. Ideally, views should be discussed openly and an equitable decision made.

The interface between occupational health and other healthcare providers should therefore be open and two way, initiated either by primary care and hospital services or by occupational health services whenever discussion of patient care in relation to employment could be advantageous.

Audit and monitoring

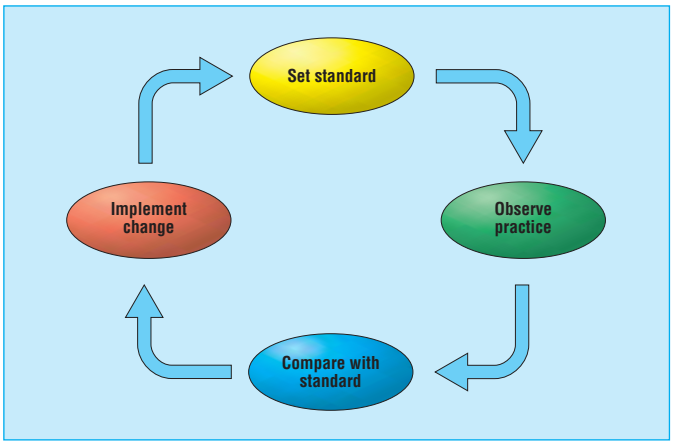
It is important that occupational health practitioners critically evaluate their practice and, through application of the iterative audit cycle, improve the quality, effectiveness, and efficiency of their service. Audit is conventionally divided into structure (resources), process (procedures), and outcome (results). The use of audit should not be confined to clinical matters, and the inclusion of occupational health practitioners from other disciplines—for example, occupational hygiene or safety, will contribute to better services for all.

For a service to report on its activities in a meaningful way there needs to be in place a basic dataset that allows comparison between time periods, different employee groups, or operational divisions. Data that may be appropriate include new appointments, review appointments, health surveillance activity, immunisations, referral reason, type of clinician (doctor or nurse), and diagnosis. This information is invaluable for presentation to management to show changes in activity or areas for which increased funding is needed when making a business case. It will also be useful when discussing issues from the perspective of occupational health in organisational meetings such as health and safety meetings, risk management, and when compiling an annual report or business plan. These data are ideally compiled in a computerised database, either bespoke or a commercially available occupational health software package.

Research

Research is an essential occupational health function. It is only through testing hypotheses that we can advance our knowledge of occupational disease causation, the effectiveness of screening programmes, the benefits of workplace health promotion, quantification of occupational risk, establishment of exposure levels, and the economic impact of occupational injuries and ill health.

Occupational health practitioners may also be faced with ethical difficulties in this field. For example, an organisation may not wish to publicise adverse information about its products or activities. If private companies or national bodies



The audit cycle

Information technology and occupational health

When implementing an occupational health computer system consider:

- The information required from the system and therefore the data entry fields that will be needed
- Data security, in the context of confidentiality and back up in the event of system failure (there are advantages of having the computer server in the organisation's IT department)
- Whether the system is to be "stand alone," networked within a department, or over multiple sites
- Compatability with other organisational systems—for example, personnel or payroll for downloads of starters and leavers, incident reporting systems, sickness absence recording
- Production of reports and database queries
- Maintenance of data quality—that is, that the information recorded accurately represents the information presented
- The use of coding systems if comparisons with other occupational health services may be useful in the future, perhaps for audit, benchmarking, or research.

Practitioners also need to ensure that they meet professional requirements for continuing development. These responsibilities are usually set by professional bodies and it is important that employers recognise that continuing professional development is a necessary component of ongoing competence

An exposure chamber for respiratory challenge studies. The subject is seated inside the metal chamber and gas or vapour is passed through a laminar flow wall into the chamber, inside which spirometry can be performed

are concerned with or participate in research, their influence on what is finally published and intellectual property rights should be formally agreed at the outset. Too often there is reluctance for employers, unions, charities, and government bodies to fund occupational health research. Each seems to feel that the responsibility belongs to one of the other parties. It therefore behoves occupational health practitioners to participate in or act as advocates for occupational health research activities.

The box showing the use of health professionals at work is adapted from the report and recommendations on improving access to Occupational Health Support, Occupational Health Advisory Committee, 2000.

Further reading

- International Labour Organisation. *Technical and ethical guidelines for workers' health surveillance*. Geneva: ILO, 1997. *A discussion of the principles and purpose of health surveillance, including consideration of the ethical implications involved*
- World Health Organization. *Health and Environment in Sustainable Development: Five Years after the Earth Summit: Executive Summary*. Geneva: WHO, 1997. Http://www.who.int/environmental_information/Information_resources/htmdocs/execsum.htm. *This is an executive summary of the WHO report* Health and environment in sustainable development: five years after the earth summit. *It contains extracts from the report, selected figures and tables, and the conclusions in full*
- Macdonald EB. Audit and quality in occupational health. *Occup Med* 1992;42:7-11
- US Bureau of Labor Statistics. http://www.bls.gov/home.htm. *A website providing statistical information and reports relating to the United States*
- Occupational Health Advisory Committee. *Report and Recommendations on Improving Access to Occupational Health Support*. London: HMSO, 2001. *A comprehensive review of occupational health provision and functions, with proposals for improving access for small and medium size employers in the United Kingdom*
- Health and Safety Executive. *Five Steps to Risk Assessment*. Sudbury: HSE Books 1998. (INDG163 (Rev1)). *A short guide to risk assessment, aimed at employers*
- MacDonald E, ed. *Quality and audit in occupational health, Report of the Faculty of Occupational Medicine*. London: Royal College of Physicians, 1995. *This publication describes the essential principles and practical requirements for audit in occupational health practice*
- *The role of occupational health in the process of managing sickness absence*. Association of National Health Occupational Physicians, 1996. *The Association of National Health Occupational Physicians provides a forum for clinical networking, education, and audit and produces guidance for members*
- *Occupational safety and health and employability programmes, practices and experiences*. Luxembourg: European Agency for Safety and Health at Work, 2001. *This report gives an overview of the different types of initiatives in the Member States that aim to increase the employability of workers by using interventions deriving from the field of occupational safety and health*

3 Investigating the workplace

Keith T Palmer, David Coggon

Investigation of the workplace is as central to the practice of occupational medicine as clinical assessment is of the individual patient. It is an essential step in the control of occupational hazards to health. Moreover, by visiting a place of work, a doctor can understand better the demands of a job, and thus give better advice on fitness for employment. Investigations may be prompted in various circumstances.

Circumstances that may prompt investigation of a workplace

- Initial assessment when first taking over care of a workforce or advising an employer
- Introduction of new processes or materials that could be hazardous
- New research indicating that a process or substance is more hazardous than was previously believed
- An occurrence of illness or injury in the workforce that suggests an uncontrolled hazard
- A need to advise on the suitability of work for an employee who is ill or disabled
- Routine review

Direct inspection and the walk through survey

One method of investigation is direct inspection of the workplace. Inspections often take the form of a structured "walk through" survey, although more narrowly targeted approaches may sometimes be appropriate.

When planning a walk through survey an unannounced snap inspection may be revealing, but is practicable only for a health and safety professional who has an established relationship of trust with the employer

Planning

Industrial processes are often complex, and hazards are plentiful. How should a walk through survey be conducted? The arrangements and context are important. The initial visit should be by appointment. Arrangements should be checked before visiting, as a planned visit saves time.

The survey should be structured, but the precise way it is organised is less important and at least three approaches are commonly adopted.

- *Following a process from start to finish*—from raw materials coming in to finished goods going out. What hazards occur at each stage? How should they be controlled? Do the controls actually work? Focusing the assessment on the process helps with basic understanding of the work and its requirements.
- *Auditing a single category of activity or hazard* (such as dusty or noisy procedures or manual handling) wherever it occurs within the organisation. Does the control policy work everywhere, or are there special problems or poor compliance in certain groups of workers or sites? This approach is useful for introducing and monitoring new policies.
- *Detailed inspection site by site*—What are the hazards in this particular site? How are they handled? The inspection moves on only when the geographical unit of interest has been thoroughly inspected. This site focused approach is often appreciated by shop stewards and workers' representatives with local ownership of the problem. They may accompany the inspection and often give insight into working practices and problems not apparent during the visit.

Arranging a walk through survey

- Visit by appointment (at least to begin with)
- Check whether you will:
  - be accompanied by someone with responsibilities for safety
  - see someone who can explain the process
  - have a chance to see representative activities
- Look at documentation on health and safety, such as data sheets, risk assessments, safety policy, accident book
- Do some preliminary research: identify sorts of hazard likely to be encountered and legal standards that are likely to apply
- If visiting because of an individual's complaint, discuss it first with complainant

A hazard represents a potential to cause harm. A risk represents the likelihood of harm. In risk assessment the hazard is put in its correct context

What to cover in a walk through survey

After listing the hazards, it is important to consider who might be exposed and in which jobs, how likely this is under the prevailing circumstances of the work (including any precautions followed), the magnitude of the expected exposures, and their likely impact on health (that is, the risks to health). The aim is to determine whether risks are acceptable, taking into account both the likelihood of an adverse outcome and its seriousness, or whether further control measures are required and, if so, what these should be.

As prevention is better than cure, can the hazard be avoided altogether, or can a safer alternative be used instead? Otherwise, can the process or materials be modified to minimise the problem at source? Can the process be enclosed, or operated remotely? Can fumes be extracted close to the point at which they are generated (local exhaust ventilation)?

Have these ideas been considered before issuing ear defenders, facemasks or other control measures that rely on workers' compliance ("Do not smoke," "Do not chew your fingernails," "Lift as I tell you to")? A realistic strategy should always place more reliance on control of risk at source than on employees' personal behaviour and discipline.

Health and safety professionals use checklists to ensure that all the major types of hazard are considered and to ensure that the control options are fully explored. They seek to verify that these options have been considered in an orderly hierarchy

Simple checklist of control measures

Option	Key questions to ask	Possible controls*
Avoidance or substitution	Does the material have to be used or will a less noxious material do the job?	Try using a safer material if one exists
Material modification	Can the physical or chemical nature of the material be altered?	Is it supplied as granules or paste rather than powder? Can it be used wet?
Process modification	Can equipment, layout, or procedure be adapted to reduce risk?	Can it be enclosed? Can the dust be extracted? If material is poured, tipped, or sieved, can the drop height be lowered?
Work methods	Can safer ways be found to conduct the work? Can it be supervised or monitored? Do workers comply with methods?	Avoid dry sweeping (it creates dust clouds). Be careful with spills. Segregate the work; conduct it out of hours
Personal protective equipment	Have all other options been considered first? Is equipment adequate for purpose? Will workers wear it?	Provision of mask, visor, respirator, or breathing apparatus suitable for intended use

*A dust hazard is used as an example. See also Verma DK, et al. *Occup Environ Med* 2002;59:205-13.

What the survey may find

The purpose of the walk through survey is to be constructively critical. When good practices are discovered these should be warmly acknowledged. Faulty ones arise from ignorance as often as from cutting corners.

In certain workplaces that we have visited, expensive equipment provided to extract noxious fumes from the workers' breathing zone was switched off because of the draught, or directed over an ashtray to extract cigarette smoke rather than the fumes, or obstructed by bags of components and Christmas decorations.

Local exhaust ventilation may be visibly ineffective: the fan may be broken, the tubing disconnected, the direction of air flow across rather than away from the workers' breathing zone. Protective gloves may have holes or be internally contaminated; the rubber seals of ear defenders may be perished with age; and so on. Poor housekeeping may cause health hazards. There may be no system of audit to check that items of control equipment are maintained and effective. Simple commonsense observations, made and recorded systematically, will go a long way towards preventing ill health at work.

The walk through survey may prompt improvements directly or highlight a need for further investigation, such as workplace measurements or a health survey.

Workplace inspection aids understanding of the job demands and risks. This stonemason is exposed to hand transmitted vibration, noise, and silicaceous dust

This industrial process (scabbling) generates a lot of dust. Formal measurements showed that respirable dust and silica levels were several times in excess of those advocated in British standards. The highest exposure arose during sweeping up

Formal assessment of exposures

More formal measurement of exposure may be required if an important hazard exists and the risk is not clearly trivial. Often a specialised technique or sampling strategy will be needed, directed by an occupational hygienist. The UK Health and Safety Executive publishes guidance on methods of measurement and acceptable exposure levels for some physical hazards, such as noise and vibration, and many airborne chemical hazards. In some cases legal standards exist. For some chemicals absorbed through the skin or lungs, exposure can also be assessed by blood or urine tests, and biological action levels have been proposed.

Action after a workplace assessment

The aim in assessing a workplace should be to draw conclusions about the prevailing risks and the adequacy of the controls. But if this is to have a lasting benefit the results must be communicated to senior managers who have the authority to set, fund, and oversee policies in the workplace. A written report is advisable, but a verbal presentation, perhaps at a meeting of the organisation's safety committee, may have more impact, as may a short illustrated slide show. Feedback on the findings of a workplace health survey can make important contributions to the promotion of change and a safer working environment.

Some exposure standards for airborne chemicals

- The UK Health and Safety Executive publishes an annual list of exposure standards (EH40) and also advice on measuring strategies (EH42) and techniques (various EH publications)
- The listed chemicals generally fall into one of two categories. Occupational exposure standards (OES) are prescribed when a level can be specified below which long term exposure is thought not to present a risk to health. In other cases, where the safe level is less certain, a maximum exposure limit (MEL) is specified. This must not be exceeded, and there is a requirement to minimise exposure as far below the MEL as is reasonably practicable
- Other international exposure limits include the threshold limit values (TLVs) published by the American Conference of Governmental Industrial Hygienists (ACGIH) (see http://www.acgih.org)

The worker is exposed to noise during grinding. He should be wearing ear defenders

Frayed electrical cable and homemade plug discovered at a work site

Investigating new occupational hazards

As well as inspecting workplaces to identify and control known hazards, health and safety professionals should be alert to the possibility of previously unrecognised occupational hazards. Suspicions may be aroused in various circumstances. The demonstration and characterisation of new hazards requires scientific research, often using epidemiological methods. The most frequent types of investigation include cohort studies, case-control studies, and cross sectional surveys.

An advantage of epidemiology is that it provides direct information about patterns of disease and levels of risk in humans. However, because of the practical and ethical constraints on research in people, it also has limitations that must be taken into account when results are interpreted. Epidemiological findings should therefore be evaluated in the context of knowledge from other relevant scientific disciplines such as experimental toxicology, biomechanics, and psychology.

Reasons for suspecting an occupational hazard

- Parallels with known hazards—for example, use of a substance that has a similar chemical structure to a known toxin
- Demonstration that a substance or agent has potentially adverse biological activity *in vitro*—for example, mutagenicity in bacteria
- Demonstration that a substance or agent causes toxicity in experimental animals
- Observation of sentinel cases or clusters of disease

Commonly used epidemiological methods

Cohort studies
People exposed to a known or suspected hazard are identified, and their subsequent disease experience is compared with that of a control group who have not been exposed or have been exposed at a lower level. Cohort studies generally provide the most reliable estimates of risk from occupational hazards, but need to be large if the health outcome of interest is rare

Case-control studies
People who have developed a disease are identified, and their earlier exposure to known or suspected causes is compared with that of controls who do not have the disease. Case-control studies are often quicker and more economical to conduct than cohort studies, especially for the investigation of rarer diseases. However, risk estimates tend to be less accurate, particularly if exposures are ascertained from subjects' recall

Cross sectional surveys
A sample of people are assessed over a short period of time to establish their disease experience and exposures. The prevalence of disease is then compared in people with different patterns of exposure. This method is best suited to the investigation of disorders that do not lead people to modify their exposures (which might occur because associated disability makes them unfit for certain types of work). Where a disease causes people to leave a workforce, cross sectional surveys may seriously underestimate the risks associated with exposure

Interpretation of epidemiological findings

In evaluating epidemiological results, consideration must be given to the following factors:

Bias
A systematic tendency to overestimate or underestimate an outcome measure because of a deficiency in the design or execution of a study. For example, in a case-control study assessing exposures by questioning participants, affected persons might tend to recall exposures better than controls (because they are more motivated). The effect would be to spuriously exaggerate any association between exposure and disease

Chance
The people included in a study may be unrepresentative simply by chance, leading to errors in outcome measures. The scope for such errors can be quantified statistically through calculation of confidence intervals. Generally, the larger the sample of people studied, the lower the potential for chance error

Confounding
This occurs when a hazard under study is associated with another factor that independently influences the risk of disease. For example, an occupational group might have high rates of lung cancer not because of the chemical with which they worked, but because they smoked more heavily than the average person (that is, exposure to the chemical was associated with heavier smoking)

Assessment of disease clusters

One starting point for investigation of a workplace may be the observation of a disease cluster. A disease cluster is an excess incidence in a defined population, such as a workforce, over a relatively short period (less than a day for acute complaints such as diarrhoea to several years for cancer).

Apparent clusters are not uncommon in occupational populations, and investigation sometimes leads to the recognition of new hazards. For example, on the one hand, the link between nickel refining and nasal cancer was first discovered when two cases occurred at the same factory within a year. On the other hand, excessive investigation of random clusters wastes resources. The extent to which a cluster is investigated depends on the level of suspicion of an underlying hazard and the anxiety that it is generating in the workforce. A staged approach is recommended.

A cluster of wheezing and rhinitis occurred on this prawn processing line. High pressure hoses (used to free the prawns from the shells) had created aerosols containing crustacean protein

Is there a true cluster?

The first step is to specify the disease and time period of interest and to confirm the diagnoses of the index cases that prompted concern. Sometimes no further action is needed. Of three cases of brain cancer, two might turn out to be secondary tumours from different primary sites. If suspicion remains, it is worth searching for further cases. Often, the number of identified cases is clearly excessive, but if there is doubt, crude comparison with routinely collected statistics such as of cancer registration or mortality should establish whether the cluster really is remarkable.

Further steps

If a raised incidence is confirmed, the next step is to find out what the affected workers have in common. Do they work in the same job or building, and do they share exposure to the same substances? If so, what is known about the risks associated with their shared activities and exposures? This information may come from published reports or manufacturers' data sheets. Scientific articles should also be searched to identify known and suspected causes of the disease of interest. Could any of these be responsible for the cluster?

Getting help

At this stage the cause of the cluster may have been identified, or suspicions sufficiently allayed to rule out further investigation. If concerns remain it may be necessary to carry out a more formal epidemiological investigation to assess more precisely the size of the cluster and its relation to work. Help with such studies can often be obtained from academic departments of occupational medicine. Also, patients may need to be referred to specialist centres for investigations such as dermatological patch testing or bronchial challenge.

Stages in investigating occupational clusters of disease

1. Specify disease and time period of interest. Confirm diagnoses of index cases
2. Search for further cases. Is the observed number of cases excessive?
3. What do affected workers have in common? Do their shared exposures carry known or suspected risks?
4. What is known about the causes of the disease?
5. Further investigation: epidemiology and clinical investigation

Hazards controlled

Over the years, investigation of workplaces has made a major contribution to public health through the identification and control of occupational hazards, and improved placement and rehabilitation of workers with illness or disability. Although some types of investigation need special technical expertise, all health and safety professionals should be familiar with the principles, and capable of inspecting and forming a preliminary assessment of working environments.

Some important occupational hazards that have been identified and controlled through investigation of workplaces

Hazard	Control measures
• Bladder cancer from aromatic amines in dyestuffs and rubber industries	• Substitution of the chemicals with non-carcinogenic alternatives
• Lung cancer and mesothelioma from asbestos	• Substitution by less hazardous materials such as manmade mineral fibres; dust control and personal protective equipment in asbestos removal
• Coal workers' pneumoconiosis from dust in mines	• Dust suppression by water spraying
• Occupational deafness from exposure to noise	• Substitution or enclosure of noisy processes; exclusion zones; personal protective equipment

Further reading

- Olsen J, Merletti F, Snashall D, Vuylsteek K. *Searching for causes of work-related disease: an introduction to epidemiology at the worksite.* Oxford: Oxford University Press, 1991
- Pittom A. Principles of workplace inspection. In: Howard JK, Tyrer FH, eds. *Textbook of occupational medicine.* Edinburgh: Churchill Livingstone, 1987:91-106.
  These two references describe in greater detail the process of workplace inspection
- Health and Safety Executive. *Five steps to risk assessment.* Sudbury: HSE Books, 1998. (INDG163 (Rev 1)). *This free leaflet suggests a simple five point plan for assessing the risks in a workplace*
- Health and Safety Executive. *Occupational exposure limits.* Sudbury: HSE Books, 2000. (Guidance Note EH40/00). *This HSE publication, which is updated annually, provides guidance on the permissible limits for exposure to a number of chemicals*
- Health and Safety Executive. *Monitoring strategies for toxic substances.* Sudbury: HSE Books, 1999. *Assessment of exposure requires a strategy of representative sampling: this booklet explains the required approach*
- Coggon D, Rose G, Barker DJP. *Epidemiology for the uninitiated,* 4th ed. London: BMJ Publishing Group, 2003. *This short primer provides a useful introduction to epidemiological methods and principles*
- Harrington JM, Gill FS, Aw TC, Gardner K. *Occupational health pocket consultant,* 4th ed. Oxford: Blackwell Science, 1998. *This concise textbook explains how to make and interpret measurements of the working environment. It also provides a very good overview of other topics in occupational medicine*
- Verma DK, Purdham JT, Roels HA. Translating evidence of occupational conditions into strategies for prevention. *Occup Environ Med* 2002;59:205-13. *This review illustrates how evidence on risks and control measures can be used to develop effective preventive strategies in the workplace*

4 Fitness for work

William Davies

Assessments of fitness for work can be important for job applicants, employees, and employers. Unfitness because of an acute illness is normally self evident and uncontentious, but assessing other cases may not be straightforward and can have serious financial and legal implications for those concerned. Commercial viability, efficiency, and legal responsibilities lie behind the fitness standards required by employers, and it may be legitimate to discriminate against people with medical conditions on these grounds. Unnecessary discrimination, however, is counterproductive and may be costly if legislation is breached. The Disability Discrimination Act 1995 makes it unlawful for employers of 15 or more staff (all employers from 2004) to discriminate without justification against those with disability as defined by the Act. The Employment Rights Act 1996 requires procedural standards and fairness before any decision to dismiss an employee. Fortunately, balancing these often complex socioeconomic and legal issues to achieve a sustainable decision on fitness is not primarily a medical responsibility. Doctors do, however, have responsibilities to assess the relevant facts competently and to assist with the decision making process.

Implications of fitness assessments

- Security of employment
- Rejection at recruitment
- Justifiable or unfair discrimination
- Retirement because of ill health
- Termination of contract
- Claim for disability discrimination
- Claim for unfair dismissal
- Employment tribunals
- Medical appeal
- Civil litigation for personal injury
- Criminal prosecution for breach of health and safety legislation
- Professional liability
- Pension entitlements
- Benefit claims

Basic principles and responsibilities

Staying on track

This chapter deals with assessing fitness for "identified employment." To avoid confusion with related issues, the following points should be noted at the outset:

- Fitness for work in relation to ill health retirement benefits will depend on the specific provisions of the pension scheme. Pivotal issues that frequently arise are the interpretations that should be given to incapacity and to permanence, and whether fitness relates to current employment or all work. General guidance has been issued and specific guidelines for all UK public sector schemes should now be available following the recommendations of a HM Treasury report in 2000
- The Disability Discrimination Act 1995 has encouraged good medical practice in assessing and deciding on fitness for work by requiring individual and competent assessments, and by obliging employers to be more accommodating to those covered by the legislation
- Key health and safety concepts—hazard, risk, negligible risk, and competence—apply to assessing fitness for work and should be clearly understood
- Rehabilitation back to work and an emphasis on capability rather than limitations are now central themes of legislation, guidance, and government policies on health and safety and occupational health.

Medical responsibilities

Doctors' responsibilities vary according to their role. General practitioners and hospital doctors acting as certifying medical practitioners have direct responsibilities to their patients to provide statutory evidence of advice given about fitness for the patient's regular occupation. Such doctors also have an obligation to provide related information to a medical officer working for the Department for Work and Pensions.

Basic principles and responsibilities—when fitness assessments may be required

- Before employment, placement, or redeployment
- Routine surveillance in safety critical jobs
- During or after sickness absence
- To identify adjustment needs
- When attendance or performance issues arise
- If health and safety concerns arise
- To examine ill health retirement issues
- If required by statute
- Benefit assessment—for example, the "own occupation test" administered by the Department for Work and Pensions (DWP)

The personal capability assessment is the medical assessment used to determine if a person is eligible for state incapacity benefit. It does not consider fitness for a specific type of employment but assesses general functional ability in relation to everyday physical and mental activities. Decision makers within the Department for Work and Pensions who apply the test will take advice from a specially trained doctor approved for the purpose by the Secretary of State

Medical responsibilities

General and hospital practitioners

To patient

- Act in patient's best health interests
- Provide advice on fitness for regular occupation
- Consider clinical management that would support employment wherever clinically reasonable
- Provide patient with statutory forms (for example, Med 3) recording the advice given

To Department for Work and Pensions (DWP)

- Supply on request relevant clinical information to a medical officer

Detailed advice for general and hospital practitioners on DWP issues is available in the guide IB204 (March 2000) and from regional Medical Services Centres

Occupational health practitioners

To patient

- Act in patient's best health interests
- Consider clinical management that would support employment wherever clinically reasonable

To employer

- Assess functional ability and occupational risks
- Make recommendations on fitness in accordance with valid predetermined standards
- Provide information and advice that enables management to make an informed decision on compatibility of subject with employer's requirements and legal responsibilities

Detailed advice on medical responsibilities of occupational health practitioners is available in *Fitness for Work. The Medical Aspects* or from accredited specialists in occupational medicine

To society and the general public

- In certain circumstances public interest will override any duty to the individual patient or employer—for example, a surgeon infected with hepatitis B who continues to work in a way that puts patients at risk

Occupational health practitioners have direct responsibilities to the employee or job applicant and the employer. Both groups have a responsibility to society.

These groups may take different approaches but have important common ground. If patients, employees, and job applicants are to be treated fairly, every medical opinion on their fitness for a job should be based on a competent assessment of relevant factors, and should satisfy the same basic criteria. Patients' interests will be best served when there is clear understanding, due consultation, and, as far as possible, agreement between doctors.

Key principles in practice

The first principle in the table opposite establishes three basic criteria for fitness: attendance and performance, health and safety risk to others, and health and safety risk to self. In this context, "without risk" reflects a fundamental ethical concept of occupational medicine that limits medical discretion. Doctors should not presume to decide for others that risks are acceptable; employers must take this responsibility, and they require medical advice and information on the nature and extent of risk to make informed decisions.

The second principle means that an appraisal of the subject's medical condition and functional ability, together with a review of the relevant occupational considerations, should provide an empirical assessment of ability and risk. This assessment may be judged against the required fitness criteria to determine what the outcome should be.

The third, fourth, and fifth principles point to the potential there may be for preventing or controlling risk, and for accommodating the needs of people with disabilities or medical conditions. Such measures may justify a conditional recommendation of fitness.

The sixth principle means that technically all decisions on fitness rest with the employer. This is because the employer determines what is required of the employment and ultimately carries responsibility for the risks.

Framework for assessing fitness for work

The terms fitness and incapacity are open to interpretation, and responsibilities for assessing and deciding on fitness issues span medical and management disciplines. A systematic

Key principles of assessing fitness for work

1. The primary purpose of the medical assessment of fitness to work is to ensure that the subject is fit to perform the task required effectively and without risk to the subject's or others' health and safety
2. The subject's fitness should be interpreted in functional terms and in the context of the job requirements
3. Employers have a duty to ensure, so far as is reasonably practicable, the health, safety, and welfare of all their employees and others who may be affected (Health and Safety at Work etc. Act 1974)
4. Legal duties of reasonable adjustment and non-discrimination in employment are imposed by the Disability Discrimination Act 1995
5. Good employment practice involves due consideration of the needs of all job applicants and employees with disabilities or medical conditions (Employment Rights Act 1996)
6. It is ultimately the employer's responsibility to set the objectives for attendance and performance, and to ensure compliance with the law on health and safety and employment

Framework for assessing fitness for work

Stage 1—Workplace assessment of ability and risk

Step 1: Assess medical condition and functional capacity
Step 2: Consider occupational factors
Step 3: Explore enabling options

Stage 2—Relate Stage 1 findings to fitness criteria

Step 4: Identify any attendance or performance limitations
Step 5: Identify nature and extent of any risks to others
Step 6: Identify nature and extent of any risks to self

Stage 3—Report on outcome in suitable terms

Step 7: Confirm fitness or unfitness
Step 8: Present assessment conclusions if 7 not possible
Step 9: Provide supplementary advice to 8 if appropriate

approach is required to ensure consistency and to avoid confusion of roles.

The framework is based on the key principles and relevant legal provisions. There are three stages and up to nine logical steps. In simple cases where no medical conditions apply, steps 1, 2, and 7 should suffice. In other cases, seven, eight, or all nine steps may be required.

Reporting the outcome

When the parameters of the fitness criteria are defined and the assessment clearly satisfies or fails to satisfy the employer's requirements and responsibilities, a confirmation of fitness or unfitness can be made (see green columns in the desktop aid on page 23).

When the parameters of the fitness criteria are uncertain (when the employer's requirements and responsibilities cannot be predetermined or presumed) the conclusions of the assessment should be made clear to the employer. In addition, an opinion on the reasonableness of any enabling options identified or the case for employment or continued employment may be given as supplementary advice (see red columns in desktop aid).

It should be noted, however, that supplementary advice offered under step 9 above relates to management rather than medical issues, and should be qualified accordingly. All reports should comply with professional standards on disclosure and consent.

Assessment of ability and risk

Medical functional appraisal

Doctors should always have a basic knowledge of the job's demands and working environment before undertaking a medical functional appraisal so that the extent and emphasis of the appraisal may be tailored accordingly. Any medical conditions that could pose a risk to the subject's or others' health and safety, or that could affect attendance and performance, should be identified and evaluated.

A suitably constructed questionnaire is the simplest form of assessment; for pre-employment screening, a questionnaire or health declaration will be sufficient to permit medical clearance in many categories of employment. Some occupations have statutory standards (for example, in the United Kingdom, there are statutory medical standards for seafarers), and appraisals must include measuring necessary factors. Others have standards set by authoritative recommendations or guidance (for example, the Health Advisory Committee of the UK Offshore Operators Association has drawn up guidelines on the medical standards for offshore work).

If no guidance exists, doctors must judge how extensive the assessment should be by taking account of the nature of any medical conditions identified, the type of work, and the reasons for management's request for medical advice.

Occupational considerations

In straightforward cases a medical functional appraisal and the doctor's existing knowledge of the job demands and working environment may be sufficient for a confirmation of fitness. However, a closer look at occupational factors is often needed to determine the precise requirements of the job, the subject's real abilities in a working environment, the nature of any hazards, and the probability of harm occurring (the actual risk in the workplace).

Medical functional appraisal

History and examination
- Pre-employment questionnaire or health declaration
- Health interview, occupationally relevant direct questions
- Physical examination focusing on job requirements

Functionally specific questionnaires
- Respiratory (MRC questionnaire)
- Pre-audiometry

Consultation and research
- Details from general practitioner and medical specialist, under Access to Medical Reports Act 1988 or non-UK equivalent
- Details from other specialists such as psychologists or audiologists
- Advice or second opinion from specialist occupational physician
- Advice or second opinion from independent specialists such as cardiologists or neurologists
- Clinical guidelines and evidenced based reviews
- Texts, journals, and research

Work related tests and investigations

Perceptual tests
- Snellen chart: special visual standards may be required for certain occupations such as aircraft pilots, seafarers, and vocational drivers
- Colour vision tests such as Ishihara plates or City University test, or matching tests, may be necessary if normal colour vision is essential—for example, for some jobs in transport, navigation, and the armed services
- Voice tests
- Audiometry: occupations such as the armed services, police, and fire service may have specific standards

Functional tests
- Lung function tests (for example, UK regulations require fire service employees to have their respiratory parameters measured before employment
- Dynamic or static strength tests
- Physical endurance and aerobic capacity (for example, the fire service or commercial divers)
- Step test
- Bicycle ergometer

Diagnostic (health on work)
- Exercise electrocardiography: needed—for example, for vocational drivers and offshore workers
- Drug and alcohol tests may be a requirement in certain safety critical industries

Diagnostic (work on health)
- Haematology, biochemistry, and urine analysis: UK commercial divers will have full blood count and haemoglobin S assessed before employment
- Radiographs: long bone radiographs are a requirement before employment for saturation diving in the United Kingdom

- A subject may be able to show satisfactory ability in a job simulation exercise despite a physical impairment that might have affected fitness—for example, a work related test of manual dexterity for an assembly line worker with some functional loss resulting from a hand injury
- In teaching, health care, and many other occupations, the perceived hazards of epilepsy are often found to be negligible when the potential for harm to others is properly assessed
- If diabetes is well controlled, the risk of injury from hypoglycaemia may be found to be very remote when the true frequency and duration of hazardous situations are taken into account.

Enabling options

A subject's potential fitness often depends on intervention. There may be unexplored treatments that can be provided. Rehabilitative support may be needed to achieve or speed recovery. Employers can make reasonable adjustments, temporary or permanent, to meet the needs of people with medical conditions. Prevention and control measures can reduce or eliminate health and safety risks that would otherwise prohibit a recommendation of fitness.

- Unexplored treatments that are often identified during assessments include physiotherapy, anxiety management, and psychotherapy
- A tailored, stepwise rehabilitative programme can make the prospect of returning to work after serious illness less daunting and may be vital for recovery from anxiety, depression, occupational stress, and other demotivating conditions
- Modifying a job specification may allow a recommendation of fitness with minimal inconvenience to the employer(for example, removing the requirement to undertake occasional lifting for an arthritic subject)
- Substituting a sensitising or irritant product may, with other sensible precautions, enable an employee with asthma or eczema to continue working as—for example, a paint sprayer or cleaner.

These measures may be applicable under the Health and Safety at Work etc. Act 1974. The Disability Discrimination Act 1995 may also require reasonable adjustments to be made. Even if intervention is not obligatory, employers may recognise the benefits of positive action. Doctors should therefore always bear these options in mind, as it may be possible to give a conditional recommendation of fitness that the employer would be willing to accommodate.

Fitness criteria in difficult cases

This approach should produce a reliable opinion in most cases, but further steps may be needed if the criteria for fitness for work are uncertain. In a fitness assessment this may occur with one, two, or all three of the criteria. Dealing with the issues in turn is advisable.

Attendance and performance

The possible impact of a medical condition on a subject's ability to meet required levels of attendance and performance is a major source of employers' requests for medical opinion. When asked by an employer about an employee's performance and attendance capabilities, the doctor's responsibility is to give the most accurate opinion that the circumstances allow. Conclusions and advice should be as positive as possible but without misrepresenting the facts, and should be discussed with the subject. This should help motivation and may improve recovery.

Occupational considerations

Ability in the workplace—consider actual effect of physical or medical condition on performance
- Confirm job requirements such as perception, mobility, strength, and endurance
- Ask employee what the work entails
- Review job description or inspect worksite
- Perform field tests of specific abilities or structured job simulation exercises
- Consider trial of employment with feedback from management

Nature of hazards—consider interaction of occupational factors and medical condition
- Harm from:
  - demands (heart attack, back strain, prolapsed disc, repetitive strain injury)
  - exposures (asthma, dermatitis, hearing loss)
  - situations (seizure, trauma, accidents)
  - infections (food handling, surgical procedures)
- How much harm is likely (temporary, permanent, minor, major, fatal)?
- Who may be affected (self, colleagues, clients, public)?

Extent of risk—focus on facts and avoid presumption
- Question employee on relevant details
- Obtain management report on material facts
- Examine documentation such as exposure records, accident reports, etc.
- Observe work, workplace, and working practices
- Identify frequencies and duration of hazardous exposures or situations
- Request technical data from hygienist, ergonomist, etc. if required
- Review relevant literature, journals, and research

Enabling options

Unexplored treatments
- Drug treatment or surgery
- Physiotherapy or occupational therapy
- Counselling or psychotherapy

Rehabilitative measures
- Graded resumption of responsibilities
- Refamiliarisation training
- Temporary reduction of workload
- Management appraisal or progress reports
- Scheduled or self requested medical reviews

Reasonable adjustments
- Modification of duties or working hours
- Redeployment to existing vacancy
- Modifying or providing equipment
- Time off for rehabilitation or treatment
- Providing supervision

Risk prevention and control
- Elimination or substitution of hazard
- Implementation of methods to reduce worker exposure to hazards
- Personal protection or immunisation
- Information, instruction, and training
- Health and medical surveillance

- Employers do not like open ended statements such as "Unfit; review in three months;" they prefer uncertainties to be expressed as probabilities: "Mr Smith has been incapacitated but is progressing well and is likely to become fit to return to work within four weeks"
- The doctor should may need to ask management for an appraisal of capabilities before making definitive conclusions on the relevance of medical factors: "I will therefore require a management report on her progress after week 6 of the rehabilitation programme"
- In cases of prolonged sickness absence, the doctor should not be pressured into recommending ill health retirement for doubtful reasons: "Mr Green is likely to remain unfit for the foreseeable future, but there are not sufficient grounds for ill health retirement under the pensions scheme."

If social or motivational factors are evident, discuss these with the subject, and advise management accordingly: "Mrs Jones' incapacitation is due to family commitments that are likely to continue for the foreseeable future. She realises that her employment could be at risk and would welcome an opportunity to discuss her situation with management."

Health and safety risk to others

Employers have a legal duty to ensure the health and safety of employees and the public. In principle, the doctor identifies hazards and quantifies any risks; management decides on a subject's fitness on the basis of the medical conclusions and advice. In practice, however, doctors confirm fitness when there is no risk, and unfitness if there are clearly unacceptable risks. For the many cases that lie in between, there may be confusion as to whether it is a management or medical responsibility to decide on fitness. A pragmatic approach is suggested.

For negligible risk, the doctor may advise that the subject be considered fit provided that the judgement of negligible risk is made objectively, is based on a competent risk assessment, and that the employer applies all reasonably practicable precautions.

For greater than negligible risk, the doctor should define the type of hazard and extent of risk as clearly as possible to enable management to make an informed decision.

Advice from a specialist occupational physician may be required to confirm the competence of the risk assessment or to assist management on acceptability.

Health and safety risk to self

The principles of assessing risk to others applies here, but medical advice can go further. In some cases employment may pose a risk of ill health but the employer is satisfied that everything possible has been done to prevent or reduce risks (for example, the risk of relapse in a teacher with a history of work related anxiety depressive disorder). To advise that in such cases the subject should always be deemed unfit because of a risk of work related illness is unrealistic. The benefits of employment for the subject, and possibly their employer, may considerably outweigh the risks. On the other hand, there could be issues of liability for both employer and doctor if the risks are overlooked.

The parameters of the fitness criteria may be uncertain when:

- Attendance or performance limitations resulting from a medical condition are identified, but the employer's willingness to accommodate them cannot be prejudged
- Health and safety risks to others exist, but they seem remote enough to ignore
- Health and safety risks to self are identified, but they do not seem to justify a recommendation of unfitness.

Reasonable adjustments under the Disability Discrimination Act (DDA) 1995 (see chapter 5)

- Reasonable adjustments are essentially any steps relating to arrangements and premises that are reasonable for an employer to take in all the circumstances to prevent the disabled person being at a disadvantage. Many of the enabling options listed above fall within this definition
- The DDA Code of Practice expands on examples given in the Act and provides guidance on the reasonableness of adjustments (Paragraphs 4.12-4.48)
- A comprehensive series of practical briefing guides on the DDA is published by the Employers Forum on Disability, Nutmeg House, 60 Gainsford Street, London SE1 2NY

Data sources for standards of fitness (see Further reading)

Key publications
For drivers, pilots, food handlers, and many other occupations: Cox et al., DVLA

General guidance
Health and Safety Executive

Professional associations
ALAMA (Association of Local Authority Medical Advisors) for firefighters, police, teachers, etc.
ANHOPS (Association of National Health Occupational Physicians) for healthcare professions

Government departments
Department for Education and Skills for teachers

Statute
Seafarers: Merchant Navy Shipping (Medical examination) Regulations 1983. Revised in 1998 [Merchant shipping notice MSN 1712(M)]

The autonomy of the subject must be reconciled with the needs and responsibilities of the employer. Legal precedent does not provide clear guidance on how this should be done; the issues are complex and the implications serious. A rational basis for providing helpful medical advice includes a full discussion of the prognosis with the subject to determine where the balance of benefits and risks lies

- If the subject thinks the benefits outweigh the risks and the doctor agrees, advice should be given in support of employment, provided that the assessment and the judgement of balance between benefit and risk have been competently undertaken
- If the subject thinks the benefits outweigh the risks but the doctor cannot agree, consider seeking a second opinion from a specialist occupational physician before providing management with definitive advice
- If the subject thinks the risks outweigh the benefits and the doctor agrees, early retirement should be considered
- If the subject thinks the risks outweigh the benefits when the hazard and risk seem disproportionately low, then motivational factors (such as a common law claim or ill health retirement incentives) may be relevant. If so, the doctor should proceed cautiously and consider obtaining a second opinion from a specialist occupational physician.

The conclusions should be presented to management in context, indicating the nature of the hazard, the extent of risk, and strength of medical consensus. This will enable the employer to discharge his or her responsibility in a complex area with the benefit of such medical support as the circumstances allow.

Definitive opinion

The conclusions, recommendations, and advice outlined above are valid only for the specific fitness criterion considered. In each case, the outcomes of all three criteria should be consolidated to provide an all embracing definitive report. The desktop aid includes a synopsis of the outcomes commonly encountered and may be adapted as a classification guide for audit purposes.

Further reading

- Cox RAF, Edwards FC, Palmer K. *Fitness for work. The medical aspects*, 3rd ed. Oxford: Oxford Medical Publications, 2000. *A comprehensive text on medical issues covering background issues, all medical systems and specific occupations*
- Benefits Agency, Department of Social Security. *A guide for registered medical practitioners.* Revised with effect from April 2000. (IB204) *Medical evidence for statutory sick pay, statutory maternity pay, and social security incapacity benefit purposes.* Supplemented in April 2002 by chief Medical Officer's Bulletin and Desk aid. Publications available on www.dwa.gov.uk/medical. *Detailed practical reference, related website has evidenced based information and guidance*
- Drivers Medical Unit, DVLA. *At a glance guide to current medical standards of fitness to drive. March 2001.* Available on www.dvla.gov.uk/ataglance/content.htm. *Regularly updated prescriptive standards for wide range of medical conditions*
- Royal College of General Practitioners. *Clinical guidelines for the management of acute back pain.* 1997, updated 1999. Faculty of Occupational Medicine. *Occupational heath guidelines for the management of low back pain evidence review and recommendations,* March 2000. *Two complementary guides providing a positive practical approach to medical management and rehabilitation*
- Health and Safety Executive. *Your patients and their work, an introduction to occupational health for family doctors.* Bootle: HSE Books, 1992. *Simple general guide*
- Health and Safety Executive. *Pre-employment screening.* London: HMSO, 1982. (Guidance note MS20.) *Reviews main principles; would benefit from updating*
- ALAMA, ANHOPS, at Society of Occupational Medicine, 6 St Andrews Place Regents Park London. *Membership gives access to website facilities and current guidance and on firefighters, police, and healthcare professionals*
- DfEE. *Fitness to teach. Occupational health guidance for the training and employment of teachers. The physical and mental fitness to teach of teachers and of entrants to initial teacher training.* London: HMSO, 2000. *Focused, up to date, working guidance supported by well balanced complementary guide for employers and managers*

Desktop aid—Framework for assessing fitness for work

Assessment of ability and risk
- Medical-functional appraisal
- Occupational considerations
- Enabling options

+

Fitness criteria
- Attendance and performance
- Health and safety risk to others
- Health and safety risk to self

=

Outcome
- Confirm fit or unfit
- Report conclusions
- Offer advice

Applying fitness criteria—Synopsis of outcomes

Attendance and performance

A	B	C	D	E
Subject's condition compatible with required levels of attendance and performance	**Attendance or performance limitations due to medical conditions or disabilities identified but likely to resolve**	**Attendance or performance limitations due to medical conditions or disabilities identified and likely to remain for foreseeable future**	**Subject's performance and capabilities cannot be determined by medical assessment alone**	**Subject's condition clearly incompatible with requirements of post and likely to remain so**
	(a) in foreseeable future because of anticipated recovery or (b) if certain enabling options can be accommodated (such as treatment, rehabilitation, reasonable adjustments, or risk prevention)	Do not overlook social or motivational factors that may be relevant. Discuss implications with subject. If necessary seek advice*	Feedback on performance is required to identify possible impact of medical conditions	Help subject come to terms with implications such as ill health retirement, termination of contract, redeployment (if available), or rejection (at pre-employment stage)
Confirm fit	*Report conclusions indicating* *(a) likely timescale and/or* *(b) relevance of enabling options* *Reviews as necessary*	*Report conclusions* *Review as necessary*	*Report on medical issues and identify need for management appraisal/ feedback. Review as necessary*	*Confirm likely to remain unfit*

Health and safety risk to others

F	G	H	I	J
No risk to others	**Risk identified but preventable**	**Negligible risk**	**Risk greater than negligible but may be acceptable**	**Risk to others clearly unacceptable and likely to remain so**
	Identify and pursue relevant enabling options such as treatment, rehabilitation, reasonable adjustment, or risk prevention	Ensure judgment of negligible risk is made objectively and based on competent assessment (if unsure seek advice*) and that management applies all reasonably practicable precautions	Inform management of nature and extent of risk as clearly as possible. Specialist occupational physician may be able to help management in deciding on acceptability*	Help subject come to terms with implications such as ill health retirement, termination of contract, redeployment (if available), or rejection (at pre-employment stage)
Confirm fit	Report conclusions and *advise fit (subject to specified conditions)*	Report conclusions and *advise fit (subject to specified conditions)* *Review if circumstances change*	Report conclusions *advise risk cannot be dismissed as negligible and that acceptability is for management to consider*	*Confirm likely to remain unfit*

Health and safety risk to self

K	L	M	N	O
No risk to self	**Risk identified but preventable**	**Risks identified which subject thinks are outweighed by benefits**	**Risks identified which subject thinks outweigh benefits**	**Risk to self clearly unacceptable and likely to remain so**
	Identify and pursue relevant enabling options such as treatment, rehabilitation, reasonable adjustment, or risk prevention	*If doctor agrees*—Ensure assessment and judgment of balance between risk and benefit have been competently undertaken (if unsure seek advice*) *If doctor disagrees*—Consider obtaining second opinion before advising	*If doctor agrees*—Consider early retirement *If doctor disagrees*—If risks seem disproportionately low consider relevance of motivational factors (such as common law claim or ill health retirement incentives) If present proceed cautiously and consider obtaining second opinion*	Help subject come to terms with implications such as ill health retirement, termination of contract, redeployment (if available), or rejection (at pre-employment stage)
Confirm fit	Report conclusions and *advise fit (subject to specified conditions)*	*Report conclusions with supplementary advice as appropriate*	*Report conclusions with supplementary advice as appropriate*	*Confirm likely to remain unfit*

Definitive opinion

The confirmations, conclusions, and advice outlined abvoe are valid only for the specific fitness criterion addressed.
In each case the outcomes of all three criteria should be consolidated to provide an all embracing definitive report

*Advice and second opinions should be obtained from doctors with training and expertise to provide proper assistance.
Specialist qualifications for occupational physicians in the UK (MFOM, FFOM) are awarded by the Faculty of Occupational Medicine of the Royal College of Physicians.

5 Legal aspects

Martyn Davidson

Society has become increasingly litigious in recent years, and the modern "blame culture" has encouraged a tendency to look for fault whenever there is harm. In all areas of medicine this has led to increased awareness of the legal process, and occasionally defensive medicine. Employment law and rights based legislation following European Union initiatives have expanded as a result of enlightened social policies, extension of a single European market, and environmental protection. Because employment and rights based legislation affect the worker and the workplace, the occupational health (OH) practitioner needs to understand the legal provisions and the framework in which they operate.

Health and safety legislation aims to prevent the workforce being injured or made ill by their work. Employers have considerable duties, including duties relating to the general public, and the role of the OH practitioner is to advise on steps to achieve compliance. An understanding of the principles is essential, and these are covered here with reference predominantly to English law. Employees also have corresponding duties to take "reasonable care" for their own safety and that of others, and to cooperate with appropriate procedures.

The OH practitioner will become involved in employment law when medical advice is needed, and it is essential that the basics are understood.

The legal framework defining the duty towards the health of the workforce was established in the 19th century. Although prompted by humanitarian concerns, these legal developments were the pragmatic result of the concerns of industry—the toll of premature death and disability threatened the supply of healthy workers required to increase productivity. Reproduced with permission from Hulton Deutsch

Ethics

The position of the OH professional

Physicians are primarily bound by the codes of their profession and in the United Kingdom they are accountable to the General Medical Council for their behaviour. For OH nurses the corresponding body is the Nursing and Midwifery Council. Difficulties sometimes arise because the OH practitioner is often an employee of the company requesting advice. The company may feel that the practitioner's contract of employment overrides professional codes. This is not so, and employers cannot insist on contractual terms that would require a physician or nurse to breach professional codes. If such terms existed, they would be difficult, if not impossible, to enforce.

Confidentiality

The duty of confidentiality applies as it does to any physician or nurse. This includes the safeguarding of all medical information, records, and results. The legal basis of the duty of confidentiality remains unclear, however, and the duty is ultimately relative rather than absolute. Material should be regarded as confidential if it has been obtained in circumstances which would indicate that this was the intention. Circumstances can arise in any medical specialty in which disclosure may be necessary; in such cases the clinician will be expected to justify his or her action, before a court if necessary.

OH practitioners may sometimes feel that they are not in a traditional nurse/doctor-patient relationship when they are acting on behalf of a third party. This might be the case with respect to a job applicant whom an OH practitioner sees in order to advise the employing company. Offers of employment are usually conditional upon "medical clearance"—is the

Major responsibilities of occupational health physicians

Professional ethical obligations

- Provide a good standard of practice and care
- Keep up to date and maintain performance
- Respect confidentiality and maintain trust
- Maintain good communications

General Medical Council. *Good medical practice.* London: GMC, 2001
See also: Faculty of Occupational Medicine. *Good medical practice for occupational physicians.* 2001

Guidance

- Health assessments
- Advice on absence
- Confidentiality
- Health records
- Relationships with others

See: The Occupational Health Committee. *The occupational physician.* London: BMA, 2001

The OH physician must exercise professional skill and judgement in giving advice, and there is an ethical duty to inform the applicant of any abnormality uncovered by the process; however, the contractual duty lies with the prospective employer

applicant fit, in medical terms, for the duties of the post? The degree and extent of the duty upon the OH physician has been explored in two leading English cases.

Medical reports

When the OH practitioner is asked to provide advice on an individual's health for employment purposes, they should obtain written consent before releasing their opinion. This is correct ethical practice.

Because the OH practitioner is not usually the clinician caring for that individual, the Access to Medical Reports Act 1988 will not generally apply. It will apply, however, if the OH practitioner seeks further information from any other specialist or general practitioner who has been providing such care. The provisions of the Data Protection Act 1998 apply to obtaining, use, and retention of any personal information, including OH records.

Health and safety law

Statutory duties upon the employer

Health and safety

The Health and Safety at Work etc. Act (HSWA) 1974 is the main statute covering the general responsibilities of the employer. It covers others who might be affected by workplace activities—contractors, visitors, and the general public. The workplace must be safe and well maintained, with safe systems and organisation of work. Equipment and tools must be suitable and well maintained. Ensuring that employees behave safely is also down to the employer, who has the responsibility for supervision. Supervisory staff must be demonstrably competent. This duty is only limited when the employee might be considered to be "on a frolic of his own," as the courts have termed it.

The underlying principle of the statutory framework is that those who generate risk as a consequence of work activities have a duty to protect the health and safety of anyone who might be affected by those risks. Occasionally the duty is absolute but more commonly the extent of the duty is "as far as reasonably practicable." This allows the employer to balance the degree of risk against the difficulty and cost of reducing it. A small employer with modest resources may therefore argue that it could not go so far in risk reduction as a multinational company, for instance.

A great deal of more recent legislation, driven largely by directives from the European Commission, has focused on particular areas.

The general move has been away from the prescriptive approach and towards a duty on the employer (and the self employed) to assess risks arising from work activities. The employer must then identify and institute preventive actions on the basis of their assessment.

Reporting injuries and disease

Fatal and major injuries, those resulting in three or more days lost from work, and certain occupational diseases must be reported to the Health and Safety Executive (HSE), as per the Reporting of Injuries, Diseases and Dangerous Occurrences Regulations 1995 (RIDDOR). For a disease to be reported, the disease must be listed in the regulations, the affected employee must be involved in a relevant task or activity, and there must be a written diagnosis from a doctor. Under-reporting is considerable—the patient or treating doctor may not realise that the condition is work related, employers have little incentive to report, or the patient may fear for their job and therefore not wish to agree to disclosure.

Duty of care at pre-employment

Baker v *Kaye* (1997)
Mr Baker, applying for a job as International Sales Director, attended for pre-employment assessment by Dr Kaye. During the assessment, Dr Kaye elicited a history of significant alcohol consumption, supported subsequently by abnormal liver enzymes. Mr Baker had already resigned from his existing post, and when Dr Kaye advised the new employer that he did not consider Mr Baker fit for employment, Mr Baker sued for loss of the new post. The court in this case held that the OH physician owed a duty of care to the prospective employee, as well as to the employer, but as Dr Kaye had taken reasonable care in making the assessment, he was found not to be negligent

Kapfunde v *Abbey National plc* (1998)
However, the Court of Appeal, in *Kapfunde* v *Abbey National plc* (1998), disagreed with the decision in the case above.

Mrs Kapfunde, who suffered from sickle cell disease, applied for a job at the Abbey National. Dr Daniel, advising Abbey National, reported that the applicant's medical history and previous absence record indicated that she was likely to have an above average sickness record. Mrs Kapfunde was not considered for the job, and subsequently sued Abbey National, arguing that Dr Daniel had been negligent. The Court, in judging Dr Daniel not negligent (because she had exercised reasonable skill and care in reaching her decision), added that neither did she owe a duty of care to Mrs Kapfunde

Modern domestic legislation since 1988 based on risk assessment

More than 20 European directives have produced a large number of specific regulations, notably the "Framework" Directive for the Introduction of Measures to Encourage Improvements in Safety of Health of Workers, which was enacted into UK law by the Management of Health and Safety at Work Regulations 1992 (updated in 1999), and together with its five "daughter" directives forms the "six pack" (marked *).

- Management of Health and Safety at Work Regulations 1992 (now MHSWR 1999)*
- Workplace (Health, Safety, and Welfare) Regulations 1992*
- Provision and Use of Work Equipment Regulations 1992*
- Personal Protective Equipment Regulations 1992*
- Display Screen Equipment Regulations 1992*
- Manual Handling Operations Regulations 1992*
- Working Time Regulations 1998

Many of these are accompanied by an approved code of practice or guidance notes. These are not legally binding in their own right. However, they bring detail to the statute, and guidance on how compliance may be achieved. An employer would have to justify a diversion from their recommendations

Information on potential health risks must be given to the workforce, with suitable instruction and training on control measures. Sometimes medical surveillance may also be specified

Failure to comply

Despite the extensive legislation to prevent them, work related injuries and illness still occur. In these cases, the legal system has two distinct roles: to punish the negligent employer, and to compensate the injured employee.

Prosecution

The enforcing authorities. The HSE is responsible for enforcement activities in most workplaces, including factories, farms, hospitals, schools, railways, mines, nuclear installations, and also driving as part of work. The exceptions are—for example, retail and finance, where responsibility lies with local authorities. The Health and Safety Commission and the HSE were established by the HSWA 1974. The HSE includes the Employment Medical Advisory Service, which comprises doctors and nurses who are OH specialists and who have the full powers of inspectors.

The law. Breach of the HSWA is subject to criminal sanctions. Prosecutions (most are undertaken by the HSE in the Magistrates' and Crown Courts) will generally result in a fine.

Manslaughter. After a fatal accident, the HSE will defer to the police. The Crown Prosecution Service may then bring a case for manslaughter. However, prosecution is rarely successful. The difficulty is that a "company" is not an individual and therefore not capable of a crime of this nature.

Manslaughter

An unsuccessful case

A total of 188 people died when the Herald of Free Enterprise capsized at Zeebrugge in 1987. The case against P&O showed failures within the management and with several individuals on the vessel. However, no single person was found sufficiently at fault for the charge to apply

A successful case

Peter Kite, the managing director of Oll Ltd, received a custodial sentence after four teenagers drowned during a canoeing trip in Lyme Bay in 1993. Kite ran the small company and was found to be the "controlling mind." There was a history of his ignoring warnings about safety and he clearly failed to adhere to accepted standards. The company was also found guilty of manslaughter and fined £60 000

Current thinking

Reform of this area has been considered since the 1996 Law Commission Report. This recommended new offences of corporate killing and individual offences of reckless killing and killing by gross carelessness. However, legislation has not been forthcoming. In May 1998 Simon Jones, aged 24, died on his first day at work at Shoreham Dockyard; the resulting unsuccessful action provoked further outcry. In 2000-1, 26 cases were referred by the Health and Safety Executive to the Crown Prosecution Service for consideration of manslaughter charges; six are proceeding. Since 1992, 162 referrals have led to 45 prosecutions and 10 convictions. Five individuals have received prison sentences

Compensation

An employee who suffers from a work related illness or injury has two possible routes to seek compensation. Firstly, they may claim from the government if they have a "prescribed disease" via the Industrial Injuries Benefit Scheme. Secondly, and entirely separately, they may claim against the employer via a personal injury claim in the civil court.

Prescribed diseases. The Industrial Injuries Scheme administered by the Department of Work and Pensions "prescribes" a number of occupational illnesses for compensation. To qualify for compensation, the applicant must have the prescribed

Main powers of enforcement authorities

Health and Safety Executive

- Enters and inspects workplaces
- Issues improvement or prohibition notices (immediate or deferred)
- Prosecutes

Employment Medical Advisory Service

- Gives advice on health and safety issues to employers and employees
- Investigates complaints or concerns about health, or after a report under Reporting of Injuries, Diseases, and Dangerous Occurrences Regulations 1995
- Has the same legal powers as inspectors (where the Health and Safety Executive is the reporting authority)
- Appoints doctors for health surveillance required by regulations (Ionizing Radiation Regulations 2000, Control of Lead at Work Regulations 1998, Diving at Work Regulations 1997)

Criminal law and enforcement activities

UK criminal law

- Arises from statute and is a punitive system for offences against society as a whole
- Acts of parliament and regulations made thereunder provide the "rules" by which employers are expected to abide
- Case law—the court's decisions in specific cases—provides guidance on the interpretation of these rules
- Decisions made in higher courts are binding upon lower courts
- The burden of proof in criminal cases is "beyond reasonable doubt;" a higher standard than that applying in civil claims "on the balance of probabilities"

Health and Safety Executive (2000-1) Activities

- 11,058 enforcement notices (70% in manufacturing and construction)
- 6673 improvement notices
- 2077 prosecutions, resulting in 1493 convictions (72%)
- Average penalty £6250

Local authorities (1999-2000)

- 4850 improvement notices
- 1250 prohibition notices
- 412 prosecutions, resulting in 358 convictions (87%)
- Average fine £4595

Zebrugge ferry disaster. Reproduced with permission from Rex Features

disease, and must also have worked in an occupation recognised to carry a risk of that particular disease. The amount of the payment depends on the degree of disability, as assessed by an adjudication officer.

Civil claims. A claim through the civil courts is a means of compensating one person for damage arising from another's action or inaction. Most claims are brought under the tort of negligence. The employer is held to have a broad, general duty of care to avoid harm to its employees. This is part of the common law (where there is no guiding statute law, but is developed over time by decisions of the judiciary).

The employee must argue that the employer failed in their duty of care to safeguard the worker's health. The applicant employee must show that:

(a) The employer owed the worker a duty of care
(b) The employer negligently breached that duty
(c) The employee suffered damage as a result of that breach.

The level of proof is the "balance of probabilities."

Employees with illnesses that may be occupationally related but are not prescribed can only pursue this route.

Large damages paid in compensation may seem impressive when reported in the media, but the adversarial system as presently practised has its problems, and the impact of new civil procedure rules introduced in 1999 in an attempt to improve the present system (on the basis of the Woolf reports on access to justice) is not yet clear. Furthermore, if state compensation is paid for an industrial disease before the personal injury claim, this may be clawed back from awarded damages in excess of £2500.

Another option in the civil courts is an action for breach of statutory duty. The HSWA expressly excludes any such civil action in sections 2-8, although some regulations made under the HSWA do support such an action. Current plans are to remove the existing civil liability exclusion from the Management of Health and Safety at Work Regulations 1999.

Employment law

Legislation

A considerable body of both European and domestic legislation exists in this area. The Employment Rights Act 1996 (ERA) consolidated employees' rights into a single statute. Other primary and subordinate legislation relates to issues of discrimination, pay, and sick pay and are supported by various influential codes of practice, such as those produced by the Advisory, Conciliation and Arbitration Service.

Complaints in this area are heard by employment tribunals, which were established so that employment disputes can be settled rapidly and without the expense of going to court. The employment tribunal comprises three members, including an experienced lawyer as the chair. Appeals are referred first to the Employment Appeal Tribunal, and ultimately to the Appeal Court.

Dismissal

The Employment Rights Act gives employees the right not to be unfairly dismissed. In general, one year's continuous employment is required before a complaint for unfair dismissal can be brought. Some types of unfair dismissal, notably certain grounds relating to discrimination or health and safety, require no such qualifying period—this might be the case if an employee were dismissed because he or she raised the issue of hazardous working conditions.

How diseases become prescribed

Thirty-nine conditions are listed in four categories; those caused by:

A Physical agents (for example, occupational deafness)
B Biological agents (for example, viral hepatitis)
C Chemical agents (for example, angiosarcoma of the liver)
D Those of a miscellaneous nature (for example, occupational dermatitis)

The list is similar to those diseases reportable under the Reporting of Injuries, Diseases and Dangerous Occurrences Regulations 1995

The Industrial Injuries Advisory Council advises on the addition of new prescribed diseases. Its criteria are narrow: the disease must be a recognised risk in a particular occupation and not to the general population, and the causal link between exposure and disease must be well established. This process may take some time

- Vibration white finger (now hand arm vibration syndrome) was considered four times between 1954 and 1985, when it was prescribed
- Occupational deafness was considered in 1961 and prescribed in 1975

Common law duty of care

The depth and breadth of the employer's duty of care has been developed over the years by landmark cases. The concept of the "*reasonable and prudent employer, taking positive thought for the safety of his workers in the light of what he knows or ought to know*" was clarified by Judge Swanwick in 1968

The duty is greater if the employee has a known vulnerability. This is known as the "eggshell skull" rule, after a 1901 case. A better example is that of *Paris* v *Stepney Borough Council* (1951). Mr Paris, a bus fitter with sight in only one eye, lost the sight in the other eye after entry of a metallic foreign body. The Council was negligent in not providing Mr Paris with eye protection, though, given that the risk of an accident was slight, they were not obliged to provide this for others in their workforce

Exactly when an employer should be aware of a particular health risk in the workplace is inevitably contentious, particularly in relation to claims for occupational illness. Courts may decide on a "date of knowledge," after which no employer could reasonably claim ignorance. This date will often relate to government guidance or other influential advice. For instance, in the case of noise induced hearing loss, the year 1963 became the watershed, after a Ministry of Labour pamphlet in that year

Areas of interest in civil litigation

Work related stress (WRS)

The 1995 case of *Walker* v *Northumberland County Council* attracted considerable attention. Mr Walker was an area social services officer. He had a heavy caseload, and frequently requested help. After five months' absence for a "nervous breakdown," he returned to a backlog of work, and the promised assistance did not materialise. After a second breakdown, he sued his employer. The Council was held not to be liable for his first breakdown as they were not aware that he was susceptible to stress. However, it was liable for the second breakdown. The risk was foreseeable and preventable, and there was a duty not to cause Mr Walker psychiatric injury. Damages were £175 000

However, successful actions for WRS are few and far between, and the burden of proof on the employee remains considerable. In February 2002 the Court of Appeal overturned three awards (*Hatton* v *Sutherland and others*) of almost £200 000. It set down 16 guidelines that it considered relevant, and which will aid both courts and employers. These include the following:

- The employer is entitled to assume that the employee can deal with the normal pressures of the job unless there is a known vulnerability
- No occupations should be regarded as intrinsically dangerous
- If the only alternative would be to dismiss or demote the employee, the employer would not be in breach of duty if the employee willingly continues in the job

The hurdle for applicants under this heading remains high and, if anything, this judgement will make a claim for WRS more difficult

Dismissal occurs when the contract of employment is terminated by the employer, when a fixed term contract expires and is not renewed, or when an employee terminates the contract as a result of the employers' conduct. The five potentially fair reasons for dismissal are given in the box.

Absence from work may generate grounds for dismissal and, if absence is attributed to ill health, OH advice will be required. It is important to differentiate between long term absence and recurrent short term absenteeism.

Long term absence—This may give rise to fair dismissal on the grounds of capability, which includes both ill health and incompetence. The employer is expected to gather enough information to assess the situation fully and to decide on a reasonable course of action. This should include consultation with the employee and will often include a medical opinion. The employer might consider alternative work, although it is under no statutory duty to do so (unless the case falls under the disability discrimination legislation, vide infra). The employer cannot know details of the illness because of confidentiality, but is entitled to ask when the employee might recover, whether the employee will be capable of returning to their former job and, if not, the likely restrictions on capability.

The final decision on employment is a management rather than a medical decision, with the physician in an advisory role. It is important to appreciate that the cause of ill health is irrelevant to the fairness of the dismissal, even if the current employment is likely to have been the cause.

Attendance—The problem of recurrent short term absenteeism may be approached rather differently. Employers may view this as an attendance issue and are entitled to expect a certain level of reliability from employees. The genuineness of the illness is not relevant, as an employer may ultimately fairly dismiss on the grounds of either capability or "some other substantial reason." However, the employer should investigate fully and act in line with its absence policy, giving due warning to the employee that attendance is expected to improve. It is good practice (although not essential, depending on the case) to take medical advice as to whether poor attendance is because of an important underlying medical condition. (If there is, the case might more properly be dealt with as a capability problem.)

Disability discrimination

The employment provisions of the Disability Discrimination Act 1995 came into force on 2 December 1996, with duties on the employer to accommodate disabled people, whether existing employees or job applicants. It is unlawful to discriminate—that is, to treat anyone with a disability less favourably for reasons relating to the disability. There is a duty to make "reasonable adjustments" to allow the disabled person to work. However, the Act can allow the employer to justify discriminatory treatment.

Awards for complaints under the Disability Discrimination Act have no upper limit; the stakes are therefore potentially high. Employment tribunals have sometimes had difficulty dealing with the medical issues, as they do not normally use medical experts. Experience of this legislation has clarified and confused in almost equal measure.

Reasons for fair dismissal

1. Relating to capability ("skill, aptitude, health, or any other physical or mental quality") or qualifications ("any degree, diploma, or other academic, technical, or professional qualification")
2. Relating to conduct (behaviour at, or sometimes outside, the workplace)
3. Redundancy
4. If employee cannot continue to work without breach of statutory duty (such as after loss of driving licence)
5. Some other substantial reason (SOSR) sufficient to justify dismissal

In February 2002 the compensatory award for unfair dismissal was limited to £52 600. The burden of proof is said to be neutral, although the employer is required to show that the dismissal was not unfair. An employment tribunal will judge the circumstances of the case—including elements such as the size, resources, consistency of behaviour, and procedural correctness of the employer—in deciding reasonableness

The Disability Discrimination Act (DDA) and some definitions

Disability—"a physical or mental impairment causing a substantial and long term adverse effect on the ability to carry out normal day to day activities"

- A physical impairment is not defined in the legislation, but is likely to encompass any "organic or bodily detriment," including severe disfigurements (facial scars or burns), but excluding deliberately acquired disfigurements (tattoos or body piercings)
- A mental impairment is any clinically well recognised condition (that is, one recognised by a responsible body of medical opinion), and must be beyond a reaction that could be described as a normal human reaction
- A substantial adverse effect is defined as one that is more than minor or trivial
- Normal day to day activities are:
  - Mobility
  - Manual dexterity
  - Physical co-ordination
  - Continence
  - Ablity to lift, carry, or otherwise move everyday objects
  - Speech, hearing, or eyesight
  - Memory or ability to concentrate, learn, or understand
  - Perception of the risk of physical danger
- Long term implies an impairment that has lasted 12 months or more, is likely to last 12 months or more, or is terminal

Certain specific conditions (for instance, nicotine or alcohol dependence) are excluded from the Disability Discrimination Act. Controlled or corrected, progressive, and recurring conditions may be included

Reasonable adjustments to allow the disabled to work

- Accessible and equitable recruitment processes
- Modifications to equipment
- Changes to job design and work environment
- Resources and cost are relevant

Justification

- The failure to adjust must be both material to the circumstances of the case and substantial
- Stricter than "reasonable"
- Requires hard evidence

The Disability Discrimination Act does not currently apply to organisations with fewer than 15 employees, but this exemption will be removed from October 2004. The provisions of the Act are also likely to be extended to include the emergency services, and other medical conditions. The justification provision will be removed

Outcomes of the Disability Discrimination Act

5662 cases brought up to March 2000 (England and Wales)—23% successful

Medical conditions
- 21% back or neck conditions
- 16% hand or arm conditions
- 14% depression or anxiety

Legal issues
- 34% concerned failure to transfer to suitable alternative work
- 26% sought to justify on the basis of the amount of sick leave
- 51% required medical evidence

Awards

Total compensation in 1999: £369 297
Average: £9981 per award
Maximum award in 2001: £278 800

Disabled person at work with appropriate aids such as a voice recognition dictation system linked to a laptop computer for an employee no longer able to type rapidly. The photograph was produced by Mr D Griffiths, with the subject's permission

Future developments

Many other areas may come to have relevance to the work of the OH practitioner, two of which are considered below.

Human rights

The Human Rights Act 1998 came into force in October 2000, bringing the European Convention on Human Rights into UK law. It makes no explicit reference to HSWA. However, under the right to privacy it may forseeably impinge on areas such as drug testing and surveillance. The lack of legal aid for employment tribunals (in England and Wales) and the fairness of the employment tribunal system may generate debate under the provisions for the right to a fair trial.

Rehabilitation

In contrast to the numerous duties to prevent ill health or injury, there is currently no requirement to rehabilitate back to the workplace. This has a huge cost: in 2000, 2.29 million people claimed incapacity benefit, and employers paid out £750 million in compensation under employer's liability insurance schemes. The United Kingdom has a poor record; a Swedish worker has an almost 50% chance of returning to work after an injury, whereas in the United Kingdom the figure is only 15%. Employers may have to develop a policy in this area as in all other health and safety fields. The Departments of Health, and of Work and Pensions are working on a pilot initiative to encourage early return to work, with its effectiveness evaluated by the National Centre for Social Research.

Frequent questions for the OH practitioner in relation to the Disability Discrimination Act
- **Is the condition covered?** In practice, employment tribunals have been hesitant to exclude a condition even when there is considerable scientific debate about the exact nature of the diagnosis (for example, chronic fatigue syndrome)
- **Does the Act apply?** This is a legal decision, and judgement rests with the employment tribunals. The OH practitioner can advise, and the employer make its own judgement, but the employment tribunal is the final arbiter
- **Is treatment unequal?** The "comparator" against which the disabled person should be considered must be an able bodied individual; not one who has another condition but does not fall under the Act
- **Have accommodations been considered?** The employer must make genuine efforts to accommodate the disabled individual. The guidance that accompanies the Act must be followed closely
- **Is unequal treatment justified?** Less favourable treatment may currently be justified, but the employer must make a properly constructed argument with evidence to support its case
- **What happens if health and safety may be compromised?** The employer has a difficult balancing act under these circumstances. However, provided the employer has undertaken a proper risk assessment and subsequently generated a rational policy, then the tribunal cannot disregard the policy on the basis of a differing medical view. It is vital, though, that the employer acts on competent advice backed up by good evidence
- **May the disabled person assume a risk to their own health?** In other words, at what threshold does paternalism on the part of the employer take over from the well informed view of the individual? Current case law suggests that when there is a significant risk to health, the employer has the right (or even duty) to exclude the employee from that work activity

Further reading

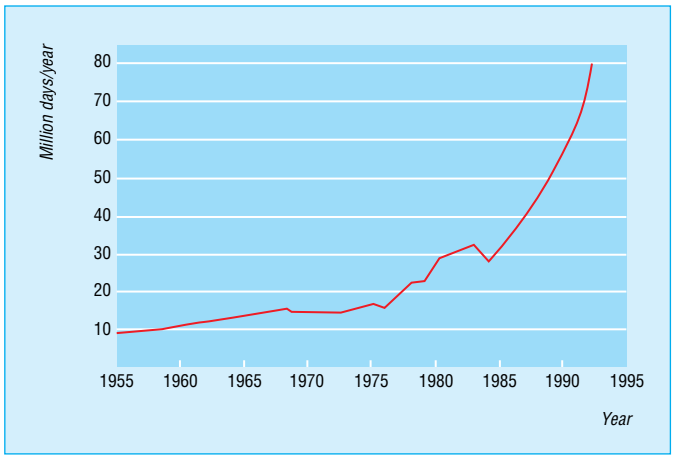
- General Medical Council. *Good medical practice*. London: GMC, May 2001 (www.gmc-uk.org). *Covers expectations of the regulatory body and duties as a member of the medical profession*
- Faculty of Occupational Medicine. *Good medical practice for occupational physicians*. London: FOM, Dec 2001 (www.facoccmed.ac.uk). *More specific advice applied to occupational medical practice, based on the GMC guidance*
- Faculty of Occupational Medicine. *Guidance on ethics for occupational physicians*. London: FOM, May 1999. *Occupational physicians undertake roles outside the traditional doctor-patient relationship. Guidance on the ethics of commonly encountered situations that differ from clinical medical practice is valuable*
- British Medical Association. *The occupational physician*. London: BMA, June 2001 (www.bma.org.uk). *More detailed consideration of the role of the OH physician, relationships with employing organisations, and terms and conditions of service*
- Kloss D. *Occupational health law*. Oxford: Blackwell Science, 1998. *Widely acknowledged as the "bible" in this area*
- Lewis D, Sargeant M. *Essentials of employment law*. City: IPD 2000. *All the basics from a legal perspective*
- Berlins M, Dyer C. *The law machine*. London: Penguin 2000. *Brief, accessible description of the judicial system and explanation of how it all works*
- Branthwaite M. *Law for doctors*. City: RSM 2001. *Concise and physician orientated, aimed primarily at clinical practice*
- Monitoring the Disability Discrimination Act 1995. First Interim Report March 2000. City: Income Data Services Ltd (www.incomesdata.co.uk). *Analysis of caseload and decisions concerning the employment provisions of the Act*
- Health and Safety Executive Annual Report 2000/1 (www.hse.gov.uk). *Information on the activities of the HSE, and the data collected on occupational ill health and accidents*
- Data Protection Commissioner. www.dataprotection.gov.uk. *How data must be managed, applied to all personal information, including health records*
- www.courtservice.gov.uk. *Employment appeal tribunal and high court decisions*
- www.lawreports.co.uk. *Judgements from the House of Lords downwards*

6 Back pain

Malcolm IV Jayson

The number of working days lost as a result of back problems has increased dramatically in recent years. Over 80 million days were lost because of registered disability in 1994, but the total estimate, including short spells, is probably in the order of 150 million—four times more than 30 years ago. This increase does not reflect an increased incidence of back problems, which has changed little over the period, but it is probably because of an altered reaction to the problem, with increases in sick certification and state benefit, perhaps reflecting patients' and doctors' expectations, concerns by employers, and social and medicolegal pressures. There is, however, recent UK evidence to suggest that the peak in this rise of back pain incapacity is now past and claims for benefit are now falling. Whether the disability has simply been reclassified as "stress" (which is rising) is uncertain.

The costs of back pain are huge. Recent estimates suggest that the overall cost to the UK economy is about £6 billion a year. Improved management and better outcomes would lead to major financial and medical benefits.



Changes in sickness and invalidity benefit for back pain since 1955

Who gets back pain?

The problem affects workers of all ages. It usually starts between the late teens and the 40s, with the peak prevalence in 45-60 year olds and little difference between the sexes. The prevalence of back disability is increased in people performing heavy manual work, smokers, and those in non-managerial positions. Clearly these factors interact in many patients. It is often difficult to determine whether heavy manual work has caused or aggravated a back problem or whether a worker cannot do the job because of back pain. Obesity and tallness are also associated with back problems. Postural abnormalities do not predict back problems, except possibly gross discrepancies in leg length.

Psychological factors are important. Psychological distress in a population without back pain predicts the development of back pain. In the Boeing aircraft factory, workers who did not enjoy their jobs had a greatly increased risk of reporting back injury.

Heavy repetitive manual work increases the risk of back problems

Causes of back pain

The major causes of back pain are mechanical strains and sprains, lumbar spondylosis, herniated intervertebral disc, and spinal stenosis. In many cases it is not possible to make a specific mechanical diagnosis. Such problems are commonly called non-specific back pain. Non-mechanical causes of back pain include inflammatory disorders such as ankylosing spondylitis and infections, primary and secondary neoplasms, and metabolic bone disorders such as osteoporosis. The patient's clinical characteristics and a general health screen will exclude systemic disease.

The physical state of the spine determines how well it functions, and use and injury of the back will alter its structure. This interrelation between structure and function is central to understanding many back problems related to work

Pre-employment screening

There is no evidence that physical build, flexibility of spine movements, or other physical characteristics are of any value in predicting the development of back problems, and they should not be used for screening purposes. In particular, lumbar radiographs are not helpful in identifying people liable to develop back pain at work.

A detailed medical and occupational history is required for all employees, and an assessment of their fitness to do the job.

The principal risk factor for back pain is a history of back pain. Those who have had back problems in the past are likely to experience further episodes in the future

The most useful single item of information in predicting potential back problems is a history of back pain, particularly if it is recent and severe enough to cause absence from work.

Preventing back injuries

Manual handling is commonly associated with strains and sprains of the back and resultant disability. *Manual handling. Guidance on regulations* lists measures that employers should take to reduce the risk of problems. These include:

- Avoiding hazardous manual handling operations as far as is reasonably possible (lifting aids may be appropriate)
- Making an appropriate assessment of any hazardous manual handling operations that cannot be avoided
- Reducing the risks of injuries from these operations as far as is reasonably possible.

Weight limits

In Britain no limits for weights that may be lifted have been stated. This is because setting a weight limit is a fallacious approach as so much depends on the individual and the circumstances of any procedure. When a load is moved away from the trunk the level of stress on the lower back increases. As a rough guide, holding a load at arms' length imposes five times the stress experienced with the same load held close to the trunk. Moreover, the further away the load is from the trunk the less easy it is to control, adding to the problems.

Guidelines to loads that may be lifted are necessarily crude, given the wide range of individual physical capabilities even among fit and healthy people. There are no truly safe loads. Present guidelines do no more than identify when manual lifting and lowering operations may not need a detailed risk assessment. If the handler's hands enter more than one of the box zones during the operation, the smallest weight figures apply. Where the handler's hands are close to a boundary an intermediate weight may be chosen. Where lifting or lowering with the hands beyond the box zones is unavoidable, a more detailed assessment should always be made.

Lifting techniques

The technique for lifting is important. Simple ergonomic principles will protect the back against excessive strains. A poor posture increases the risk of injury. Examples include stooping and twisting while weight bearing, carrying loads in an asymmetric fashion, moving loads excessive distances, and excessive pushing and pulling. Repeated or prolonged physical effort may carry additional risk. Many episodes of back pain develop after sudden or unanticipated movements such as a stumble on the stairs or an unexpected twist.

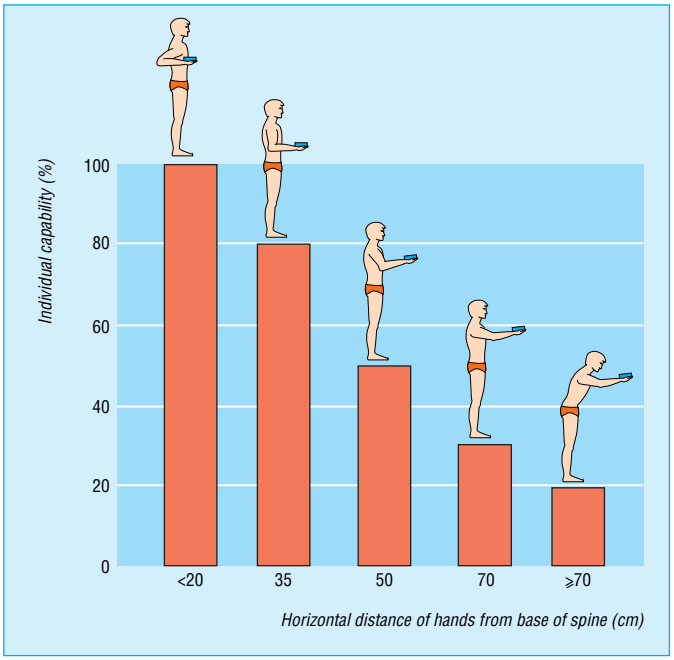
Wherever manual handling occurs employers should consider the risks of injury and how to reduce them by reviewing the task required, the load carried, the working environment, and individual capability. Redesigning the job and providing mechanical assistance may be appropriate, and individual workers should be trained in safe manual handling.

Diagnosis and prognosis

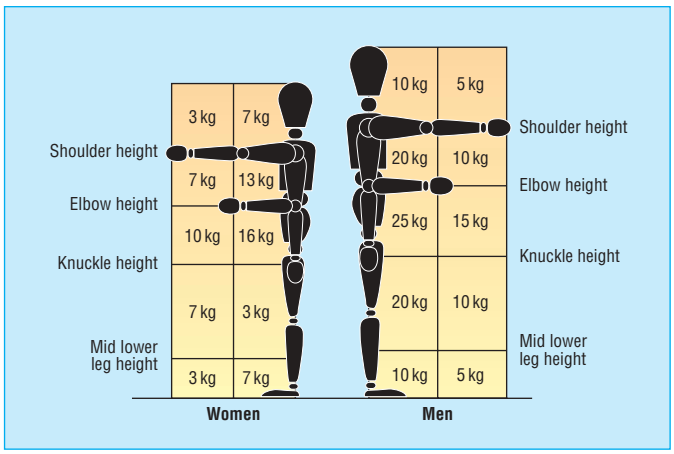
Diagnostic triage

On simple clinical grounds, patients with acute back problems can be triaged into simple backache, nerve root pain, and possible serious spinal conditions. Simple back pain will be managed by an occupational health physician or general practitioner. Nerve root pain will initially be dealt with by a general practitioner in a similar way to simple backache, although at a slower pace, providing there is no major or

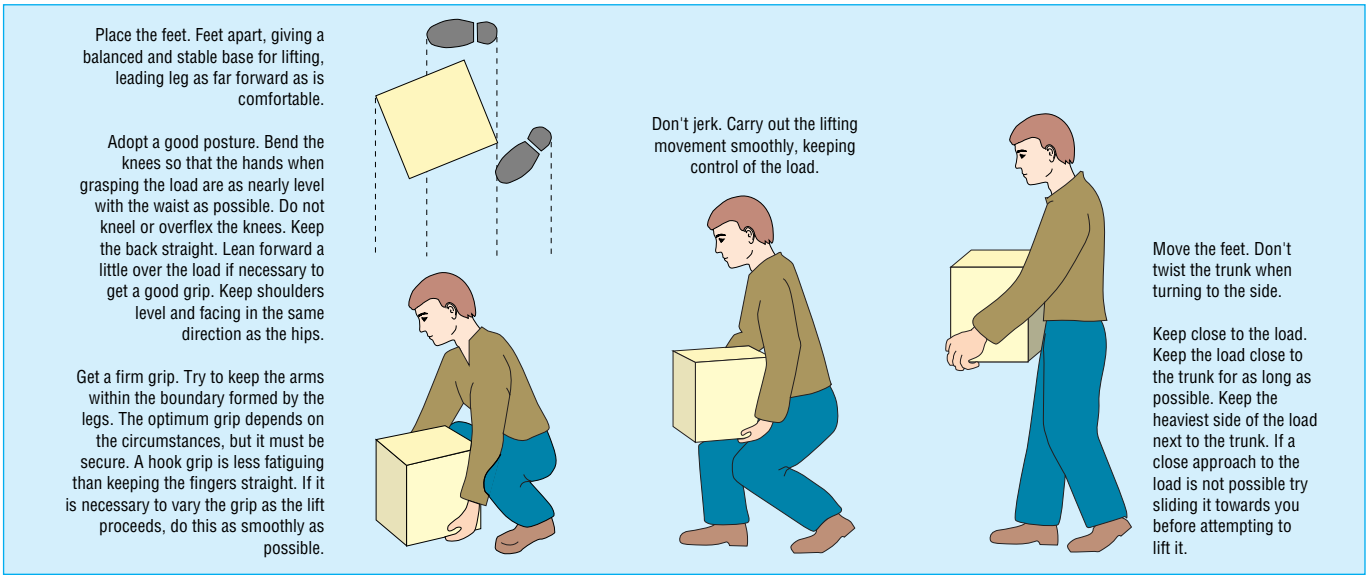
Excessive loading—simple mechanical aids can eliminate this



Reduction in handling capacity as hands move away from trunk



Guide to loads that may be lifted in various positions, assuming that the load is easily grasped with both hands



Principles of lifting and carrying a load

progressive motor weakness. However, early referral to a specialist may be required. Patients with possible serious spinal conditions require urgent referral, and emergency referral is needed for those with widespread or progressive neurological changes.

Prognosis

Most patients have simple backache. The exact condition and source of the pain are rarely identifiable, but the principles of management are now well established. Nearly all episodes of acute back pain resolve rapidly. Most patients return to work within a few days, and 90% return within six weeks. Some patients, however, develop chronic back pain, and this small proportion with prolonged disability is responsible for most of the costs associated with back injuries.

With longer time off work, the chances of ever getting back to work decrease rapidly. Only 25% of those off work for a year and 10% of those off work for two years will return to productive employment.

Investigations

Routine radiographs of the lumbar spine should be avoided. Apparent degenerative changes are common and correlate poorly with symptoms: they are better considered as age related changes. Radiographs are necessary when there is the question of possible serious spinal conditions, but a negative result does not exclude infection or tumour.

Imaging with computed tomography or magnetic resonance imaging is of no value for simple backache. These techniques also often display age related changes that correlate poorly with symptoms. The presence of these changes does not influence management.

Management

Simple backache

The early management of acute back pain is important. Much of the traditional management of back pain seems to promote chronicity. In view of the increasing toll of back disability, the Clinical Standards Advisory Group of the UK Departments of Health has published guidelines on managing back problems. These emphasise the importance of maintaining physical activity and minimising the period off work.

Indications for emergency referral

- Difficulty with micturition
- Loss of anal sphincter tone or faecal incontinence
- Saddle anaesthesia about anus, perineum, or genitals
- Widespread (more than one nerve root) or progressive motor weakness in legs or disturbed gait

Characteristics of simple backache

- Onset generally at ages 20-55 years
- Pain in lumbosacral region, buttocks, and thighs
- Pain is mechanical in nature—varies with physical activity and with time
- Patient is well
- Prognosis is good—90% of patients recover from acute attack in six weeks

Characteristics of nerve root pain

- Unilateral leg pain worse than back pain
- Pain generally radiates to foot or toes
- Numbness and paraesthesia in same distribution
- Signs of nerve irritation—reduced straight leg raise which reproduces leg pain
- Motor, sensory, or reflex change—limited to one nerve foot
- Prognosis reasonable—50% of patients recover from acute attack in six weeks

Red flags suggesting possible serious spinal pathology

- Age at onset <20 or >55 years
- Violent trauma—such as fall from height, or road traffic accident
- Constant, progressive, non-mechanical pain
- Thoracic pain
- History of cancer
- Use of systemic corticosteroids
- Misuse of drugs, infection with HIV
- Patient systematically unwell
- Weight loss
- Persisting severe restriction of lumbar flexion
- Widespread neurological signs
- Structural deformity

The natural course of simple backache is spontaneous resolution within a short time. Treatment is directed at relief of symptoms, a minimum period of rest, physical activity, and a rapid return to work.

Pain relief is with simple analgesics such as paracetamol or non-steroidal anti-inflammatory drugs. Narcotics should be avoided if possible, and never used for more than two weeks.

Rest is prescribed only if essential. Bed rest should be limited to three days as longer periods increase the duration of disability.

Early activity is encouraged. Patients should be reassured that exercise promotes recovery. The particular type of exercise is less important. There may be some increase in pain, but the patient should be reassured that hurt does not mean harm, and that those who exercise have fewer recurrences, take less time off work, and require less healthcare in the future.

Physical therapy should be arranged if symptoms last for more than a few days. This may include manipulation, exercises, and encouraging physical activity. Other techniques such as short wave diathermy, infrared treatment, ice packs, ultrasonography, massage, and traction provide only transient symptomatic benefit, but may enable patients to exercise and mobilise more rapidly. Some factories employ therapists so that physical therapy is available early in the work environment. This approach seems promising in promoting quick recovery and reducing risks of chronicity.

Risk factors for back pain becoming chronic

- History of low back pain
- Previous time off work because of back pain
- Radicular pain, possibly with reduced straight leg raise and neurological signs
- Poor physical fitness
- Poor general health
- Smoking
- Psychological distress and depression
- Disproportionate pain behaviour
- Low job satisfaction
- Personal problems—alcohol intake, marital, financial problems
- Medicolegal proceedings

Persistent back pain

By six weeks, most patients will have recovered and be back at work. A detailed review is required for those with persistent problems. These patients should undergo a biopsychosocial assessment. There are particular risk factors for chronicity and for back pain and more prolonged disability, and their early identification will help in planning treatment.

Biological assessment includes reviewing the diagnostic triage, seeking evidence of nerve root problems or possible serious spinal conditions with appropriate referral. At this stage, measurement of the erythrocyte sedimentation rate, and radiographs, are indicated.

Psychological assessment should include the patient's attitudes and beliefs about pain. Many patients will not attempt to regain mobility because of unjustified fears about the risks of activity and work. Patients may have psychological distress and depressive symptoms, and develop characteristics of abnormal illness behaviour.

Social assessment includes patients' relationships with their families (who may reinforce the patient's disability), and work problems related to the physical demands of the job, job satisfaction, compensation, and medicolegal issues.

Referral

When a patient with simple backache does not return to work within three months, specialist referral is required to provide a second opinion about the diagnosis, to arrange investigations, and to advise on management, reassurance, multidisciplinary rehabilitation, and pain management. If pain in the back is referred to the buttocks or thighs the appropriate speciality is rheumatology, pain management, or rehabilitation medicine. For nerve root pain, the patient should be referred to orthopaedics or neurosurgery.

Psychological and social factors are increasingly recognised as important, and a multidisciplinary rehabilitation programme is likely to be effective. This may include incremental exercise and physical reconditioning, behavioural medicine, and encouragement to return to work.

Further reading

- Clinical Standards Advisory Group. *Back pain.* London: HMSO, 1994. *The first UK evidence based review containing broad guidelines for the management of back problems. This publication has led to radical changes in the management of back pain in primary care. Reviewed 1999, see www.rcgp.org.uk*
- Croft PR, Papageorgiou AC, Ferry S, Thomas E, Jayson MIV, Silman A. Psychological distress in low back pain: evidence from a perspective study in general practice. *Spine* 1996;20:2731-7
- Bigos SJ, Battie MC, Spengler D, Fisher LD, Fordyce WE, Hansson TH, et al. A prospective study of work perceptions and psychosocial factors affecting the report of back injury. *Spine* 1991;16:1-6. *These two papers have emphasised the importance of the pre-back pain psychological state in predicting the future development of back problems in both primary care and in industry*
- Heliovaara M, Makela M, Kenkt P, Impivaara O, Aromaa A. Determinants of sciatica and back pain. *Spine* 1991;16:608-14 *Another predictive study highlighting the importance of the back history*
- Health and Safety Executive. *Manual handling. Guidance on regulations*, 2nd ed. London: HSE Books, 1998. *Provides helpful advice on manual handling techniques and provides crude guidelines that are useful in industry*
- Deyo RA, Diehl AJ, Rosenthal M. How many days of bed rest for acute low back pain? *New Engl J Med* 1986;315:1064-70. *The first study indicating that bed rest should be minimised in the management of back pain and that longer periods tend to be harmful*
- Waddell G, Burton AK. *Occupational health guidelines for the management of low back pain at work—evidence review.* London: Faculty of Occupational Medicine, 2000. *The principles underlining the management of back pain in relation to work (whether caused by work or impeding work, or both) have been subject to a number of reviews. This review has been carefully evidence based and is an invaluable source of current knowledge on this subject. Also available at www.facoccmed.ac.uk*

Work modification

Early return to work should be a priority because the physical and psychological consequences of inactivity and unemployment contribute to further dysfunction. Although patients should be encouraged to exercise, some are not capable of undertaking heavy manual work. Careful ergonomic assessment is necessary to avoid excessive stresses on the back. In particular, care should be taken to minimise tasks that require bending, lifting, and twisting. Light work—such as reception or inspection duties that require sitting, standing, and walking but avoid long periods in any one position—may be appropriate. At this point a coordinated approach with an Occupational Health Department is likely to be very helpful.

The figure showing changes in sickness and invalidity benefit for back pain since 1955 is adapted from a report of the Clinical Standards Advisory Group. *Back pain*. London: HMSO, 1994. The figures showing reduction in handling capacity as hands move away from trunk, guide to loads that may be lifted in various positions, and the principles of lifting and carrying a load are all adapted from *Manual handling. Guidance on regulations*. Health and Safety Executive. London: HSE Books, 1998.

7 Upper limb disorders

Mats Hagberg

Improved management of patients with work related neck and arm disorders can reduce the number of working days lost and the incidence of work related illness. A patient's quality of life and potential economic loss is largely dependent on the medical consultation.

The consultation

Every patient who seeks medical attention for neck and arm problems is entitled to a thorough medical examination. It is important for the patient—even when the disorder is non-specific—to get a clear message from the treating physician as to whether progressive disease is present, and for the physician to get the patient to engage with and have control over their rehabilitation and return to work.

The assessment of work related musculoskeletal disorders consists of a clinical examination, an exposure history, a workplace assessment, and suitable further tests.

History

The type, onset, and localisation of symptoms should be explored in detail. The use of a manikin ("bodymapping") to let the patient mark the type and location of pain has good reliability. It is important to distinguish between nociceptive and neurogenic pain. Nociceptive pain usually originates from peripheral pain receptors reacting to mechanical or chemical stimuli. Muscle pain can be regarded as nociceptive. Neurogenic pain is caused by a dysfunction in the nervous system. Accompanying sensory disturbances are common, and they can be caused by entrapment of nerves. Neurogenic pain may follow the sensory distribution of a nerve, whereas nociceptive pain is usually more diffuse and does not correspond to a single nerve distribution. Examples of questions to be asked are: "Does the pain radiate?" "Where to?" Diffuse symptoms may indicate musculoskeletal referred pain, whereas pain radiating towards specific dermatomes suggests a cervical root lesion (radiculopathy). For each single symptom the character, quality, distribution, intensity, frequency, and duration should be described. Information should be elicited about the relation between symptoms and posture, about movements and loading during occupational activity, and the relationship of symptoms to recreational activities and rest.

Special efforts should be made to identify red flags. Examples of red flags are weight loss and severe pain in the mornings. This may indicate a severe systemic disease, endocrine disorder, infection or malignancy. The family and medical history, and questions about morning stiffness and signs of inflammatory activity (joint swellings) may suggest a rheumatoid disorder.

Work and exposure history

A person's job title usually supplies insufficient information to determine whether the disorder is work related and whether the patient can return to their job. The actual work task has to be described in terms of what the patient produces, work posture, repetition, material handling, and work organisation. Any history of sudden events of high energy transfers (formerly termed "accidents") that could have resulted in clinical or subclinical injury should be explored.

Characteristics of non-specific musculoskeletal pain in neck, arm, and hand

History
- Pain and stiffness gradually increase during work and are worst at the end of the working day and week
- Pain localised to cervical spine and the angle between the neck and shoulder or to the upper part of forearm
- Usually no radiation of pain
- Symptoms are improved by heat and worsened by cold draughts

Signs
- Tenderness over neck and shoulder muscles or tenderness over forearm extensor muscles
- Reduced range of active movement of cervical spine (normal passive movement)
- No neurological deficits

Differential diagnosis
- Tendonitis
- Nerve entrapments
- Systemic diseases

Management of work related neck and arm disorders

Clinical management
- Non-steroidal anti-inflammatory drugs can reduce pain and inflammation
- Acupuncture can reduce pain
- Corticosteroids—a single subacromial injection of corticosteroid mixed with local anaesthetic may cure shoulder tendonitis. For tennis elbow and carpal tunnel syndrome, corticosteroids should be used by specialists only
- Heparin (15 000 IU/day in a single intravenous dose) given for 3-4 days is an effective treatment for acute crepitating peritendinitis
- Surgery—surgical division of the carpal ligament is the first choice of treatment for carpal tunnel syndrome. For chronic severe shoulder tendonitis, surgical removal of the lateral part of the acromion may relieve pain at night
- Splints—whether splints should be used to treat early hand and wrist tendonitis and carpal tunnel syndrome is still debated

Modifications to working environment
- Job analysis—to assess work relatedness of a patient's symptoms it is necessary to evaluate working posture, repetition, force and handling of loads, psychological and social factors, and static posture or task invariability
- Job design—job enlargement can reduce the duration and frequency of awkward postures and load handling. Job enrichment reduces poor work content and task invariability. Layout of workplace and technical aids should be improved
- Technique training—ergonomists and supervisors can improve working technique to reduce stressors of postures, motion, and load handling
- Rests and breaks should be organised to allow recovery

For a better assessment of exposure, the patient should be encouraged to bring photographs of their station, products, and tools. Direct observation of the task at the worksite is valuable and can also be used as the basis for suggestions about job redesign and return to work policies during rehabilitation. Direct evaluation can also be enhanced by video recording.

Clinical examination

The physical examination should include the following steps: (1) inspection; (2) testing for range of motion; (3) testing for muscle contraction, pain, and muscle strength; (4) palpation of muscle tendons and insertions; and (5) specific tests. The physician must have a diagnostic strategy to identify and rule out systemic diseases. As a general rule when tests are used for screening or to rule out disease, the test with the highest sensitivity is preferred. When tests are used to confirm or rule in disease, the test with the highest specificity is preferred. Serial (multiple) tests with results that are all normal tend to rule out disease convincingly; serial tests with results that are all abnormal tend to confirm disease convincingly. Several textbooks cover the physical examination of the musculoskeletal system.

Further investigations

Blood tests such as sedimentation rate and rheumatoid factor can be used to rule out general inflammatory disorders. Imaging tests such as radiographs, ultrasound, and magnetic resonance imaging to detect morphological changes should be done if there are red flags present. Radiographic findings such as spinal degeneration, cervical ribs, etc. should be interpreted with caution because they may be normal physiological findings unrelated to back, neck, or arm symptoms. Patients who are told that their radiograph shows that their back or cervical spine is "worn out" may be resistant to rehabilitation. Even advanced magnetic resonance imaging of the spine may show severe degenerative changes that are not related to the patient's symptoms. A patient may deduce from the radiographic findings that they have a progressive disease and thereby become "medicalised." This may, in turn, influence their participation in active rehabilitation and impair the process of returning to work.

Common work related musculoskeletal disorders may constitute a disturbance of sensory neural processing. In the future both neurosensory testing—for example, vibratory perception threshold—and biochemical markers, may become a part of clinical musculoskeletal assessment.

Principles of managing hand and arm pain in keyboard operators

- Exclude clear pathological causes such as carpal tunnel syndrome
- Explore psychological profile, including attitudes to work, and support from management and colleagues
- Reassure patient that the condition will improve and is likely to resolve
- Keep the patient physically active and at work. Both aerobic and strength training will reduce pain and increase performance
- If necessary reduce keyboard work
- Liaise with patient's workplace—if possible, with an occupational physician or nurse
- Consider variation of work tasks, reduced work intensity, encouraging short breaks from keyboard work, or job rotation
- Ensure that workstation ergonomics have been evaluated and are satisfactory and that the patient has been taught to use the equipment properly and has the right glasses
- Monitor patient's progress with regular follow up
- When symptoms have subsided advise gradual increase in normal activities
- Exercise may improve blood flow and reduce pain. Strength training may reduce pain and increase performance. Heat application may be worth trying
- Advice from an experienced physiotherapist may assist in rehabilitation
- Those few patients who do not respond to this multidisciplinary management may be at risk of developing chronic symptoms. Revisit the biopsychosocial aspects
- Consider specialist referral (for example, to an occupational physician, rheumatologist, or pain or rehabilitation specialist)
- In extreme cases where long term disablement seems likely, retraining may be necessary. Voice activated software is now widely available

Classification of disease (ICD-10)

The terminology of common musculoskeletal disorders is confusing. The use of terms such as repetitive strain injury (RSI) and cumulative trauma disorder (CTD) should be avoided. The evidence base is often weak or non-existent for these terms. In industrial settings ergonomics may modify the symptoms and signs of disorders and diseases. In a task involving repetitive arm elevation, signs of both tendonitis and non-specific disorders may be present, which are probably related to both concurrent strain on rotator cuff tendons and static strain on neck and shoulder muscles. The occurrence of musculoskeletal symptoms and clinical signs in working and mixed populations has been described. If the different musculoskeletal symptoms and signs do not wholly comply with the criteria for a disease, the recommendation is to choose an ICD label that focuses on the symptoms rather than on the disease. An example of this for non-specific neck and shoulder

No consensus accepted criteria exist for most ICD-10 (international classification of diseases, 10th revision) musculoskeletal related diagnoses for manual work. When considering the criteria for different musculoskeletal disorders it is reasonable to look first at proposed criteria for surveillance, and epidemiological studies

pain with or without radiation to the forearm would be the label "cervicobrachial syndrome M53.1" (ICD-10, nerve root entrapment is excluded).

Risk factors

Multiple factors

Certain occupations are associated with a high risk for neck and arm pain. Some risk factors can be identified, but the interaction between different risk factors is not understood, and there are not enough data yet to set accurate limits for disease effects. It is important to recognise that personal characteristics and other environmental and sociocultural factors usually play a role in these disorders. A patient with neck pain may be exposed to an awkward posture at work but also to social stress at home: both factors contribute to sustained contraction of the trapezius muscles, inducing pain and stiffness. The cause of a work related disorder can sometimes be attributed to a specific exposure in a job, but there is often simultaneous exposure to several different factors. Individual factors must also be considered when assessing the history of a patient with a work related disorder, and when redesigning a job before such a patient returns to work.

Risk factors for work related neck and arm disorders

- Working posture
  - Awkward postures or task invariability
  - Static postures
- Repetitive motion
- Force—handling loads or tools
- Psychological and social factors
  - Work organisation
  - Stress
- Working environment

Awkward postures

Working with hands at or above shoulder level counts as an awkward posture and may be one determinant of rotator cuff tendonitis. Awkward postures may cause mechanical trauma or compression, reducing blood flow and tissue nutrition.

The pathogenesis of rotator cuff tendonitis is mainly impingement—compression of the rotator cuff tendons when they are forced under the coracoacromial arch during elevation of the arm. The supraspinatus tendon is forced under the anterior edge of the acromion, causing both a compression that impairs blood circulation through the tendon and mechanical friction to the tendon. Reduced blood flow because of static muscle contraction may contribute to degeneration of the rotator cuff tendons.

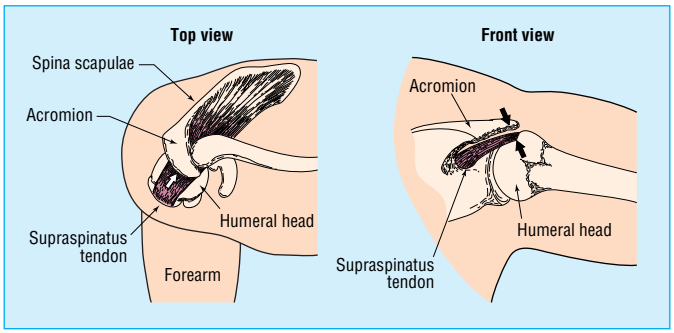
Abduction and forward flexion of more than 30° may also constitute a risk factor because the pressure induced within the supraspinatus muscle will exceed 30 mm Hg, impairing blood flow. The vessels to the supraspinatus tendon run through the muscle, and so raised intramuscular pressure can affect the tendon vasculature.

Poultry dressing involves forceful and repetitive manipulation in cold conditions—ergonomic assessment is essential

Static postures (task invariability)

It used to be argued that to prevent work related musculoskeletal disorders it was necessary to minimise the load that workers were exposed to. This concept has led to the creation of jobs with low external load, but some of these are still not ideal because poor work content usually leads to a job with invariable tasks, resulting in constrained postures and a low static load for the neck and arms. Ergonomists now try to design jobs that are not only physically variable but also psychologically variable and stimulating.

The health problems caused by task invariability may result from prolonged static contraction of the trapezius muscle during work or daily activity, resulting in an overload of type I muscle fibres, explaining the neck pain. At a low level of muscle contraction, the low threshold motor units (type I fibres) operate. A low static contraction during work may result in a recruitment pattern in which only the type I muscle fibres are used, causing selective fatigue of motor units and damage to the type I fibres. Biopsies of the trapezius muscle from patients with work related trapezius myalgia show enlarged



Impingement of the supraspinatus tendon against the surface of the anterior part of the acromion when the arm is raised to shoulder height. Pressure and mechanical friction are centred on the tendon (thick black arrows)

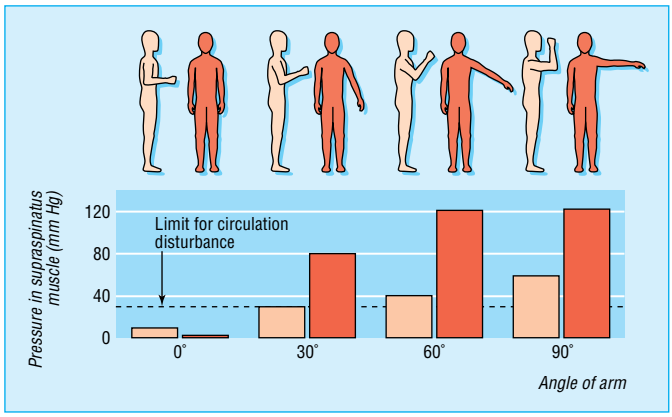
type I fibres and a reduced ratio of type I fibre areas to capillary areas. Strength training improves the performance of the type 2 fibres and there is reduced perceived exertion during work in patients with non-specific neck pain.

Another pain hypothesis is a relative shortage of energy in the muscle cells. When the energy demand in the muscle fibre is excessive, pain can result. The postural pain syndrome associated with sagging shoulders is a type of cervicobrachial pain that may be caused by prolonged stretching of the trapezius muscle or the brachial plexus. In cervical brachial pain syndromes, pain may be triggered by a pain locus in muscles, tendons, joint capsules, ligaments, or vessels. Nociceptors (pain receptors) in these loci may be the origin not only of the neck and shoulder pain but also of the referred pain to the arm and hand. The nociceptive pain may trigger a chronic pain syndrome that can affect the sympathetic nervous system. A possible pathogenic mechanism is that a small injury caused by a strain or a microrupture during some activity (work or leisure time) does not recover properly. Pain receptors induce a pathway of signals to the central nervous system by increasing the susceptibility to stimuli. The neurological response to normal activity is perceived as pain, and a chronic pain syndrome is the result. The predominant clinical symptom is activity related pain. Stiffness and severe pain at extreme postures are also common. The patient affected by chronic pain must be recognised as soon as possible for proper treatment and rehabilitation, preferably in a pain clinic.

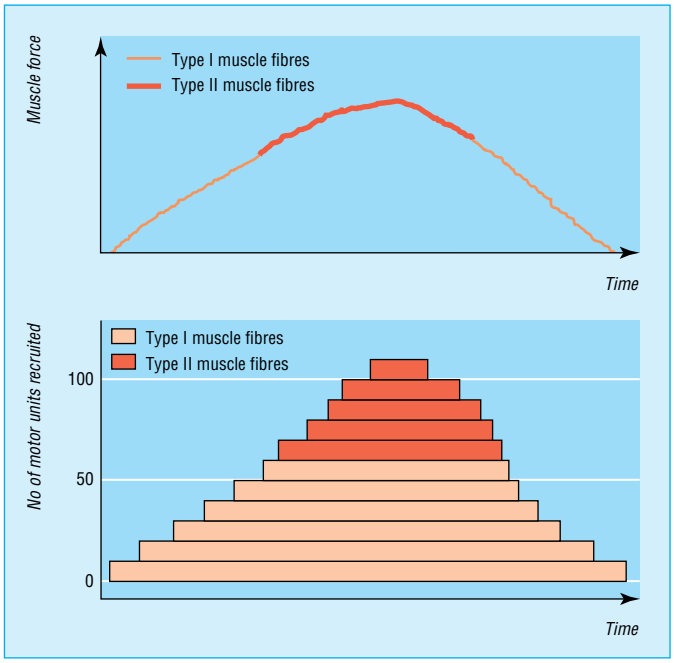
Awkward and static postures are common in players of musical instruments. Pain in the neck and arm have been related to gripping an instrument in an awkward posture. Pain in the left shoulder and arm in professional violinists can be the result of static holding of the violin with the left arm.

Neck flexion while working at a visual display terminal may be associated with non-specific shoulder symptoms. A prospective study showed that a non-optimal sight angle with the head overextended was related to neck symptoms, and extreme radial deviation of the hands was related to hand and arm disorders. An exposure-response relation has been found for neck pain and angle of neck flexion in keyboard operators: neck pain was more prevalent among operators who flex their necks more acutely. Incorrect glasses or the need for glasses when working at a visual display terminal may result in neck and shoulder pain, by affecting posture and because of muscle activity in the trapezius muscle caused by a reflex mechanism of oculomotor strain during sustained visual work at short distances.

The development of non-keyboard input devices, such as the computer mouse, has resulted in new postures that may cause a combination of symptoms from the wrist to the shoulder. Work tasks of long duration with a flexed and, to some extent, extended wrist have been reported as risk factors for carpal tunnel syndrome.



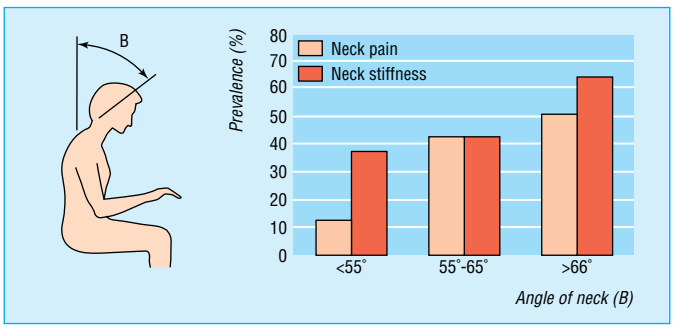
Intramuscular pressure in the supraspinatus muscle at different angles of abduction and forward flexion



Differential recruitment of muscle fibres with different levels of contraction. At low level static contraction, only type I muscle fibres may be recruited, leading to their selective fatigue and damage

Repetition motion

Repetitive motions of the shoulder may constitute a risk for rotator cuff tendonitis. An experimental study showed that women performing repetitive forward flexions of the shoulder developed shoulder tendonitis. Clinical signs of tendonitis were present up to two weeks after the experiments. Repetitive motions by industrial assembly workers (truck making, meat packing, and circuit board assembly) have been associated with the development of shoulder tendonitis, lateral epicondylitis, and tendonitis at the wrist (de Quervain's disease). Excentric exertion with injury of the extensor carpi radialis brevis muscle is one mechanical model for the pathogenesis of lateral epicondylitis.



Association between neck flexion and pain and stiffness in the neck

Repetitive motion, being a causal factor for tendonitis, is consistent with the high risk of shoulder tendonitis in competitive swimmers, and epicondylitis in tennis players.

Force—handling load or tools

Only a few studies have investigated the effect of handling loads on neck and arm symptoms. Handling heavy loads seems to be associated with osteoarthrosis and cervical spondylosis. Low frequency vibration exposure of high magnitude is associated with osteoarthrosis of the elbow, wrist, and acromioclavicular joints, and symptoms in the elbow and shoulder. Impacts, jerks, and blows with high energy transfer to the hands at low frequency might have the potential to result in musculoskeletal disorders, considering the general model for injuries. Furthermore, the observed associations with vibration exposure and musculoskeletal disorders might result from the strong dynamic and static joint loading and the repetitive hand and arm motions required in tasks where handheld machines are used.

Psychological and social factors

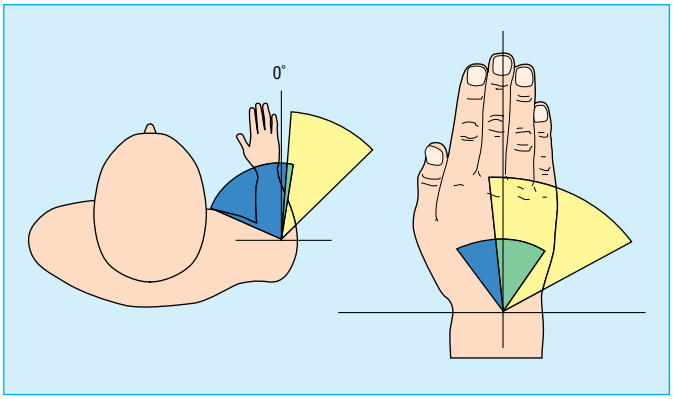
Psychological and social factors are generally more strongly associated with back pain than with shoulder pain. Furthermore, the association is stronger for non-specific pain than for pain with a specific diagnosis. This means that a diagnosis of general cervicobrachial pain may be more strongly related to psychological and social factors than are carpal tunnel syndrome or shoulder tendonitis. Highly demanding work and poor work content (repetitive tasks with short cycles) have been identified as risk factors for neck and shoulder pain. Psychological factors and personality type may be determinants of muscle tension and the development of myofascial pain.

Piece work is associated with neck and arm disorders when compared with work paid by the hour. This may be because of an increased work pace in addition to high psychological demand and low control in the work situation. Management style, in terms of social support to employees, is claimed to be associated with increased reporting of neck and shoulder symptoms. Social support from management obviously affects turnover of workers, and sick leave.

Psychological stress and burnout are associated with depression. Depressive moods are associated with musculoskeletal pain. It is likely that both psychological stress and chronic musculoskeletal pain can cause depressive moods. When assessing a patient with chronic musculoskeletal pain, a psychological evaluation and identification of possible affective disorders should be done. Treatment of depression can reduce musculoskeletal pain and facilitate return to work.

Individual susceptibility

Individuals may have increased vulnerability to injury because of disease, genetic factors, or lack of fitness. This individual susceptibility may result in a lower threshold for given exposures to cause work related musculoskeletal disorders. Additionally, the exposure may trigger symptoms earlier and at an unusual location because of localised vulnerability in a person who has preclinical systemic disease. As examples, a worker exposed to repetitive flexion in the shoulder developed tendonitis one year before developing rheumatoid arthritis. An electrician exposed to repetitive power grips and vibration developed symptoms and signs of carpal tunnel syndrome: at surgery these were found to be caused by amyloidosis. For work related musculoskeletal disorders individual factors usually have a low magnitude of risk compared with relevant ergonomic factors.



Outward rotation of the shoulder and ulnar deviation of the wrist may be found with use of a computer mouse (yellow) and keyboard (blue)

Work related musculoskeletal disorders found in blue collar and white collar workers

Shoulder pain

Blue collar workers—assembly workers
- Usually shoulder tendonitis due to working with hands above shoulder height
- Repetitive forward flexions

White collar workers—keyboard operators
- Usually non-specific cervicobrachial pain, which may be caused by task invariability leading to static tension of trapezius muscle

Hand and wrist pain

Blue collar workers—assembly workers
- Repetitive power grips may cause repetitive strain of extensor tendons and tendonitis
- Carpal tunnel syndrome may also be related to repetitive power grips

White collar workers—keyboard operators
- Intensive keying may cause repetitive strain of extensor tendons and tendonitis
- Carpal tunnel syndrome may also be related to intensive keying

Individual susceptibility to musculoskeletal disorders

Age
- For most musculoskeletal disorders, risk increases with age

Sex
- Among both the general population and industrial workers, women have a higher incidence of carpal tunnel syndrome and muscular pain in the neck and shoulder than men
- Whether this is due to genetic factors or to different exposures at work and at home is not clear

Anatomical differences or malformations
- A rough surface and the sharp edge of the intertubercular sulcus on the humeral head increases wear on the tendon of the long head of biceps muscle, which may make a person more prone to biceps tendonitis
- A cervical rib is a common cause of neurogenic thoracic outlet syndrome: a repetitive task may be the occupational exposure that triggers clinical disease
- Width of the carpal tunnel has been proposed as a risk factor for carpal tunnel syndrome, but there is no consensus

Further reading

- Ohnmeiss DD. Repeatability of pain drawings in a low back pain population. *Spine* 2000;25:980-8
- Lundeberg T. Pain physiology and principles of treatment. *Scand J Rehabil Med Suppl* 1995;32:13-41
- Swenson R. Differential diagnosis: a reasonable clinical approach. *Neurol Clin* 1999;17:43-63
- Black ER, Bordley DR, Tape TG, Panzer RJ, eds. *Diagnostic strategies for common medical problems*, 2nd ed. Philadelphia: American College of Physicians, 1999. *Gives basic information on how to evaluate tests and test performance—for example, predictive values and likelihood ratios*
- Hoppenfeld S. *Physical examination of the spine and extremities.* Connecticut: Appletom Century-Crofts, 1976. *Includes a detailed description of examining different parts of the musculoskeletal system, with extensive illustrations, with emphasis on neurological evaluations*
- McRae R. *Clinical orthopaedic examination.* Edinburgh: Churchill Livingstone, 1983. *Includes brief descriptions of musculoskeletal disorders in addition to extensive illustrations of examination technique*
- Saxton JM. A review of current literature on physiological tests and soft tissue biomarkers applicable to work-related upper limb disorders. *Occup Med (Lond)* 2000;50:121-30. *Concludes by proposing new ways that testing might be implemented during occupational health surveillance to enable early warning of impending problems and to provide more insight into the underlying nature of soft tissue disorders*
- Harrington JM, Carter JT, Birrell L, Gompertz D. Surveillance case definitions for work related upper limb pain syndromes. *Occup Environ Med* 1998;55:264-71. *Describes the consensus case definitions that were agreed for carpal tunnel syndrome, tenosynovitis of the wrist, de Quervain's disease of the wrist, epicondylitis, shoulder capsulitis (frozen shoulder), and shoulder tendonitis. The consensus group also identified a condition defined as "non-specific diffuse forearm pain," although this is essentially a diagnosis made by exclusion. The group did not have enough experience of the thoracic outlet syndrome to make recommendations*
- Hagberg M, Silverstein B, Wells R, Kuroinka I, Smith M, Forcier L, et al. *Work related musculoskeletal disorders (WMSDs): a reference book for prevention.* London: Taylor and Francis Ltd, 1995. *Themes are identification, evaluation, action, and change. The various chapters link work with tendon, nerve, muscle, joint, vascular, and non-specific or multiple tissue disorders; explore individual susceptibility; assess occupational risk factors; describe health and hazard surveillance techniques; discuss the management of change; outline training and education programmes; and give an overview of medical management*
- Health and Safety Executive. *Upper limb disorders in the workplace,* 2nd ed. Sudbury: HSE Books, 2002. *A practical guide on how to assess and minimise workplace risks through positive action*

8 Work related stress

Tom Cox

It is clear from large scale surveys of working people, and of those who have recently worked, that stress is currently one of the two main work related challenges to health. (The other is musculoskeletal disorders.) It is therefore not surprising that a plethora of guidance on work stress is available from government bodies, the social partners, and professional and scientific organisations, and it is unlikely that any individual or organisation could successfully claim ignorance of the topic or a lack of basic knowledge.

Myths and facts

"Work related stress is not a serious problem"
Wrong—in the United Kingdom, as many as one in five people report themselves to be suffering from high levels of work related stress. That's around 5 million workers. An estimated half a million individuals report experiencing stress at a level they believe has made them ill. The cost to Britain's economy is estimated at. **6.7 million working days lost per year**. It costs society between **about £3.7 billion and £3.8 billion**

Health and Safety Executive

What is and what is not stress?

Stress is not an illness; neither is it a meaningful descriptive term to apply to a situation such as a domestic scenario or a workplace, although they might be described as "stressful" or containing "stressors."

Stress is an emotional state that is very real for many people, and poses a major threat to the quality of their lives and to their health. Although that experience is rooted in the way the person sees and thinks about their world, it is essentially *emotional* in nature, normally involving a mixture of negative feelings, such as unpleasant arousal, apprehension, shame, guilt, or anger. It is not necessarily trivial.

The Ad hoc Group on Work Stress of the European Commission offered the following definition of work stress

Work stress is the emotional reaction to aversive and noxious aspects of organisations, work, and the work environment. It is a state characterised by extremes of arousal, and by discomfort and distress. It is often characterised by feelings of being out of control and helplessness. Stress can arise as a result of exposure to both physical and psychosocial hazards and may, in turn, affect not only psychological, physical, and social health, but also availability for work and work performance

Why do people experience stress?

Stress is the emotional state that results from someone's perceptions of an imbalance between the demands (pressures) on them, and their ability to cope with those demands. The control they have over related events and the support that they receive in coping are very important factors in this equation. Demands can be internally (self) generated as well as externally imposed, and a person's needs and expectations can be important in their experience of stress.

Classically, people at risk experience events that place demands on them with which they cannot cope. Their inability to cope may be because of lack of relevant knowledge or skill. They feel out of control and without support. Under these circumstances, they are more likely to experience stress and show the commonly associated patterns of cognitive, behavioural, and physiological change. Interestingly, although some of these changes may represent attempts at coping, others may be detrimental to coping. It is easy to see how a vicious cycle can quickly become established in that the person's ability to cope may be degraded by their experience of stress.

Stress can occur through work. It may be experienced as a result of exposure to a wide range of work related hazards and, in practice, often coexists with adverse influences operating outside the workplace

The correlates of stress

The experience of stress alters the way people think, feel, and behave. Many of the changes that occur are modest and potentially reversible, although detrimental to the person's quality of life at the time. Other changes may be more enduring, and have substantial consequences for health.

Behavioural changes may include increases in health risk behaviour, such as smoking and drinking, and decreases in health positive behaviour, such as exercise and relaxation. Many behavioural changes represent attempts to cope with the

Some factors affecting individual susceptibility to stress

- Individual constitution
- Lifestyle and work style
- Coping mechanisms
- Emotional stability
- Previous experiences
- Expectation
- Self confidence

emotional experience of stress—for example, by drinking more. However, this type of coping can easily become a secondary source of stress and ill health if sustained.

Evidence shows that cognitive stress is associated with poor decision making, impaired concentration, reduced attention span, impaired memory, and confusion. People who report "being stressed" also tend to admit to "not being able to think straight." Social behaviour and interpersonal relations may also be affected, possibly reflecting these and other psychological changes such as exhaustion and increased irritability.

The effects of stress are thought to contribute to a range of disorders as wide as cancer, heart disease, musculoskeletal conditions, skin disease, gastrointestinal disorders, and sexual problems. The evidence is strongest for links between certain types of prolonged stress and ischaemic heart disease, hypertension, and mental illness. Evidence also suggests that stress plays a part in the aetiology, course, and outcome (recovery from disability) of musculoskeletal disorders. Most of the evidence for such links is epidemiological. The pathophysiological mechanisms are not clear—perhaps the effects are direct (chemical mediators, effects on immunity) or indirect (the results of secondary, damaging behaviour).

It is likely that what is bad for the individual employee is also bad for their organisation. Organisational concerns associated with work related stress include high absenteeism, increased staff turnover, low job satisfaction, low morale, poor organisational commitment, poor performance and productivity, possible increased accident and near miss rates, and, in some cases, an increase in employee and client complaints and litigation.

Causes of stress at work

"Psychosocial and organisational" hazards refer to those aspects of the design and management of work and of its social and organisational contexts that are known to contribute to employee stress—so, to a lesser extent, do "physical" hazards such as noise and extremes of temperature. There is a reasonable consensus on the nature of the psychosocial and organisational hazards, and they have been divided into nine broad categories.

Managing stress at work

Work stress can be managed from two different perspectives: the individual and the organisational. The occupational health practitioner has a role to play in each approach.

Education, treatment, and rehabilitation: the individual

Much of little value has been written about individual stress management, and many weird and wonderful treatments are offered commercially. A healthy scepticism is warranted here as few of these treatments are based in scientific knowledge and even fewer have been evaluated.

Three strategies that might help the individual experiencing stress through work are: further education and training in relevant work or life skills, short term treatment for any medical condition, and managed rehabilitation to a normal pattern of working life.

Without doubt, the most effective form of stress management training is through a proper analysis of training needs in relation to the person's job; lifestyle counselling can also be valuable. Fundamental problems in the demands-ability balance may need to be examined. At the same time, reducing health risk behaviour and strengthening health positive

Some possible self reported symptoms of work stress

- Anxiety about work, continually agitated
- Continual complaints of unreasonable or unrelenting work demands
- Deep exhaustion
- Disturbed sleep and daytime tiredness
- Expressed dislike of work or work colleagues and low job satisfaction
- Feelings of being out of control or helpless
- Feelings of lack of support and care from others
- Forgetfulness
- Inability to concentrate, continually distracted
- Inability to think straight
- Irritability, being short tempered
- Loss of sexual interest, or impaired sexual performance
- Loss of the "big picture:" unable to get events into perspective
- Repeated absences from work

Psychosocial and organisational hazards: a taxonomy

Content of work

- **Task content:** lack of variety or short work cycles, fragmented or meaningless work, underuse of skills, high uncertainty
- **Workload and workpace:** work overload or underload, lack of control over pacing, time pressure
- **Work schedule:** shift working, inflexible work schedules, unpredictable, long or unsociable hours
- **Control:** low participation in decision making, lack of control over work

Context to work

- **Organisational culture and function:** poor communication, low levels of support for problem solving and personal development, lack of definition of organisational objectives
- **Role in organisation:** role ambiguity and role conflict, responsibility for people
- **Career development:** career stagnation and uncertainty, under or over promotion, poor pay, job insecurity, low social value to work
- **Interpersonal relations at work:** social or physical isolation, poor relations with superiors, interpersonal conflict, and lack of social support
- **Home-work interface:** conflicting demands of work and home, poor support at home, dual career problems

Adapted from Cox (1993)

Work related factors and ill health: the Whitehall II Study

This research concentrated on how the design of work affected people's mental well being and related health outcomes. The key findings were as follows:

- Having little say in how the work is done is associated with poor mental health in men and a higher risk of alcohol dependence in women
- Work requiring a fast pace and the need to resolve conflicting priorities is associated with a higher risk of psychiatric disorder in both sexes, and poor physical fitness or illness in men
- A combination of putting high effort into work and poor recognition of employees' effort by managers is associated with increased risk of alcohol dependence in men, poor mental health in both sexes, and poor physical fitness or illness in women
- A lack of understanding and support from managers and colleagues at work is associated with higher risk of psychiatric disorder. Good social support at work, particularly from managers for their staff, has a protective effect
- Aspects of poor work design is also associated with employees taking more sickness absence

Causes of stress and possible solutions

Poor management culture
Examples of good management are when:
- An organisation is committed to promoting the wellbeing of employees through good management practice
- The people who work in the organisation are valued and respected
- They receive support from the organisation if they wish to raise problems affecting their work

Poor relations
Examples of good relations are when:
- There is good communication between employer and employees, so that the employees understand what is expected, and the employer reacts to any problems experienced by the employees
- Employees are not bullied or harassed, and policies are in place to manage this

Role uncertainty
People understand their role when:
- They know why they are undertaking the work and how this fits in with the organisation's wider aims and objectives
- Jobs are clearly defined to avoid confusion

Too many demands
Demands are at the right level when:
- Staff are able to cope with the volume and complexity of the work
- The work is scheduled sensibly so that there is enough time to do allocated tasks; shift work systems are agreed with employees or their representatives; people are not expected to work long hours over an extended period

Poor management of change
Good change management includes when the organisation:
- Communicates to employees the reason why change is essential
- Has a clear understanding of what it wants change to achieve
- Has a timetable for implementing change, which includes realistic first steps
- Ensures a supportive climate for employees

Lack of control
People feel in control when:
- They are given a say in how they do their work
- The amount of control they have is balanced against the demands placed on them

Lack of training and support, and failure to take account of individual factors
Examples of good practice:
- Employees receive suitable and sufficient training to do their jobs
- Employees receive support from their immediate line management, even when things go wrong
- The organisation encourages people to share their concerns about health and safety and, in particular, work related stress
- The individual is fair to the employer—they discuss their concerns and work towards agreed solutions

behaviour such as exercise and relaxation may both improve the person's psychological and physical health, and offer a distraction from their problems.

If the person is affected by anxiety, depression, or some other stress induced illness, then that should be treated in the conventional way, possibly with drugs or psychological treatments, but always appropriately combined with education and rehabilitation. Managed rehabilitation is critical to the success of any treatment for work stress, and necessarily entails a dialogue between the occupational health practitioner and line management.

Prevention and an appropriate response: the organisation

Employers have a duty of care under common law to take reasonable and practicable steps to protect their employees' safety and health at work. This duty clearly extends to psychological as well as physical health, and to psychosocial and organisational as well as physical hazards. It is clear that an employer's failure to consider stress seriously can result in legal challenge. Employers also have duties under statutory health and safety law. Such law has evolved to prevent harm to employees through work, whereas common law allows for financial redress when harm has occurred. These two bodies of law are complementary, as are the duties they impose.

The occupational health practitioner can advise employers on two issues: prevention through risk management, and provision of employee support systems.

According to guidance from the Health and Safety Executive in the United Kingdom and the European Commission, work stress is to be treated as a health and safety issue and dealt with in organisations through the application of a risk management approach (essentially systematic problem solving). Organisations will need to include methods of assessing the risk from exposure to psychosocial and organisational hazards in their routine assessments and develop ways of reducing such risks if necessary. Methods to do this are

Expectations of a person experiencing stress through work
- Timely and appropriate support from both management and occupational health
- A professional and sensitive approach
- Help in solving the problem at source: moderating work pressures, providing education and training, increasing control over work events, and improving support
- Advice, if necessary, on lifestyle
- Short term treatment for any associated medical problems
- Active management of rehabilitation to work

available, and occupational health practitioners have a major role to play both as expert advisers and organisational champions.

The successful provision of employee support (to deal with stress) depends on three things: a broad based and competent system, an accessible system, and an integrated system. Most large organisations provide good employee support in theory, but fail themselves and their employees in practice because the overall system is fragmented, often competitive for resources, and territorial, and lacks internal collaboration at the case level. Much can be achieved by bringing existing systems together, by training staff in relation to work stress, and by marketing what is available within the organisation.

The box containing information on psychosocial and organisational hazards is adapted from Cox 1993. The box containing causes of stress and possible solutions is adapted from Health and Safety Executive. *Work related factors and ill health: the Whitehall II study*. Sudbury: HSE Books, 2000 (CRR 266/2000).

Further reading

- Cox T, Griffiths A, Rial-Gonzalez E. *Work-related stress.* Luxembourg: Office for Official Publications of the European Communities, 2000. *A compact overview of the professional and scientific literature on work stress; incorporates a discussion of the risk management approach to dealing with stress problems at work. Can be downloaded free from the website of the European Agency for Safety and Health at Work: http://osha.eu.int*
- Cox T. From environmental exposure to ill health. In: McCaig R, Harrington M, eds. *The changing nature of occupational health.* Sudbury: HSE Books, 1998. *This chapter in an edited volume in memory of Dr Thomas Legge provides a more detailed account of the model of stress referred to here, with more information from the organisational perspective. It also includes further discussion of the individual perspective*
- Griffiths A. The psychosocial work environment. In: McCaig R, Harrington M, eds. *The changing nature of occupational health.* Sudbury: HSE Books, 1998. *This chapter, in what has become known as the Thomas Legge book, focuses on the psychosocial work environment and provides an informed and detailed discussion of the psychosocial and organisational aspects of work, their design, and management. This chapter is usefully read in conjunction with that by the author in the same volume*
- Cox T, Griffiths A, Randall R. A risk management approach to the prevention of work stress. In: Schabracq MA, Winnubst JAM, Cooper C, eds. *Handbook of work and health psychology*, 2nd ed. Chichester: Wiley and Sons, 2002. *A detailed discussion of the risk management approach to work stress; includes a short series of organisational case studies to illustrate that approach in practice. The chapter touches on risk reduction strategies that focus on the individual employee as well as those that operate at the organisational level*
- Griffiths A, Randall R, Santos A, Cox T. Senior nurses: interventions to reduce work stress. In: Dollard M, Winefield A, eds. *Occupational stress in service professionals.* London: Taylor and Francis, 2002. *Provides a relatively detailed case study of an organisational intervention to manage work stress in a group of hospital based senior nurses applying the risk management approach. As with the above chapter, there is some discussion of risk reduction strategies that focus on the individual employee*
- Health and Safety Executive. *The scale of occupational stress: the Bristol stress and health at work study.* Sudbury: HSE Books, 2000 (CRR 265/2000)
- Health and Safety Executive. *Work related factors and ill health: the Whitehall II study—CRR 266/2000*, Sudbury: HSE Books, 2000

9 Mental health at work

Rachel Jenkins

Introduction

Mental illness affects about a tenth of all adults at any one time—about 450 million people worldwide. Lifetime prevalence is much higher. Mental disorders now account for about 12% of the global burden of disease and this is expected to rise to 15% by the year 2020. Neuropsychiatric conditions account for 30% of all years lived with disability.

Mental disorders and substance abuse are important issues in the workplace, partly because they are so common in the general adult population and partly because increasing rates of employment in many countries mean that the less able are entering the workforce. Mental ill health at work seems to be rising and in the United Kingdom at least, it has overtaken musculoskeletal disorders as the main cause of absence from work, long term sickness, and retirement on medical grounds. Whatever the cause of mental disorders, they have consequences for work performance and economic productivity. Appropriately tailored work is generally beneficial for people suffering mental illness, and the workplace can be an important setting for mental health promotion and the prevention of illness.

Positive mental health is not just the absence of mental disorder but has been defined as a positive sense of wellbeing, implying the presence of self esteem; optimism; a sense of mastery and coherence; the ability to initiate, develop, and sustain mutually satisfying personal relationships; and the ability to cope with adversity (resilience). Factors such as these enhance a person's capacity to contribute to family and other social networks, the local community, and society at large. They are also qualities that may be expected to influence work performance.

The spectrum of mental health problems

Mental disorder is common in the adult population of the United Kingdom, as elsewhere in Europe and the rest of the world.

Those who are unemployed have higher rates of mental illness than people in employment, partly because of the socially stressful aspects of unemployment and partly because people with mental illness experience more difficulty in finding and maintaining work.

Rates of mental illness in employed and unemployed, Great Britain 2000

	Working full time	Working part time	Unemployed	Inactive
	Rate per thousand (se)			
Neurosis (per thousand in past week)	136 (7)	161 (11)	196 (30)	270 (12)
Probable psychosis (per thousand in past year)	1 (1)	6 (2)	–	17 (3)
Alcohol dependence (per thousand in past 6 months)	94 (5)	53 (8)	146 (25)	67 (7)
Drug dependence (per thousand in past year)	40 (4)	32 (6)	137 (23)	40 (6)

Source: ONS survey of psychiatric morbidity among adults living in private households, 2000

The spectrum of mental health disorders

Disorder	Rough prevalence
Psychological distress usually connected with various life situations, events, and problems	Most of us from time to time
Common mental disorders (depression, anxiety disorders in adults, and emotional and conduct disorders in children)	10-20% of adults in general population but 40-50% in highly vulnerable populations; 30% of primary care attenders; 10% of children in general population
Severe mental disorders with disturbances in perception, beliefs, and thought processes (psychoses)	0.5% of general population
Substance abuse disorders (excess consumption and dependency on alcohol, drugs, and tobacco)	Highly country specific; 5% and above, increasing
Eating disorders	1-5%; mostly women
Abnormal personality traits that are handicapping to the individual and/or others	Not known; existing studies suggest 5%
Progressive organic diseases of the brain (dementia)	Senile dementia: 5% of over 65s and 20% of over 80s (hence the demographic time bomb)
Tropical organic dementias	Situation specific
AIDS dementia	A growing problem in countries where people with AIDS live long enough to develop it
Toxic organic brain syndromes	Industry specific (mercury, lead, carbon monoxide) or environmental

Prevalence of psychiatric disorders per 1000 population in adults aged 16-64 years in Great Britain 2000

	Women	Men	All adults
	Rate per thousand in past week (se)		
Mixed anxiety and depressive disorders	112 (6)	72 (5)	92 (4)
Generalised anxiety disorder	48 (3)	46 (4)	47 (3)
Depressive episode	30 (3)	26 (3)	28 (2)
Phobias	24 (2)	15 (2)	19 (2)
Obsessive-compulsive disorder	15 (2)	10 (2)	12 (1)
Panic disorder	7 (1)	8 (2)	7 (1)
Any neurotic disorder	202 (8)	144 (7)	173 (6)
	Rate per thousand in past year (se)		
Probable psychosis	5 (1)	6 (1)	6 (1)
Drug dependence	24 (3)	60 (5)	42 (3)
	Rate per thousand in past 6 months (se)		
Alcohol dependence	32 (3)	130 (6)	81 (4)

The prevalence of mental disorders in the workplace

Study	Number studied	Population	Instrument	Male	Female	Total prevalence per 1000
Fraser R, 1947	3000	Light and medium engineering workers	Medical assessment	283	360	300
Heron and Braithwaite, 1953	184	Colliery workers: Sedentary Surface manual Underground workers	Middlesex questionniare	 334 452 522		
Jenkins R, et al., 1982	162	Times journalists: 1 month after redundancy notice and 2 months before closure date 3 months after redundancy notice, when redundancy revoked and new owner arrived 12 weeks after threat of redundancy removed	Clinical interview schedule General health questionnaire			 378 378 324
MacBride R, et al., 1981	274	Air traffic controllers during an industrial dispute 4 months later 10 months later	General health questionnaire			 480 270 310
Jenkins R, 1985	184	Executive officers in civil service	Clinical interview schedule	362	343	
McGrath A, et al., 1989	171	Nurses Teachers Social workers	General health questionnaire			270 310 370
Stansfeld S, et al., 1994	10 314	Whitehall civil servants: Admin grades 1-7 Senior executive officer, Higher executive officer, Clerical	General health questionnaire	 248 247 216	 353 310 252	

Causes and consequences of mental disorder

Causes

The causation of mental disorder is multifactorial, being half genetic and half environmental for psychoses but largely environmental for the non-psychotic disorders.

Some disorders have a genetic basis, especially the major psychoses. Malnutrition can be a direct cause, whether in childhood or as an adult (for example, pellagra). Rarely, endocrine disorders such as myxodema may be causative. Occupational and environmental causes include infection (for example, encephalitis), the toxic effects of exposures at work (for example, mercury poisoning), and trauma (head injury).

Psychological factors—for example, poor coping skills and persistently low self esteem—also contribute. Such routine adverse life events as bereavement or job loss can lead to at least temporary mental disorder in the vulnerable. Unusually distressing or life threatening events may predispose towards the development of post-traumatic stress disorder. Such mechanisms are exacerbated by inadequate social support networks. Chronic social adversity (unemployment, poverty, illiteracy, child labour, violence, and war) is also often responsible, especially among underprivileged people.

Longitudinal studies have shown that unemployment, redundancy, or even the threat of redundancy cause mental illness, although naturally, employees who are already mentally ill are more likely to lose their jobs—either voluntarily or involuntarily. Given what is known about the mean rates of illness in the population as a whole and a higher rate in the unemployed, one would expect to find comparatively lower rates of illness in people at work. However, those studies that have been done in particular groups of workers have shown quite high rates of mental illness. It has been suggested,

Risk factors associated with common mental disorders: odds ratio (OR) of sociodemographic correlates of revised clinical interview schedule (CIS-R) score of 12 or more; *p < 0.05 **p < 0.01

	Adjusted odds ratio	95% confidence interval
Sex		
Male	1.00	—
Female	1.28**	1.11 to 1.47
Age (years)		
16-24	1.00	—
25-34	1.14	0.89 to 1.45
35-44	1.27	0.99 to 1.64
45-54	1.31*	1.01 to 1.69
55-64	0.71*	0.53 to 0.94
Family unit type		
Couple, no children	1.00	—
Couple with 1 + children	0.89	0.75 to 1.06
Lone parent + child	1.41*	1.08 to 1.83
One person only	1.23*	1.00 to 1.51
Adult with parents	0.44	0.26 to 0.75
Adult with one parent	0.71*	0.53 to 0.95
Employment status		
Working full time	1.00	—
Working part time	1.16	0.96 to 1.39
Unemployed	1.44*	1.02 to 2.01
Economically inactive	2.26**	1.92 to 2.66
Tenure		
Owner-occupier	1.00	—
Renter	1.41**	1.22 to 1.64
Locality		
Semi-rural or rural	1.00	—
Urban	1.16*	1.01 to 1.34

however, that some bias may have occured in studying working populations that were chosen because they are perceived to be particularly stress prone. The table on page 46 shows the strength of some of these risk factors in relation to mental illness in the United Kingdom.

The clinical interview schedule is a semistructured standardised clinical interview for use in epidemiological studies in the community, primary care, and workplace settings. It was originally devised to be used by mental health researchers, but has since been revised for use by lay interviewers with no mental health training.

Consequences

The development of mental illness is often followed by a series of psychosocial problems. Physical illness may occur, partly as a result of self destructive behaviour. Suicide is now the tenth leading cause of death worldwide. A descent in the social order is common, and with this comes poverty and secondary effects on social relationships, especially family ones. These potential developments are paralleled by effects on working life—for example, loss of job status or unemployment. The employer incurs the costs of sickness absence, impaired productivity, and increased devotion of time to human resources issues. The table shows the high level of social disability associated with mental disorder, both psychotic and non-psychotic.

Difficulties in activities of daily living in household samples

	% with any difficulties	N
People assessed as having …		
Suicidal thoughts in the past week	59	45
Probable psychosis in the past year	58	54
Neurosis in the past week	41	1376
None of the above	13	5919

The role of the employer

Whether or not a person's illness is contributed to by work, their workplace bears the consequences of the illness in terms of reduced productivity, sickness absence, labour turnover, accidents, and so on. It should be in the employer's interest to provide a good working environment, supportive if necessary, and to enlist some kind of occupational health service to detect and, sometimes, to help rehabilitate people with mental disorders in collaboration with other health and social agencies. In fact, such is the negative attitude of employers towards potential employees with such an illness that, far from offering support, they usually attempt instead to exclude. Mental disorders that have a substantial impact on everyday life are regarded as disabilities in the United Kingdom, and employers are forbidden to discriminate unreasonably against such people when offering employment. Instead, adjustments to working life must be entertained.

Less serious disorders that have little influence on everyday life, and drug and alcohol dependence, which are not covered by the Disability Discrimination Act, may nevertheless cause immense problems for employers and fellow employees. Mental disorders can be screened for but, rather like back pain, the lifetime prevalence is so high that excluding candidates with a history of mental disorder will simply reduce the potential workforce to unmanageable levels. Certain conditions, if declared, do probably render applicants potentially unfit for certain occupations: psychotic illness, personality disorder, and substance abuse for the caring professions; personality disorder and dependency disorders in safety sensitive jobs.

The role of Government

To support a successful economy and to make an appropriate contribution to the prevention of discrimination against people with mental illness, government agencies and other national bodies may need to take action on environmental conditions at work; access to employment, including sheltered employment for those who need it; opportunities for employment rehabilitation; the promotion of workplace mental health

Mental disorder is already prevalent within the workplace. Working conditions are known to have a considerable influence on mental health. Therefore, to minimise the damage from this source to both employees and employers, the most sensible course would seem to be for employers to institute mental health policies as part of their human resources framework

Workplace mental health policy

A workplace mental health policy is agreed between employers, employees, and their representatives—for example, trade unions, and includes:

- A statement that the organisation is committed to a course of action which might include
  - increased understanding of causes of mental health problems in the workforce
  - action to combat workplace stressors and helping staff to manage their stress
  - action to manage mental health problems effectively through early recognition and appropriate management
  - action to manage the return to work of those who have suffered mental health problems to ensure their skills are not lost to the enterprise
- Commitment to a healthy workforce, placing a huge value on both physical and mental health
- Acknowledging that mental health problems may have many causes, including stressors in the workplace and in the outside world
- Listing factors that may lead to increased stress levels in the organisation (customised, based on discussion with staff and needs assessment)
- Recognising that domestic factors (such as housing, family problems, and bereavement) may add to levels of stress experienced by employees

policies; and the provision of occupational health for the workforce. This is especially important now that many governments encourage the return to work of those who have suffered mental health problems, as well as those recovering from physical illness as part of "welfare to work" schemes. In times of full employment this may well increase the proportion of those at work who are psychiatrically vulnerable.

The role of schools in supporting subsequent occupational health initiatives

Besides their primary educational role, schools are important settings for mental health promotion. They can teach children important life skills aimed at reducing acute and chronic social stressors and enhancing social supports, all of which have a direct influence on mental health, and which may be expected to influence subsequent mental health in adult working life. Thus, employers as a body have an interest in encouraging mental health promotion in schools in the same way that they encourage mathematical and literacy skills, as well as physical health. Such mental health promotion should include teaching of coping skills, citizenship skills, exam skills and techniques, stress management, achieving potential in relationships and working situations, recognising and combating bullying, learning to say no to risky behaviours, and education about parenting and child rearing in collaboration with a health education and addiction programme.

Many schools now teach children "values"—respect for others', feelings, acceptance of differences in race, religion, etc. This can be established equally well in a workplace with a set of "company values" that go beyond the usually facile "mission statement"

The role of health professionals

Health professionals, including occupational health professionals, need to be adept at detecting and assessing mental health problems in the workplace. Managers may suspect mental health problems but they cannot be expected to diagnose or assess them, and they need help from health professionals in understanding and managing them. An occupational physician should be able to take an adequate psychiatric history, identify any possible physical agents responsible or stressors (in or out of work), and then perform a mental state examination to complete a risk assessment.

An occupational health professional's most important and unique contributions to helping manage people at work who have had or are experiencing mental health problems are to try to reduce stigma and discrimination, foster an understanding among managers and work colleagues, and advise on adjustments to the workplace when employees decompensate or when they return after a period off work because of mental illness.

The high rate of suicidal thoughts in people with depression means that teaching good assessment and management techniques to health and social care professionals should be a priority, as should national and local action to minimise environmental risk factors for suicide.

Mental state examination

- *Appearance and behaviour*—Grooming, hostility, restlessness, pupils, alcohol smell
- *Communication*—Rapid, sparse, confused
- *Mood*—Low or high, feelings of self worth, hopelessness, concentration, biological aspects (sleep, energy levels, appetite, libido), suicidal ideation
- *Thoughts*—Thought formation, thought content
- *Perceptions*—Hallucinations, etc.
- *Cognitive aspects*—Orientation, short term memory, knowledge of current affairs, neurological deficits
- *Insight*—Individual aware they are ill? Prepared to be treated?

Common mental disorders that may present in the workplace

Mixed anxiety or depression

Mixed anxiety or depression is the commonest disorder seen in occupational settings. People with this disorder may present with one or more physical symptoms—for example, various pains, poor sleep, and fatigue, accompanied by a variety of psychological symptoms. It is a prime cause of absence from

Diagnostic features of mixed anxiety and depression

- Low or sad mood
- Loss of interest or pleasure
- Prominent anxiety or worry
- Multiple associated symptoms
- Disturbed sleep
- Tremor
- Fatigue or loss of energy
- Palpitations
- Poor concentration
- Dizziness
- Disrupted appetite
- Suicidal thoughts and acts
- Dry mouth
- Loss of libido
- Tension and restlessness
- Irritability

work "due to stress." Together with related states, it contributes considerably to the disability accompanying musculoskeletal disorders, especially back pain, and to fatigue states.

Depression

Depression is common, with a lifetime prevalence in the United States of 17% for a major episode. The sufferer may present with physical symptoms, irritability, anxiety or insomnia, worries about social problems such as financial or marital difficulties, increased drug or alcohol use, or (in a new mother) constant worries about her baby or fear of harming the baby. Some groups are at higher risk—for example, those who have recently given birth or had a stroke, and those with physical disorders such as Parkinson's disease or multiple sclerosis.

Differential diagnosis
The differential diagnosis includes acute psychotic disorder if hallucinations or delusions are present; bipolar disorder if there is a history of manic episodes; poisoning or substance misuse if heavy alcohol or drug use has occurred; and chronic mixed anxiety-depression. Some medications may produce symptoms of depression (for example, β blockers, other antihypertensives, H2 blockers, oral contraceptives, corticosteroids). Unexplained somatic symptoms, anxiety, or alcohol or drug disorders may coexist with depression.

Diagnostic features of depression

- Low or sad mood
- Loss of interest or pleasure
- At least four of the following:
  - disturbed sleep
  - disturbed appetite
  - guilt or low self worth
  - pessimism or hopelessness about future
  - decreased libido
  - diurnal mood variation
  - poor concentration
  - suicidal thoughts or acts
  - loss of self confidence
  - fatigue or loss of energy
  - agitation or slowing of movement or speech
- Symptoms of anxiety or nervousness are also frequently present

Essential information about depression for patient, family, work colleagues, and managers

- Depression is a common illness and effective treatments are available
- Depression is not weakness or laziness
- Depression can affect a person's ability to cope

Information leaflets or audiotapes can be used to reinforce the information

Alcohol and drug misuse

Employees (or employers) with alcohol problems may present with depression, nervousness, insomnia, physical conditions such as peptic ulcer, gastritis, liver disease, hypertension, accidents or injuries, poor memory or concentration, and evidence of self neglect (for example, poor hygiene). They may be people in whom treatment for depression has failed. Patients may also have legal and social problems resulting from alcohol—for example, marital problems, domestic violence, child abuse, or missed work. Signs of alcohol withdrawal may be present—for example, sweating, tremors, morning sickness, hallucinations, and seizures. Those with alcohol problems often deny or are unaware of their problems, and it may be others who request professional help.

Management by the occupational health department
Employees may be referred with a suspicion of an alcohol problem or the possibility may be raised at the first interview. Assessment may be aided by simple well validated screening questionnaires such as the CAGE questionnaire and, for less excessive but still harmful drinking, the alcohol use disorders identification test (AUDIT) questionnaire.

The assessment should be conducted in a straightforward non-judgmental way and cover drinking pattern, amount, type, circumstances, and duration, as well as symptoms; convictions for drink driving should be specifically asked about. Laboratory tests may help diagnosis but have a limited use in isolation. They can help in patient education and in monitoring alcohol reduction, as can a drink diary.

Managing alcohol problems at work
This is best done in the context of an alcohol and drugs policy at work, which will always include a ban on the use of illegal drugs at work but which may have a variable attitude to alcohol at work, perhaps allowing alcohol to individuals whose jobs are not safety sensitive, for social occasions, or after the working day is over, etc. Whatever the policy, it needs to be signed up to

Alcohol dependency

The presence of three or more of the following suggests alcohol dependency

- Strong desire or compulsion to use alcohol
- Difficulty controlling alcohol use
- Withdrawal (anxiety, tremors, sweating, hallucinations) when drinking has ceased
- Tolerance—drinking large amounts of alcohol without appearing intoxicated
- Continued alcohol use despite harmful consequences

Presentation of alcohol problems at work

- Poor attendance—frequent sickness absence, certified or uncertified—may be regular—for example, after weekends or breaks
- Lateness for work
- Poor performance—mistakes, slowness, poor judgement, frequent mishaps
- Prolonged lunch hours, afternoon sleepiness
- Poor personal hygiene, scruffiness, smelling of alcohol
- Irrational or noisy behaviour, inappropriate comments, irritability
- Frequent disappearances during the day
- Signs of violence—cuts and bruises
- Dishonesty or deviousness
- Frequent sickness absence because of gastrointestinal upsets

by management and workers' representatives. If there is an Occupational Health Department or some kind of welfare service, then referrals by managers or individuals themselves for alcohol related problems should be possible, and the condition treated initially as a health problem, and only when there is refusal or inability to stop or reduce drinking to reasonable levels are disciplinary procedures invoked. Time off work as sick leave may be required. Referral to a general practitioner or alcohol misuse specialist will be necessary. Compliance with undertakings can be managed by an occupational health department, using random testing if required. The same process can be used for employees who use illegal drugs, although the very illegality of the drugs can lead to disciplinary measures much more quickly.

Both alcohol and drug abuse are chronic conditions, and any employer or Occupational Health Service has to realise the high probability of relapse, although research shows that rehabilitation is more likely to be successful when the problem is dealt with in a work context when the individual is threatened with potential job loss. Early recognition, assessment, and active management of the situation also help.

Alcohol and drug abuse is a serious problem for society and is clearly increasing in incidence; it is also a huge problem for employers. The yearly cost to industry of alcohol misuse has been estimated at about £3 billion in the United Kingdom through accidents, reduced productivity, and absenteeism. Hangovers alone have been estimated to cost industry £50-100 million.

Drugs of abuse other than alcohol can have serious effects on performance, probity, and so on. Testing for drugs at pre-employment or randomly is practised in some safety sensitive industries. The testing has to be done using proper chain of custody techniques and in the context of an agreed drugs policy, which may or may not allow for rehabilitation while still employed. Employers and occupational health professionals who undertake coercive testing for drugs of abuse must ask themselves whether by instituting this programme they are trying to exclude "undesirables" from this workplace or to identify those who, while under the influence of drugs, may present a safety or security risk? This issue raises concerns about human rights.

Women's issues

Women, by virtue of their increased exposure to acute life events, chronic social stresses, lower social status and income, and smaller social networks, are often particularly vulnerable to common mental disorders. This is reflected, hardly surprisingly, in higher rates of sickness absence because of psychological causes. Disorders associated with menstruation, pregnancy, and childbirth are additional disorders specific to women. New mothers often feel pressured to return to work early after childbirth, and one of the most important preventive actions that can be taken in the mental heath arena is to recognise postnatal illness and ensure adequate and prompt treatment.

Eating disorders

An eating disorder may be declared at a pre-employment screening. The two main types, anorexia and bulimia, of which the latter is more common, occur mainly in young women. An individual may present with binge eating and extreme weight control measures such as self induced vomiting, and excessive use of diet pills and laxatives, usually covert. In the case of employees, management may ask occupational health professionals for help because of concerns about an

CAGE questionnaire

Four questions:

- Have you ever felt you ought to **Cut** down on your drinking?
- Have people **Annoyed** you by criticising your drinking?
- Have you ever felt bad or **Guilty** about your drinking?
- Have you ever had a drink first thing in the morning to steady your nerves or get rid of a hangover ("**Eye-opener**")?

Over 90% of dependent drinkers answer "yes" to two or more of these questions

Alcohol use disorders identification test (AUDIT)

See Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the alcohol disorders idenitfication test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption. *Addiction* 1993;88:791-803

Management of alcohol dependence

Essential information for employees, managers, and families

- Alcohol dependence is an illness with serious consequences
- Ceasing or reducing alcohol use brings mental and physical benefits
- Drinking during pregnancy may harm the baby
- For people with alcohol dependence, physical complications of alcohol abuse or psychiatric disorder, abstinence from alcohol is the preferred goal
- In some cases of harmful alcohol use without dependence, or where the individual is unwilling to quit, controlled or reduced drinking is a reasonable goal
- Relapses are common. Controlling or stopping drinking often requires several attempts. Outcome depends on the motivation and confidence of the patient

Advice and support to patient and family

- Discuss costs and benefits of drinking from individual's perspective
- Give feedback about health risks, including the results of gamma glutaryl transferase and mean corpuscular volume measurements
- Emphasise *personal responsibility* for change and give clear advice
- Consider targeted counselling

For patients willing to stop now

- Set a definite day to quit
- Discuss symptoms and management of alcohol withdrawal (may require time off or even hospitalisation)
- Discuss strategies to avoid or cope with high risk situations (for example, how to face stressful events without alcohol, ways to respond to friends or colleagues who still drink)
- Help identify colleagues, friends, and family who will support ceasing alcohol use
- Discuss support after withdrawal
- Mention self help organisations such as Alcoholics Anonymous, which are often helpful

Concern has arisen that a history of such disorders makes candidates unsuitable for caring professions such as nursing or school teaching, but this is not necessarily the case. In this context, attention should be paid to any accompanying behavioural disorders including self harm—for example, and personality disorders, rather than uncomplicated eating disorders

individual's weight loss. Both anorexia and bulimia may present as physical disorders (for example, seizures or cardiac arrhythmias) that may have employment consequences and need treatment. Bulimia is, in general, a much more transitory condition with a better record of successful treatment. Anorexia nervosa is often more chronic and intractable and may involve prolonged sickness absence because of hospitalisation.

Bipolar disorder

Patients may present with a period of depression, mania, or excitement, or referral may be made by others because of the individual's lack of insight.

The diagnostic features of bipolar disorder are given in the box. Periods of either mania or depression may predominate. Episodes may alternate often or may be separated by periods of normal mood. In severe cases, patients may have hallucinations (hearing voices or seeing visions) or delusions (strange or illogical beliefs) during periods of mania or depression. The differential diagnosis includes poisoning or drug or alcohol misuse, which may cause similar symptoms.

Individuals often enter the hypomanic state rapidly, with danger to themselves and to others at work, especially if their job is safety sensitive. Some kind of early warning system should be instituted by the Occupational Health Department with the individual's consent and the cooperation of managers or sympathetic work colleagues.

High risk occupations

Certain occupations are at high risk for work related mental illness (and, incidentally, for fatigue states). These include occupations such as teaching, nursing, and the police force where there is a need for emotional commitment in the personal problems of other people and where there are considerable staff shortages, high demand, and poor locus of control.

Certain occupations are also at high risk for suicide. These include vets, doctors, dentists, pharmacists, and farmers—they have greater access to the means of suicide and better knowledge about effective methods of suicide, as well as being in demanding occupations.

Health professionals lead stressful lives, and epidemiological studies have confirmed the high levels of depression and anxiety in healthcare staff, indicating the need to address the support of this key group.

Employers are becoming increasingly worried—mainly for legal reasons—about the effect of demanding work on the mental health of vulnerable employees. This is a contentious area with little in the way of legal precedent but one where advice is frequently asked of occupational health professionals. Careful psychological assessment, knowledge of the job stressors, and a traditional risk assessment approach offer the best way forward. Attempts have been made, using a partially evidence based approach, to define health standards, including medical criteria, for entry into certain demanding professions—the armed forces, medicine, nursing, teaching, and civil emergency services. This can be helpful.

The potential for violence and bullying at work has also concerned employers, but such behaviour does not in fact usually emanate from those with mental illness but from those with problematic personality types or drug and alcohol problems.

Diagnostic features of bipolar disorder

Periods of mania characterised by
- Increased energy and activity
- Elevated mood or irritability
- Rapid speech
- Loss of inhibitions
- Decreased need for sleep
- Increased importance of self
- Persistent distraction

Periods of depression characterised by
- Low or sad mood
- Loss of interest or pleasure
- Disturbed sleep
- Guilt or low self worth
- Fatigue or loss of energy
- Poor concentration
- Disturbed appetite
- Suicidal thoughts or acts

Further reading

- World Health Organization collaborating centre for research and training for mental Health, eds. *WHO guide to mental health in primary care*. London: Royal Society of Medicine Press, 2000. *This is a pocket guide for the assessment, diagnosis, management, and criteria for referral of common mental disorders at primary care level. It has been specifically tailored for the United Kingdom by the WHO Collaborating Centre at the Institute of Psychiatry and contains the evidence base, resources including voluntary agencies, and a discussion of how to audit training needs*
- Andrews A, Jenkins R. *Management of mental disorders*. Aldershot: Datapress, 2000. *This is a two volume manual for the management of mental disorders, and is suitable for all members of the multidisciplinary team. It contains specific guidance on psychological therapies*
- Jenkins R, Bebbington P, Brugha TS, Farrell M, Lewis G, Meltzer H. British psychiatric morbidity survey. *Br J Psych* 1998;173:4-7. *This paper summarises the key findings from the first national survey of psychiatric morbidity in Britain. It shows the high prevalence of the psychiatric disorder and its association with sociodemographic and social risk factors*
- Jenkins R, Macdonald A, Murray J, Strathdee G. Minor psychiatric morbidity and the threat of redundancy in a professional group. *Psych Med* 1982;12:799-899. *This paper shows the psychiatric impact of the threat of redundancy on* Times *journalists*
- Jenkins R. Minor psychiatric morbidity and labour turnover. *Br J Ind Med* 1985;42:534-9. *The paper shows that the presence of minor psychiatric morbidity (depression and anxiety) is associated with substantially increased labour turnover, with associated costs for employers*
- Jenkins R. Minor psychiatric morbidity in civil servants and its contribution to sickness absence. *Br J Ind Med* 1985;42:147-54. *This paper describes the substantial association between minor psychiatric morbidity and sickness absence both retrospectively (the 12 months before the assessment) and projectively (the 12 months after the assessment), again with associated costs for employers*
- Jenkins R, Warman D, eds. *Developing mental health policies in the workplace*. London: HMSO, 1993
- Jenkins R, Coney N. *Promoting mental health at work*. London: HMSO, 1992. *These two books look at the business case for action to develop mental health policies in the workplace, from the CBI and TUC perspective, and examines a range of good practice examples from different companies*
- Department of Health. ABC of mental health in the workplace. London: HMSO, 1996. *This government publication sets out the key elements of a workplace mental health policy*
- Jenkins R. Public policy and environment. In: Gelder M, ed. *Oxford textbook of psychiatry*. Oxford: Oxford University Press, 2000. *This chapter sets the issue of mental health in the workplace in the context or overall public policy and mental health*
- Miller DM, Lipsedge M, Litchfield P, eds. *Work and Mental Health—an employer's guide*. Gaskell and Faculty of Occupational Medicine, 2002. *Straightforward and comprehensive account of the impact of mental health problems on work and how to deal with them.*

The third area of concern is safety. The potential problems with psychotic or dementing employees and employees who misuse drugs or alcohol will be obvious. The assessment of less serious mental disorders and their relation to safety is a job for the occupational physician using a similar approach to that described for assessing those entering demanding jobs. Psychotropic medication often affects cognition, especially at the beginning of treatment; drowsiness and lack of concentration are common and should be anticipated.

Mental ill health at work is likely to become the dominant occupational health issue of the future. There is enormous scope for research and enormous need for public education and the destigmatisation of mental illness.

The table showing the prevalence of psychiatric disorders and the table showing rates of mental illness employed and unemployed are adapted from ONS survey of psychiatric morbidity among adults living in private households. London: HMSO, 2000. The table showing the prevalence of mental disorders in the workplace is adapted from Jenkins R. Public policy and environment. In: Gelder M, ed. *Oxford textbook of psychiatry*. Oxford: Oxford University Press, 2000. The table showing risk factors associated with common mental disorders and the table showing difficulties in activities of daily living are also adapted from ONS survey of psychiatric morbidity among adults living in private households. London: HMSO, 2000.

10 Human factors

Deborah Lucas

The term "human factors" is often invoked after an accident, whether a minor incident in the workplace or a major disaster entailing significant loss of life. In many respects "human factors" is regarded by the layman as being synonymous with "human failure"—an unavoidable aspect of the human condition. Although there is a long list of major accidents across all hazardous industrial sectors where human failures were causal factors, this is not to imply that human errors are inevitable. Research over the past 20 years has shown much about the origins of different types of error and the best means of reducing their occurrence. However, the loss of life in disasters such as the Clapham Junction rail crash in 1988, the Southall and Ladbroke Grove train crashes in 1997 and 1999, respectively, and the sinking of the Herald of Free Enterprise in 1987 are high in the British public's mind. All of these disasters had human factors as a cause: a maintenance worker not disconnecting a wire, a train driver passing a red danger signal, and a bosun failing to close the bow doors of a ferry. The nuclear industry faced up to the issue of human factors after Three Mile Island in 1979 and the Chernobyl accident in 1986. The oil sector recognised the issue after the Piper Alpha tragedy in 1988. The aviation, rail, and marine transportation sectors are all actively considering the issue of human factors. Proper consideration of human factors is a key ingredient of effective health and safety management in all industrial sectors.

Modern health care is also a complex and, at times, high risk activity where adverse events are inevitable. However, a substantial proportion of adverse events results from preventable human failure by medical staff. Adverse events occur in about 10% of admissions to hospital in the United Kingdom—a rate of 850 000 adverse events a year. In the United Kingdom, 400 people die or are seriously injured every year in adverse events involving medical devices. Hospital acquired infections are estimated to cost the NHS nearly £1 billion every year, but about 15% of such infections may be avoidable. In the United States it is estimated that between 44 000 and 98 000 people die annually because of medical errors. Yet health care is not unique. There are many parallels with other high risk sectors, which have been examining the need to reduce human failures in complex systems for over two decades.

Definition

Human factors are often described as the thread that runs through all the key health and safety management issues, and numerous definitions of human factors and the related term ergonomics exist. The definition given by the UK Health and Safety Executive is "Human factors refer to environmental, organisational and job factors, and human and individual characteristics, which influence behaviour at work in a way that can affect health and safety." Key elements have been identified by psychologists and ergonomists after an incident or accident, and in the military field human factors programmes explicitly consider six aspects or domains during the design or procurement of a system. These domains have been found to be useful in other industrial contexts.

"Human error" is often cited immediately after a disaster

Examples of human failures in medicine

- A patient is inadvertently given a drug that they are known to be allergic to
- A clinician misreads the results of a test
- A child receives an adult dose of a toxic drug
- A patient is given medicine that has a similar sounding name to that prescribed
- A toxic drug is administered by the wrong route—for example, intrathecally
- A heart attack is not diagnosed by emergency room staff in an older patient with ambiguous symptoms

Common errors relating to drugs

- Unavailable drug information (for example, lack of up to date warnings)
- Miscommunication of drug orders (for example, through poor handwriting, confusion between drugs with similar names, misuse of zeros and decimal points, confusion between milligrams and micrograms)
- Incomplete patient information (such as not knowing about other medicines they are taking)
- Lack of suitable labelling when a drug is repackaged into smaller units
- Workplace factors that distract medical staff from their immediate tasks (such as poor lighting, heat, noise, and interruptions)

Human factors considered in the development of military systems

Domain	Issue	Issues to consider
Staffing	How many people are needed to operate and maintain the system?	Workload Job descriptions Staffing levels Team organisation
Personnel	What human characteristics, including aptitudes and experience, are needed to operate and maintain the system?	Selection and recruitment Career development Required qualifications, competences, and experience Specific characteristics
Training	What is the best way to develop and maintain the required knowledge, skills, and abilities to operate and maintain the system?	Training needs analysis Documentation Assessment Team training Skill maintenance and update
Human factors engineering	How can human factors be built into the system design to optimise human performance?	Equipment design Workstation design Workplace layout User interface design Maintenance access
Health hazards	What are the short term and long term health hazards from operation of the system?	Minimising exposure to health hazards such as toxic materials, electricity, musculoskeletal injury, noise and vibration, extremes of temperature
System safety	How can safety risks that humans might cause when operating or maintaining the system be avoided?	Sources of human errors Effects of misuse of equipment Abnormal and emergency situations

Human failure

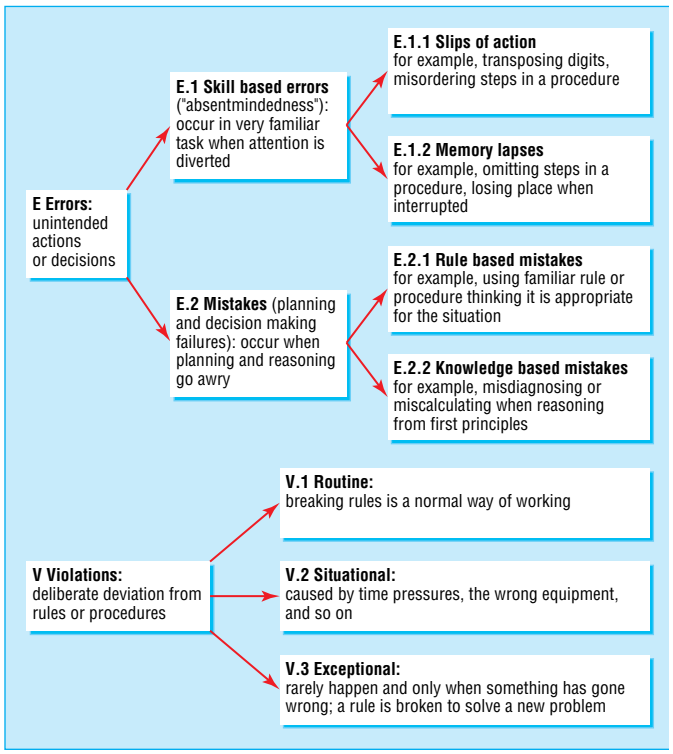
Research across industries has shown much about the types of human failure and the underlying psychological mechanisms. A key distinction can be made between unintended human errors and deliberate rule violations. However, even deliberate violations can result from system pressures such as shortage of time because of a lack of staff, or the correct equipment not being available. In high hazard industries it is no longer acceptable to attribute a safety incident just to a "human error" with the assumption that this was somehow beyond the control of managers and safety management systems. A detailed

Typical causes of human failures in accidents

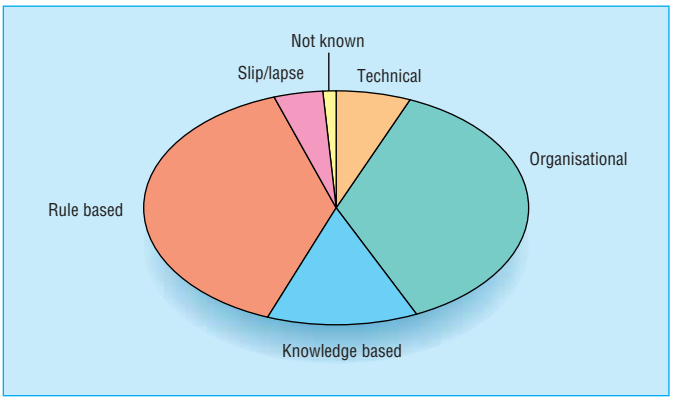
Job factors
- Illogical design of equipment and instruments
- Constant disturbances and interruptions
- Missing or unclear instructions
- High workload
- Noisy and unpleasant working conditions

Individual attributes
- Low skill and competence levels
- Tired staff
- Bored or disheartened staff
- Individual medical or fitness problems

Organisational aspects
- High work pressure because of poor work planning
- Poor health and safety culture
- One way communications (messages sent but no checks to ensure they are received or are appropriate)
- Lack of safety systems and barriers
- Inadequate responses to previous incidents



Classification of the types of human failure



Causes of incidents in a department of surgery

investigation into the causes of incidents involving human failure will show a number of immediate and underlying causes and contributing factors. Many of these will be problems with organisational systems rather than with the individual member of staff.

Control measures

There is no "magic bullet" for the problems of human fallibility. However, thoughtful, multifaceted approaches can reduce the probability of human failures leading to serious consequences. In medicine, knowledge and tools to enhance patient safety are emerging, and much can be learnt from other industries, particularly the high hazard sectors such as the nuclear industry, aviation, and transportation.

Case study: reducing errors in the administration of intravenous heparin

The intravenous administration of heparin (an anticoagulant) is a complex procedure, and this drug has been the subject of serious drug errors. A US hospital wanted to reduce errors in the administration of heparin in cardiac care units. They developed a form that combined the ordering and recording of the use of heparin. In addition, they improved communication with the hospital laboratory, converted all heparin protocols to pharmacy managed protocols, introduced pre-typed heparin orders and the double-checking of pump programming, and encouraged the use of low molecular weight heparin instead of standard heparin. These control measures were claimed to have reduced drug errors by 66%

Designing for people

Many sources of human error can be removed through effective design of equipment and procedures. Such "error tolerant" designs consider the tasks that the equipment is intended for and the errors the user may make. To give an example, in the early days of automatic teller machines, the user's bank card was returned to them by the machine *after* their cash had been issued. Banks found that many people took their money but forgot to take their card. This error was prevented by returning the card before the cash appeared.

Consideration of human factors is an important aspect of overall design and equipment procurement, and should be considered early in the design process. If left too late, then complicated procedures, added warnings, and requests for the user to "take care" can be the unfortunate result. Compliance with instructions and procedures differs according to the situation, the risks, the element of personal choice, and the probability of being detected. Written warnings are usually noticed, read less often, and complied with infrequently.

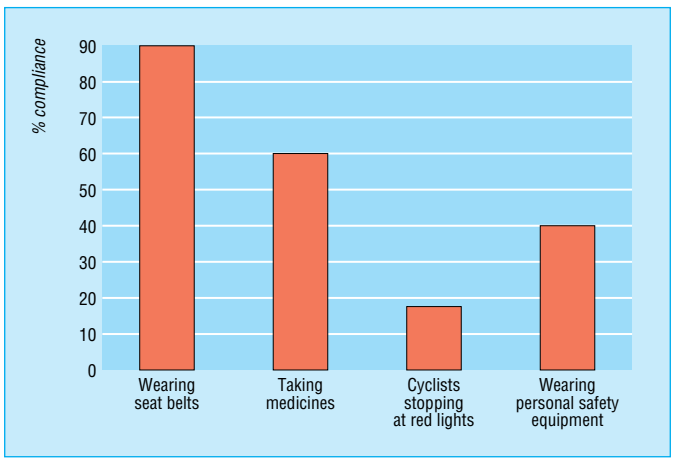
Poorly designed equipment can directly influence the chance of human errors occurring. For example, the layout of controls and displays can influence safety if switches are placed so that they can inadvertently be knocked on or off, if controls are poorly identified and can be selected by mistake, or when critical displays are not in the user's normal field of view. The controls of different equipment may not be compatible: for example, a switch in the up position may be "on" in one

"Human beings make mistakes because the systems, tasks, and processes they work in are poorly designed"

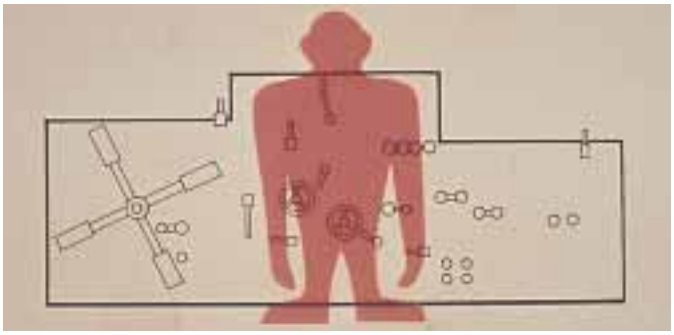
Dr Lucian Leape, testifying to the US President's Commission on Consumer Protection and Quality in Health Care

Examples of ergonomic criteria for procuring equipment

- Does the equipment suit the body size of all users?
- Can users see and hear all they need to easily?
- Is it easy to understand the information displayed?
- Would the equipment cause discomfort if used for any length of time?
- Is it easy to learn how to use the equipment? Are instructions and any warning signs clear? Is the language used appropriate for the users?
- What errors may occur? Can these be detected easily, and corrected?
- Is the equipment compatible with other systems in use?
- Can users reach controls easily?
- Can users move safely between operating positions?
- Is the equipment too noisy, does it vibrate too much, is it paced too fast?



Compliance rates in different situations



Arrangement of controls on a lathe and the "ideal" operator, who should have the following dimensions—4 feet 6 inches tall, shoulders 2 feet across, and an 8 foot arm span!

case but "off" in another. Alarm systems may be designed so that high priority alarms are not clearly differentiated and are thus easily missed.

Designing tasks, equipment, and workplaces to suit the users can prevent or reduce human errors and thus reduce accidents and ill health. A key message is that effective use of ergonomics will make work safer and more productive.

Sleep and human performance

Although it is often feasible to prevent human failures by the effective design of jobs and equipment, in other situations human performance problems may arise as the result of fatigue, shiftwork, poor communications, lack of experience, or inadequate risk perception. These aspects all need to be managed effectively to reduce the potential risks. One of the most commonly cited problems is lack of sleep for staff carrying out safety critical tasks. The decision to launch the Challenger space shuttle was partly attributed to the effects of fatigue on the decision making team. The rail crash at Selby in the United Kingdom in 2001 occurred because a car driver fell asleep and drove onto a railway line.

A significant proportion of road traffic accidents occur between 2 am and 5 am and are attributed to drivers falling asleep at the wheel. As we are not a nocturnal species, this is the time when our biological clock programmes us to sleep. Such circadian rhythms are hard to adjust to, even when working regular night shifts. Many people work shift systems, do night work, or work very extended hours including significant levels of overtime. Such working patterns can have adverse effects on their health as well as being associated with poorer performance on tasks that need attention or sustained vigilance, decision making, or high levels of skill. Sleep is a powerful biological need, and night work or certain shift systems can disrupt both the quantity and the quality of sleep. Sleeping during the day is never as satisfactory as sleeping at night. Sleep loss of just a few hours over a few days can lead to a build up of a sleep debt and reduced performance, but the person may not be aware of this.

A large body of research on shiftwork exists, but often the findings are not put into practice. Working patterns are usually seen as matters to be negotiated between employees and the employer, and additional overtime can be perceived as a financial advantage, and not as a potential health and safety issue. However, in high hazard industries awareness of the relation between sleepiness and accidents is growing.

Organisational influences

A number of factors within an organisation are associated with good safety performance. These affect not only human factors issues but also the "safety culture" of the organisation. A "culture" means shared attitudes, beliefs, and ways of behaving. An effective culture will be shown through good ways of informing and consulting all staff, recognition that everyone has a role to play in safety, visible commitment by managers to involving all staff, cooperation between members of the workforce, open two way communications, and high quality of training. The organisation that continually improves its own methods, and learns from mistakes (including accidents and "near misses") will tend to have a better safety performance than one that blames individuals for "being careless" when accidents happen.

The relationship between sleepiness and accidents: best practice approaches to managing the problem

- Plan shift rosters to take biological rhythms into account
- Set limits for maximum hours of duty and time needed for recovery afterwards
- Educate shift workers on sleep routines, nutrition, and exercise
- Make environmental changes to the workplace including lighting, temperature, and comfort level, which can all influence alertness
- Plan safety critical tasks to avoid night shifts
- Provide medical advice for shift workers
- Recognise the possibility of true sleep disorders (sleep apnoea, narcolepsy) and referral for investigation and treatment

High hazard industries are becoming increasingly aware of the importance of proper consideration of human factors

The Herald of Free Enterprise sank because no effective system was in place to ensure the bow doors were closed

Key principles

Human factors is a broad concept that can be seen as too complex or difficult to do anything about. However, there are five key principles to be remembered, and these are ones that many regulatory bodies are promoting:

- Recognise that people do not make mistakes because of "carelessness" and accept that even the most experienced members of staff are vulnerable to unintentional errors.
- Learn from adverse events including "near misses." Understand that usually there will be no single cause of an incident but a number of causes and contributing factors.
- Anticipate the influences on human performance. Key themes will include time pressure, experience, staffing levels, fatigue, and risk communications.
- Defend against paths to failure. In particular, appreciate the role of designing equipment and systems that are error tolerant.
- Encourage a "culture of safety."

Further reading

- NHS. *An organisation with a memory*. London: NHS Publications, 2000
- Reason J. *Human Error*. Cambridge: Cambridge University Press, 1990
- van der Schaaf TW, Shea CE. *MECCA: Incident reporting lessons from industry applied to the medical domain*. Conference on examining errors in health care, California: Rancho Mirage, 1996
- Reason J. *Managing the risks of organisational accidents*. Ashgate Publishing, 1997. *Seminal work on the causes of major accidents. A key influence for those looking at medical errors*
- Institute of Medicine. *To err is human: building a safer health system*. Washington DC: IOM, 1999. *Significant US report on medical errors; draws attention to the scale of the problem of potentially avoidable events that result in unintended harm to patients*
- *Building a safer NHS for patients*. London: NHS Publications, 2001. *Describes how promoting patient safety by reducing error is becoming a key priority of health services around the world. Sets out steps to implement a programme to reduce the impact of error within the NHS*
- HSE. *Reducing error and influencing behaviour*. Sudbury: HSE Books, 1999. *Guidance to industry on understanding and control of human factors in health and safety management. Covers understanding human failures, designing for people, and control measures for human errors*
- Noyes J. *Designing for humans*. London: Taylor and Francis. 2001. *Overview of human-machine interaction and the design of environments at work, with focus on health and safety at work*
- Moore-Ede M. *The 24-hour society: the risks, costs and challenges of a world that never stops*. London: Piatkus, 1993. *Introduction to the role of sleep in accidents. Covers biological aspects of sleeping and shiftwork*

11 Physical agents

Ron McCaig

The use of the term "physical agents" is not always clear. Sometimes it is taken to mean dusts and fibres whose effects are determined by their physical properties as well as their chemical composition. However, the term usually refers either to those agents that impart energy to the body by physical means (for example, the effects of radiation, heat, or noise and vibration), or to the effects of environments that differ in their physical characteristics from that existing at ground level on dry land (for example, found in diving and compressed air work, at altitude, and in flight).

The effects of physical agents have been well studied, and for many of these exposure criteria are now established at an international level. Fatalities are only likely to occur where established safety procedures are broken

The body offers some protection against physical agents experienced in the normal environment, such as heat and radiation—for example, by the physiological changes of heat acclimatisation or, at a cellular level, the operation of DNA repair mechanisms. Such mechanisms are limited in their effectiveness and can be overwhelmed if challenged by an exposure of sufficient magnitude. Even in artificial environments, such as work in compressed air tunnelling or high accelerations in flight, it is possible that a certain amount of physiological adaptation can take place. For example, the incidence of decompression illness often reduces after the first few days of exposure of a work force tunnelling in compressed air—an effect that is thought to be a form of acclimatisation—and some G tolerance can develop with physical fitness training.

Many physical agents have a threshold of exposure below which the body is unlikely to be harmed. Beyond that, it is necessary to restrict exposures, often by administrative controls such as limiting the duration of exposure (as in work rest schedules in the heat), providing shielding or protective clothing and equipment, or limiting the potential for harm by procedures such as staged decompression. Exposures must be carefully managed as some physical agents can kill within quite short periods.

Before exposure to hot, cold, or hyperbaric environments it is important to ensure that individuals have no predisposition to suffer from the effects of the environment. Fitness standards may be available, published by a variety of agencies. For ionising radiation it is important to know that individuals are medically fit for the type of work that they are expected to perform. (They may need to wear protective equipment—for example.)

Authorities that set exposure standards for physical agents

- International Standards Organisation (ISO)
- American Conference of Governmental Industrial Hygienists (ACGIH)
- International Commission on Radiological Protection (ICRP)
- International Commission on Non-Ionising Radiation Protection (ICNIRP)
- Other national, transnational, and international authorities

The principles of managing work in hot environments

- An assessment of the risk should be undertaken and ways sought to reduce the environmental heat load, paying attention to humidity and radiant heat, as well as air temperature
- Individuals should be screened for medical conditions that may predispose to heat illness, and should be physically fit, well hydrated, and ideally below 40 years of age
- Work-rest regimes should be established from published standards and adhered to, with regular opportunities taken for the worker to cool down
- Workers should be educated about heat illnesses, and first aid facilities should be available
- In planning work, the state of acclimatisation of the workers and the resistance to heat loss provided by their clothing has to be taken into account

Heat

Regulation of the central (core) body temperature is an essential physiological function—core temperature must be within the range 36-38°C for the body to perform efficiently. In the face of heat gain from the environment or as a result of exercise, the body defends the core temperature by vasodilatation (increasing skin blood flow) and by sweating.

Heat acclimatisation increases the magnitude of these responses. Any factor that impairs either the circulation or the ability to sweat will compromise thermoregulation

If heat gain is greater than heat loss by the evaporation of sweat, convective cooling, and thermal radiation, then the body stores heat. As it does so, the temperature of the brain and central organs (such as the liver)—the core temperature—increases and this threatens the survival of the individual. Eventually external cooling must be provided to prevent death. Heat hyperpyrexia (heat stroke) is the most serious effect of exposure to heat. It is generally characterised by a body

temperature of 40-41°C, an altered level of consciousness, and a hot dry skin resulting from failure of the sweating mechanism. These features are not invariable, however, so treatment should not be delayed if heat stroke is suspected.

Heat exhaustion results from a combination of thermal and cardiovascular strain. The individual is tired and may stumble, and has a rapid pulse and respiration rate. The condition may develop into heat stroke if not treated by rest, cooling, and fluids. Other effects are heat syncope (fainting), heat oedema, (often in the unacclimatised), heat cramps, and heat rash (prickly heat). Working in high temperatures can also result in fatigue and an increased risk of accidents.

Workers in fire and rescue services may be exposed to extreme heat in an unpredictable manner. Their safety depends on proper selection, training, and monitoring of the duration of exposure. Personal heat stress monitors are not yet widely available, but their use in these circumstances may confer some benefit.

Cold

In cold conditions the problem is to balance heat produced by physical activity with heat lost to the environment. The rate of heat loss depends on the insulation of the clothing and the external climate, including air temperature and wind velocity. The windchill index (derived in units of kcal/m^2/hour) relates to the risk of freezing of superficial tissues, and this, or the related chilling temperature (expressed in °C), is quite widely used as a measure of the discomfort of cold conditions.

The insulation of clothing may be impaired by moisture in the form of condensed sweat or by precipitation. Protection is generally easier in cold dry environments such as mountains or arctic regions than in cold wet conditions. The protection of individuals who are active in cold wet environments, and who need waterproof external garments, is only partly solved by the introduction of "breathable" fabrics. A particular problem occurs in those environments where there is a risk of immersion in cold water, with resulting catastrophic loss of insulation. Where this risk can be anticipated—for example, in helicopter flights over water, protective immersion suits should be used.

Large numbers of workers are employed indoors in conditions of moderate to severe cold, mostly in food preparation and storage. Only a few people are exposed to cold in scientific and testing laboratories. Cold stores can operate at temperatures as low as −30°C. Workers in cold stores must be provided with proper insulated clothing, and they must have regular breaks in warm conditions. A major problem in severe cold, indoors or outside, is to keep the hands and feet warm. The necessary insulation is bulky, which is less of a problem for footwear than for hand wear. Mitts provide better thermal protection than gloves, but limit dexterity.

Indoors, in moderately cold conditions—that is, temperatures below 15°C, it may also be hard to maintain comfort of the extremities, and exposure to draughts can be particularly troublesome. Limited evidence indicates that workers regularly exposed to cold conditions such as these may have worse than average general health.

Serious hypothermia should not occur in occupational settings. If there is a risk, people should not work alone, should have good communications with others, and should be trained in first aid management of the effects of cold. Hypothermia is treated by slow rewarming using the individual's own metabolism, and copious insulation, possibly supplemented by body heat from another person.

Groups of people at risk from heat illness

- Unacclimatised workers in the tropics
- Workers in hot industries who have had a break from exposure
- Workers with an intercurrent illness
- Workers in the emergency services—for example, fire or mines rescue
- People undertaking very heavy physical activity—for example, military recruits
- People working even moderately hard at normal temperatures in all enveloping protective clothing—for example, fire crews dealing with chemical spills
- Older people and the very young when ambient temperatures are raised for prolonged periods

The wet bulb globe temperature

- The wet bulb globe temperature (WBGT) index is an index of heat stress. It is derived from the natural wet bulb temperature (WB), the dry bulb temperature (DB), and the globe temperature (GT) (a measure of radiant heating) in the ratio:

$$WBGT = 0.7\,WB + 0.2\,GT + 0.1\,DB$$

- The WBGT index is measured using a "Christmas tree" array of thermometers, or purpose built electronic sensors and integrating apparatus
- The index was originally derived to protect troops exercising outdoors by relating environmental conditions to the risk of heat illness. It has since been developed and used extensively in industry and is the basis for International Standard 7243 and guidance by the ACGIH. These documents give upper boundaries of WBGT value for continuous and intermittent work of different intensities. Other standards apply in relation to thermal comfort—for example, ISO 7730

Heat stroke

Heat stroke is a medical emergency. The body temperature should be lowered by tepid sponging and fanning with cool air.

Intravenous fluids may be necessary. The following may predispose to heat exhaustion and heat stroke:

- Obesity
- Lack of fitness
- Age 50 years or more
- Drug or alcohol abuse
- History of heat illness
- Drug treatment (for example, antihistamines, tricyclic antidepressants, or antipsychotics)
- Pre-existing disease of cardiovascular system, skin, gastrointestinal tract, or renal system

Frostbite in an outdoor worker

The peripheral effects of cold are frost nip, frost bite, and non-freezing cold injury. Frost nip appears as a white area on the skin, and in frost bite the appearance is of marbled white frozen tissue that is anaesthetic to touch. Treatment is by slow rewarming, often using body heat. Non-freezing cold injury often does not manifest until exposure to cold ceases, and it results in warm painful swollen extremities, usually the feet. Chilblains are a minor form of cold injury.

Conditions that preclude work in moderate to severe cold

- History of ischaemic heart disease
- Peripheral vascular disease
- Hypertension or Raynaud's phenomenon
- Asthma
- Metabolic disorders
- Sickle cell disease
- Arthritis

Ionising radiation

Ionising radiation displaces electrons from their normal orbits around the nucleus of the atom. The resulting ionisation alters the nature of biological molecules, especially DNA, resulting in gene mutation or cell death. α Small particles are relatively large and easily stopped. β Small particles are small and can penetrate up to a centimetre in tissue. Neutrons are smaller than α particles but are much more penetrating. γ Small radiation and *x* rays are packets of energy transmitted as electromagnetic radiation, and are highly penetrating.

External irradiation is that arising from a source—either a radiation generator, such as an *x* ray machine, or a radioactive substance—that is separate from the body. The irradiation ceases when the generator is switched off or the source is moved away or shielded. The body can be *contaminated* by particles of radioactive material that lie on the skin externally or are incorporated into the tissues, resulting in *internal* irradiation. The latter will persist as long as the radioactive material is in the body. Alpha emitters such as plutonium are particularly harmful sources of internal irradiation.

Large doses of ionising radiation cause death by damage to the brain, gut, and haemopoietic system. Such exposures only occur in the event of accidents or deliberate release in nuclear warfare. Lower doses can damage the skin or the lens of the eye. This may occur if sources are mishandled or exposures are prolonged—for example, in industrial radiography or interventional radiography. The *direct* effects of radiation are considered to have a dose threshold for their occurrence, and the severity of the effect is related to the dose received.

The *stochastic* effects of radiation (including the induction of cancer and hereditary effects) do not have a threshold, and the likelihood of the effect is related to the dose. Risk estimates for the stochastic effects of radiation have been derived from epidemiological studies (cancer) and animal studies (hereditary effects). The most important epidemiological data are from the Life Span Study of survivors of the atom bombs used in 1945. The risk estimates are published by a number of bodies of which the ICRP is the most influential. The ICRP also publishes dose limits derived from the risk estimates, and these are the basis of the statutory dose limits applied in many countries. Risk estimates and dose limits are regularly updated as the underlying science develops.

Workers who are substantially exposed to ionising radiation are subject to regular medical surveillance. This is to ensure that they are fit for their proposed work with radiation—for example, the need to work with unsealed sources or to use respiratory protective equipment. They are also subject to dose monitoring. Exposure to ionising radiation should be as low as reasonably practicable (ALARP) by the provision of appropriate controls, including shielding and reduction of exposure time. As legislative controls have been tightened, so the typical exposure to ionising radiation of workers has fallen. In the United Kingdom, average annual occupational doses are 1-2 millisieverts per year (about the same as background radiation).

Doses and units of radiation

- Absorbed dose—the energy of ionising radiation a body absorbs, measured in gray
- Dose equivalent—an adjustment of the absorbed dose, using a quality factor for the type of radiation involved, to take account of the effectiveness of the different types of radiations in harming biological systems; measured in sieverts
- Effective dose—an integrated index of the risk of harm, derived by multiplying the dose equivalent for each of the major tissues by a weighting factor based on the tissue's sensitivity to harm by radiation. The weighted values are summed. The unit is the sievert

The probabilities of harm from exposure to ionising radiation derived by the ICRP

Values are expressed as percentage risk per sievert dose received (the values in the table are multiplied by $10^{-2}\,Sv^{-1}$ to give the actual risk)

	Whole population	Working population
Fatal cancers	5	4
Hereditary disorders	1	0.6
Total risk	6	4.6

The ICRP recommends an effective dose limit of 20 mSv (averaged over a defined five year period) for workers, and 1 mSv per year for the public. Limits are also set for exposure of the eye lens, the skin, and the hands and feet. The dose limit for the fetus is the same as the public dose limit of 1 mSv a year

Studies of large cohorts of workers occupationally exposed to radiation consistently show a healthy worker effect. Nevertheless, cases of cancer of types known to be produced by ionising radiation do occur in these populations, sometimes with a slight excess. Individuals may be compensated for such disease on the basis of presumption of origin or probability of causation.

Electromagnetic fields

Electromagnetic fields with wavelengths shorter than 0.1 mm—that is, ultraviolet and below, contain insufficient energy to break molecular bonds and so do not result in ionisation. This "non-ionising radiation" does, however, have other frequency dependent effects on biological tissues. Broad divisions of this radiation include microwave and radio frequency radiation, as well as extremely low frequency, which includes the frequencies of power distribution.

At high frequencies—for example, microwaves used in communication systems—the main effect is tissue heating, a phenomenon made use of in the microwave oven. This effect is quantified by the specific absorption rate of energy into the body, and in most situations there are unlikely to be ill effects. This might not be the case where the individual is also working hard, or is exposed to a hot environment. At lower frequencies the effects of electric and magnetic fields are considered separately. Exposure to magnetic fields can set up circulating currents within the body, which have the potential to interfere with physiological processes if sufficiently great. For example, muscle activation could potentially occur during magnetic resonance imaging. Low frequency electric fields do not penetrate the body, but can generate charges on the body surface.

Other recognised but rarer effects include the phenomenon of microwave hearing. Some people hear repeated clicks when exposed to pulsed sources of electromagnetic fields, usually radars. A visual illusion of flickering lights (magnetophosphenes) can be produced when the retina is exposed to intense magnetic fields. Exposure standards, which reflect the frequency dependence of effects, have been derived to protect against the established effects of electromagnetic fields.

Since the late 1970s there has been increasing public concern about exposure to electromagnetic fields. This was prompted by epidemiological studies of the association between childhood cancer and residential exposure to magnetic fields. In 2001 the International Agency for Research on Cancer concluded that there was limited evidence that residential magnetic fields increase the risk of childhood leukaemia, resulting in a classification of "2B" "possibly carcinogenic" for extremely low frequency magnetic fields. It is thought that any risk relates to those exposed to fields at or above 0.4 microtesla, which are relatively large. The UK Childhood Cancer Study (UKCCS), the world's largest case control study on the causes of childhood cancer, found no evidence to support the association between residential magnetic field exposure and childhood leukaemia or other cancers. Any real effects must be very small in magnitude.

Public concern also extends to the possible effects of exposure to electromagnetic fields from mobile phone hand sets and base stations. In the United Kingdom an independent expert group was commissioned to study the evidence in relation to mobile phone technology. This group concluded that exposure to radio frequency radiation below the ICNIRP guidelines did not adversely affect population health, but in

Typical magnetic and electrical fields

Typical magnetic fields
- Natural fields—70 microtesla (static)
- Mains power—200 nanotesla (if not close to power lines), 20 microtesla (beneath power lines)
- Electric trains—50 microtesla
- Cathode ray tubes—700 nanotesla (alternating)

Typical electric fields
- Natural fields—200 V/m (static)
- Mains power—100 V/m (in homes), 10 kV/m (under large power lines)
- Electric trains—300 V/m
- Cathode ray tubes—10 V/m (alternating), 15 kV/m (static)

ICNIRP 1998 Exposure guidelines to time varying electric and magnetic fields

- These specify basic restrictions in terms of current density for the head and trunk, whole body and localised specific absorption rates, and power density
- Reference levels below which the basic restrictions are unlikely to be exceeded are specified in terms of electric field strength (E), magnetic field strength (H), magnetic flux density (B), and power density (S). These are given separately for occupational exposure and for the general public, with lower values for the latter. Reference levels are also given for contact currents from conductive objects and for induced current in any limb

Exposure from mobile phones and base stations

- Public exposures from base stations are low; typical power densities have been measured as 1 mW/m^2, with maximum power densities of 10 mW/m^2
- For comparison, the ICNIRP public exposure guidelines are a power density of 4.5 W/m^2 at 900 MHz and 9 W/m^2 at 1.8 GHz
- Power densities can exceed guidelines very close to the antenna, and for this reason public access to these antennae has to be controlled
- Hand sets can generate power densities of up to 200 W/m^2, but the resulting fields inside the body are appreciably less then those measured externally

Units for electromagnetic fields

- Electric field strength (E)—volts per metre
- Magnetic field strength (H)—amps per metre
- Power density (S) (vector product of E and H)—watts per square metre
- Magnetic flux density (B)—Tesla (1 Tesla is about equal to 10 000 Gauss)

view of other biological evidence it concluded that it was not possible to say that exposures below current guidelines were totally without potential adverse health effects. The group therefore advocated a precautionary approach in the use of this technology—for example,, suggesting that the use of mobile phones by children for non-essential calls should be discouraged.

There is no evidence that exposure to electromagnetic fields from the use of display screen equipment has any harmful effects.

Optical radiations

Optical radiation comprises ultraviolet, visible, and infrared radiation, which have wavelengths between 100 nm and 1 mm. Their harmful effects are largely restricted to the skin and the eye. Ultraviolet radiation is implicated in non-melanoma and melanoma cancers. Outdoor workers—for example, farmers and the deck crews of ships—have an increased risk of non-melanoma cancer. Fortunately this is usually curable. As a sensible precaution, all those who work outdoors should avoid overexposure of the bare skin to sunlight and sunburn in order to reduce their risk of melanoma cancer. Some evidence suggests that exposure to ultraviolet radiation can impair the function of the immune system.

Ultraviolet radiation is responsible for the painful symptoms of arc eye (photokeratoconjunctivitis), which occurs some hours after exposure to a bright source of ultraviolet radiation such as a welding arc. Often, bystanders who are adventitiously exposed get this condition.

Infrared radiation can cause thermal damage to the skin and eyes, both of which are easily protected, the latter with appropriate goggles. In developed countries occupational cataract from exposure to infrared radiation is largely of historical interest, given proper protection. In developing countries, however, cataracts may occur as a result of overexposure to infrared radiation, possibly exacerbated by episodes of dehydration.

Sources of optical radiation where the light waves are in phase (for example, from lasers) can cause serious thermal damage to the retina, and skin burns. Engineering and administrative controls and personal protection are needed to prevent damage where high powered lasers are in use. Routine eye examination is not appropriate for laser workers, although a baseline assessment of visual acuity is useful to identify the functionally monocular individual, for whom a greater duty of care exists.

If unusual skin symptoms are reported in workers exposed to optical radiation the possibility of photosensitisation should be considered, as can occur with exposure to plant products—for example, psoralens released in parsley cutting. Photosensitisation can also occur from certain drugs. If workers complain of "sunburn" from working in the vicinity of ultraviolet sources such as insect killing lamps, it is important to check that the bulbs have the correct frequency spectrum.

Altered ambient pressure

Compressed air is used in civil engineering to stabilise the ground and to remove water from workings. Alternative methods of doing so are available, and should always be considered before opting to use compressed air. The effects of hyperbaric exposure in diving and compressed air work are different. Surface diving usually entails short exposures to high pressures, whereas compressed air work generally entails

Possible effects of optical radiation on the eye

- Ultraviolet C/B—arc eye
- Ultraviolet B—pigmentation of lens
- Ultraviolet A—retinal damage in aphakia
- Visible—accelerated ageing (high power sources), burns of retina (lasers)
- Infrared—corneal burns, usually prevented by blink reflex, cataract, retinal burns, from infrared A sources including lasers

Wavelengths of optical radiation

- Ultraviolet C (UVC)—100-280 nm
- Ultraviolet B (UVB)—280-315 nm
- Ultraviolet A (UVA)—315-400 nm
- Visible—400-760 nm
- Infrared—760 nm – 1 mm

The most potent sources of optical radiation are those in which the light waves are coherent or in phase, typically coming from laser sources

Working at pressure

- Atmospheric pressure is 14.7 psi
- 1 atmosphere, 1 bar, 10 m (or 33 feet) of sea water, are broadly equivalent pressures
- Absolute pressure is that of the working environment added to atmospheric pressure
- Decompression illness is very rare at pressures below 1.7 bar absolute. There is no risk from slight elevations of pressure such as in clean rooms
- Typical pressures experienced in civil engineering works are in the range 2-3.5 bar absolute
- Saturation diving techniques become necessary at depths below 50 m, 6 bar absolute

prolonged exposures at relatively low pressures. In diving, the physical effort required for the task may be limited, often using only the arms, whereas heavy manual work may be undertaken in compressed air work.

One effect that differs little in either situation is barotrauma—damage to an air containing organ by pressure exerted across a structure, typically in the ear or respiratory tract. Individuals exposed to raised pressures must be able to equalise such pressures—for example, by steady exhalation, during ascent from diving. The risk of barotrauma is minimised by excluding individuals with upper respiratory tract infections and by careful control of the rate of change of pressure during compression and decompression.

Decompression illness and osteonecrosis

More serious health effects are decompression illness and osteonecrosis. Under pressure, inert gas (principally nitrogen) dissolves in the tissues. When the pressure is reduced, this gas will come out of solution and form bubbles, in much the same way that bubbles form when pressure on carbonated drinks is released. These bubbles in turn cause effects which, if they are in the circulation or central nervous system, can be life threatening.

Decompression illness occurs in two types: pain only (previously type 1), in which symptoms occur in the skin (niggles) or around joints (bends), and serious (previously type 2), in which symptoms can occur in the circulation or nervous system. Symptoms can arise from gas bubbles in the pulmonary or coronary circulations (for example, the chokes), or from damage to the brain or spinal cord (for example, the staggers). Serious decompression illness can be life threatening.

To reduce the potential for bubble formation during decompression, pressure is reduced in a controlled, staged manner, the details of which depend on the duration and pressure of the preceding hyperbaric exposure. At its simplest this can be achieved by a series of timed stops at specified depths during ascent to the surface.

Decompression regimens inevitably entail a compromise between the long times needed for nitrogen to evolve from the tissues and the practical constraints arising from keeping a group of workers (in the case of civil engineering work) in the decompression chamber for long periods. The decompression chamber is an airlock between the working chamber and the external environment. Workers remain seated, resting, while the ambient pressure is reduced in a controlled fashion over one or more hours. Breathing oxygen during decompression helps to remove nitrogen from the body and shortens decompression times. As exposures increase in terms of both depth and time, longer decompression periods are required. At some of the higher pressures encountered in diving, the only practical approach is to adopt saturation methods, where individuals live and work under pressure for long periods, avoiding the need to decompress between working exposures.

With careful control of decompression and oxygen breathing, the incidence of decompression illness in offshore diving work has been kept very low. Further advances are needed in civil engineering work, where oxygen decompression is not yet always routine.

When decompression illness occurs it should always be treated by therapeutic recompression, as such events increase the risk of osteonecrosis. This serious complication of hyperbaric work results from compromise of the blood flow within bone structures. A section of normal bone dies and is replaced by softer material. If this occurs below the surface of a joint, such as the hip joint, there is a real risk of the joint surface collapsing, resulting in permanent disability.

Deep sea diver

Decompression chamber

Guidance on exposures, and international standards

- ICNIRP. Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). *Health Phys* 1998;74:494-522
- ICRP. 1990 *Recommendations of the international commission on radiological protection.* Annals of the ICRP 21,1-3. Oxford: Pergamon Press, 1991
- International Standards Organisation. *Hot environments—estimation of the heat stress on working man, based on the WBGT index (wet bulb globe temperature).* Geneva: ISO, 1989 (ISO 7243)
- International Standards Organisation. *Moderate thermal environments—determination of the PMV and PPD indices and specification of the conditions for thermal comfort.* Geneva: ISO, 1993 (ISO 7730)
- International Standards Organisation. *Ergonomics of the thermal environment—Medical supervision of individuals exposed to extreme hot or cold thermal environments.* Geneva: ISO, 2001 (ISO 12894)

Risk factors for osteonecrosis are not clearly established. It can occur after one "bad" decompression but is normally seen only after higher pressure exposures. Risk factors in compressed air work include the number of hyperbaric exposures and the number of episodes of decompression illness.

Barotrauma and decompression illness may occur in aviation environments. They are most likely to occur if an aircraft pressurisation system fails at an altitude above 20 000 feet, after a high altitude ejection, or after flight at altitude in an unpressurised aircraft. The risks can be minimised by breathing 100% oxygen (denitrogenation) before flights or training exposures in an altitude chamber carrying a risk of decompression illness. Osteonecrosis after decompression in aviation is exceedingly rare.

Living and working at altitude

Living and working at altitude carries different risks—namely, acute mountain sickness, high altitude pulmonary oedema (HAPE), and high altitude cerebral oedema (HACE). Symptoms of acute mountain sickness can occur at altitudes of 2500 m, with the prevalence reaching 40% at altitudes over 4000 m. The symptoms include headache, nausea and vomiting, sleep disturbance, and muscle weakness, and are thought to arise from a mild oedema of the lungs, the splanchnic circulation, and the brain. The condition is treated by descent to a lower altitude. Breathing oxygen, and taking acetazolamide and dexamethasone can also help. The main preventive measure is to limit the rate of ascent to altitude. Unlike acute mountain sickness, both high altitude pulmonary oedema and high altitude cerebral oedema are life threatening. The former is treated by descent and the use of oxygen.

People who live at high altitude show physiological adaptations to their environment, although even these may fail with time. Chronic mountain sickness (Monge's disease) is a loss of tolerance to hypoxia, which occurs particularly in middle aged men. It results in an erythropoiesis, with the haematocrit rising as high as 80%. Clinical effects include cyanosis, dyspnoea, cough, palpitations, and headache. The condition can only be alleviated by moving to a lower altitude.

Acceleration

Exposure to sustained acceleration is experienced on fairground rides (2-3 G) or in flight, and then only significantly in aerobatic or military flying. Radial acceleration occurs during banked turns. When the head is to the inside of the turn the acceleration is positive in the "z" axis. With the head on the outside of the turn the acceleration is negative in the same axis. Positive G increases the hydrostatic weight of the column of blood above the heart, reducing arterial pressure and perfusion of the retina and the brain. Negative G has the opposite effect, increasing arterial pressure and resulting in engorgement of the head and neck

Protection from positive G is provided by posture, keeping the body nearer the horizontal plane than the vertical, by lifting the legs up and lowering the backrest. Valsalva type manoeuvres are used slightly in anticipation of acceleration to increase the pressure in the arterial system, and protective anti-G suits are routinely worn by military pilots. These prevent pooling of blood in the peripheries and limit the descent of the heart and diaphragm under acceleration

Further reading

- Ashcroft F. *Life at the extremes.* London: Flamingo, 2001. *A journalistic account by a professor of physiology of the science of survival, including chapters on altitude, diving, heat, and cold*
- Case RM, Waterhouse JM. *Human physiology: age, stress and the environment.* Oxford: Oxford University Press, 1994. *An undergraduate textbook with a series of short chapters on topics including the thermal environment, altitude, diving, and acceleration. Useful academic introduction to the areas covered*
- Edholm OG, Weiner JS. *The principles and practice of human physiology.* London: Academic Press, 1981. *A bit dated, but still a valuable reference on the physiology of diving, altitude, the thermal environment, and other topics. Covers the basics in much more detail than Case and Waterhouse*
- Bennett PB, Elliott DH. *The physiology and medicine of diving,* 4th ed. London: WB Saunders, 1993. *A comprehensive textbook, which includes a chapter on compressed air work. A standard reference covering all aspects of hyperbaric exposures including clinical hyperbaric oxygen therapy*
- Cummin AR, Nicholson AN. *Aviation medicine and the airline passenger.* London: Arnold, 2002. *A multiauthored text considering the aeromedical implications of a range of common medical conditions*
- Ernsting J, Nicholson AN, Rainford DJ. *Aviation medicine,* 3rd ed. London: Butterworths, 2000. *A comprehensive text covering all aspects of aviation physiology, psychology, and clinical aviation medicine; suitable for students of specialised aviation medicine diplomas*
- Harding RM, Mills FJ. *Aviation medicine,* 3rd ed. London: BMJ Publishing Group, 1993. *An introductory text for the general reader which gives a good overview of the main topics relevant to clinical practice*
- Mettler FA, Upton AC. *Medical effects of ionising radiation,* 2nd ed. Philadelphia: WB Saunders, 1995. *A comprehensive and well referenced review of the science underlying the medical effects of ionising radiation. Covers direct effects and carcinogenesis at length*
- National Radiation Protection Board. *Living with radiation.* London: NRPB and HMSO, 1998. *A book written for the lay reader which sets out a good introduction to the science and social context of exposures to both ionising and non-ionising radiations*
- Parsons K. *Human thermal environments,* 2nd ed. London: Taylor and Francis, 2002. *A standard text on responses to hot, moderate, and cold thermal environments, presented as an integrated approach incorporating physiology, psychology, and environmental physics*
- Report of the Advisory Group on Non-ionising Radiation. *ELF Electromagnetic fields and the risk of cancer.* London: NRPB 2001;Doc12:3-179. *Scientific report covering exposures to electromagnetic fields, studies on cancer induction, epidemiological studies, and occupational exposures. Includes recommendations for further research*
- Stewart W. *Mobile phones and health.* Chilton Independent Expert Group on Mobile Phones, 2000. *Report of a Government appointed review group with good coverage of mobile phone technology and the scientific evidence for health effects. Makes numerous recommendations for action*
- Ward MP, Milledge JS, West JB. *High altitude medicine and physiology,* 3rd ed. London: Arnold, 2000. *A comprehensive review covering history, physiology, biochemistry, and the clinical effects of altitude and cold*
- Barry PW, Pollard AJ. Altitude illness. *BMJ* 2003;326:915–9. *A well-referenced up-to-date clinical review*

Effects of positive headwards acceleration

- 3-4 G—darkening of visual fields
- 3.5-4.5 G—loss of peripheral vision
- 5-6 G—loss of consciousness

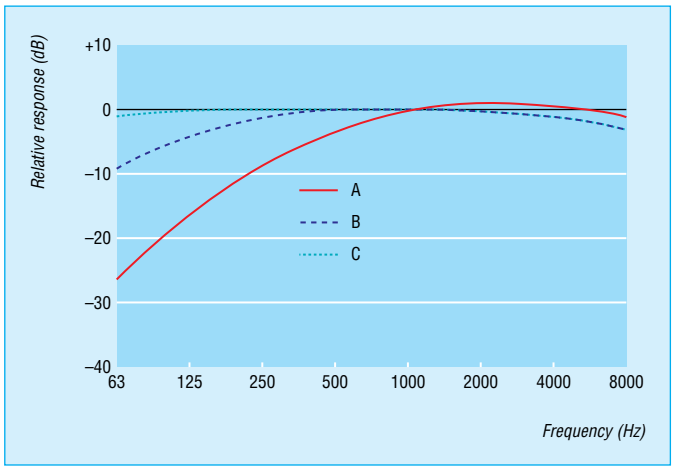
If the rate of onset of acceleration is high, loss of consciousness will be the first symptom

12 Noise and vibration

Paul Litchfield

Sound is generated when a vibrating source transmits energy to the surrounding air, creating small changes in pressure. If the frequency of the sound produced lies between about 20 and 16 000 Hz it may be perceived by the hearing mechanism and is classed as being "audible." Sound levels are measured in decibels (dB), a logarithmic unit in which the faintest sound detectable by the human ear is set at 0 and the level doubles for every 3 dB. In assessing audible sound it is conventional to use a weighted scale that filters the actual pressure level in specified octave bands by an agreed amount to resemble the response of the ear over those frequencies. The most commonly used weighting is the "A" network, and resultant units are expressed as dB(A). Noise is simply unwanted sound.

The body is also susceptible to non-acoustic vibration transmitted by direct contact with oscillating surfaces. As with sound, frequency is important: vibration below 2 Hz and above 1500 Hz is not thought to be harmful; motion between 5 Hz and 20 Hz is considered potentially most damaging. Vibration can be measured in various ways, but is normally expressed as acceleration in metres per second squared (m/s^2) averaged over the three axes. As vibration at frequencies below 2 Hz and above 1500 Hz is not thought to cause damage, weighting is applied to measurements of vibration magnitude to allow for this frequency dependence of the risk of harm.

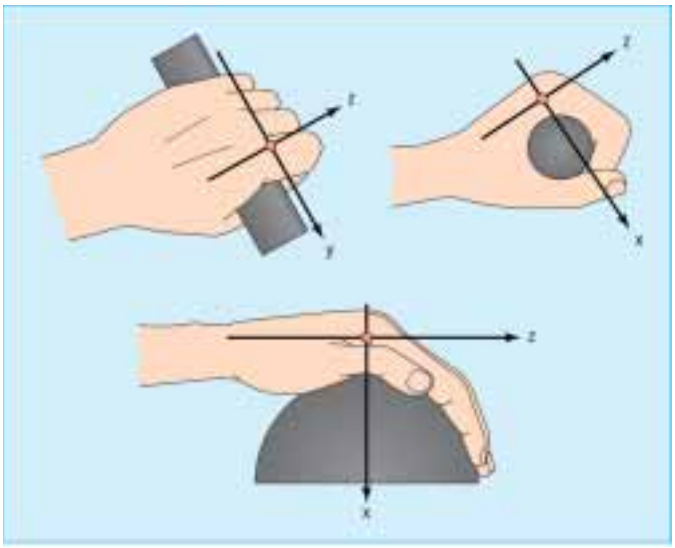


The human ear is more sensitive to certain frequencies, and in order to approximate the response of the ear it is possible to suppress certain frequencies and boost others in the electronic circuitry of sound level meters. This technique is known as "weighting," and the most commonly quoted weighting network is the A weighting

Health effects of noise

The principal hazard from noise is impairment of hearing. This may be confined to a reversible alteration in hearing levels, known as temporary threshold shift, which resolves spontaneously in the quiet. It may last from a few minutes to months depending on the noise level encountered. If exposure to high noise levels is sustained for a prolonged period a permanent shift can occur, termed noise induced hearing loss. Short bursts of very high intensity sound (such as an explosion or gunfire), known as impulse noise, can also cause additional harm to the ear by rupturing the tympanic membrane or even disrupting the ossicles.

There has been considerable interest in recent years in the non-auditory effects of noise. Comprehensive literature reviews



Vibration is usually measured in three orthogonal directions at the interfaces between the body and the vibrating surface

Non-auditory health and physiological effects of noise

- *Cardiovascular effects*: in laboratory studies, noise has been shown to produce increases in diastolic blood pressure. However, there is no clear evidence that long term exposure to noise is a risk factor for hypertension
- Some studies suggested an association between noise exposure or noise annoyance and the frequency of *psychiatric symptoms* but these findings have been questioned in later studies. There is some evidence that noise sensitivity is an indicator of vulnerability to minor psychiatric disorder, and that annoyance responses are stronger among individuals with psychiatric disorders
- The effect of noise on *performance* is complex. Some research found no clear evidence of effects at noise levels below 95 dB, whereas other research suggests that performance may be affected at much lower levels
- *Fatigue, headaches, and irritability* have been found to be over-represented in groups exposed to noise, but methodological flaws in study design have made valid conclusions difficult

Range of instruments for measuring noise and vibration levels

have been published, but much of the evidence remains weak or equivocal.

Noise induced hearing loss

Noise induced hearing loss is caused by damage to the cilia on the basilar membrane in the organ of Corti in the inner ear. This damage is progressive and irreversible and results in loss of both absolute sensitivity of the ear and in frequency selectivity. Characteristically, loss is initially predominant in the higher frequencies (3-6 kHz) and classically, a depression at 4 kHz may be seen on audiometry. With continuing exposure hearing loss extends to both higher and lower frequencies and is frequently superimposed on the effects of age related hearing loss, also known as presbyacusis.

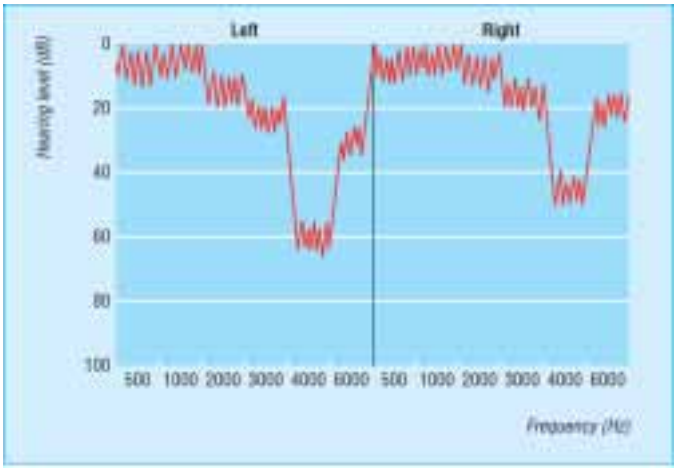
The development of noise induced hearing loss is insidious, and deafness may be considerable by the time an individual seeks assistance. Initially those affected describe difficulty in hearing conversations against a noisy background. Because consonants have a higher frequency than vowels, they are more difficult for a person with noise deafness to recognise, with resultant degradation of discrimination of speech. Hearing loss is frequently associated with tinnitus, which may be more disabling than deafness. On examination, the tympanic membranes usually seem normal, but testing with a tuning fork shows a sensorineural deafness. Industrial noise usually gives rise to bilateral hearing loss but specific activities, such as use of firearms, may produce unilateral deafness depending on the location of the noise source in relation to the ears. Audiometry shows a hearing loss that is predominantly high frequency, although in severe cases lower frequencies are affected. This latter case produces far greater disability because of the impact on the speech range (0.5-2.0 kHz).

Noise induced hearing loss is common. Data from the UK National Household Survey indicate that in excess of 130 000 people have hearing problems arising from noise at work, and the Occupational Safety and Health Administration estimates that there are 10 million people with similar hearing problems in the United States. Manufacturing industry has been the source of most cases in the past, but noise levels can be high for those working in many other sectors including construction, transport, and the armed forces. More recently, concern has been raised in relation to call centre operatives, but any potential problems seem to relate to extraneous noises received through headsets (acoustic shock) rather than to ambient noise levels.

Risk management

Noise induced hearing loss is a preventable condition and, as with any hazard, the first step is to assess the risk. As a general guide, noise levels are likely to be hazardous if communication without shouting is difficult at a distance of two metres. If there is reason to believe that there may be a problem then noise levels should be measured by a competent person. The risk of developing noise induced hearing loss is a function of both the level of noise exposure and its duration. Noise levels are therefore often expressed as daily personal noise exposure (L EP,d), which averages the dose over an eight hour working day. L EP,d action levels of 85 dB(A) and 90 dB(A) have been set in both the United States and European Union, above which certain control measures are mandatory. However, at the time of writing, negotiations are far advanced in Europe for a new Noise Directive, which will replace the existing directive (86/188/EC, implemented in the United Kingdom by the Noise at Work Regulations 1989) with tougher legislation that will reduce the action levels to 80 dB(A) and 85 dB(A), and introduce a limit value on exposure of 87 dB(A).

The best means of hazard control is elimination, and machinery noise can often be reduced substantially by better



Audiogram showing noise induced hearing loss with classical depression at 4 kHz

Noise induced hearing loss in the United Kingdom (adapted from Health and Safety Executive statistics 2000-1)

- UK Health and Safety Executive statistics are obtained from a variety of sources, including the occupational physicians reporting activities (OPRA), occupational surveillance scheme for audiologists (OSSA), and industrial injuries scheme (prescribed diseases)
- The industry groups with the highest annual average incidence rates of new cases qualifying for benefit were extraction, energy, and water supply (7.9 cases per 100 000 employees), manufacturing (3.9), and construction (2.3) (based on 1999 and 2000 data). Of cases qualifying for benefit, 11% were in shipbuilding, repair, or breaking, and 9% were in the coal mining industry. Of new cases qualifying for benefit in 2000, 52% were in the occupational group of metal machinery and related trades workers
- Noise induced hearing loss is not reportable under the Reporting of Injuries, Diseases and Dangerous Occurrence Regulations 1995 (RIDDOR)

	Number of cases (OSSA/OPRA, estimated for 2000)
Sensorineural hearing loss	627
Tinnitus	161
Balance problems	5
Tympanic disorder	3
Other problems	1
Total	797 (648 individuals)
Prescribed diseases*	226

*To qualify for benefit, there must be at least 50 dB of hearing loss. The degree of disability is calculated from the hearing loss in such a way that 50 dB in both ears equates to 20% disability. Under current guidelines, a worker must have been employed for at least 10 years in specified noisy occupations. Of the almost 2000 disallowed claims in 1998, 800 claimants had 35-49 dB hearing loss

Differential diagnosis of noise induced hearing loss

- Conductive—Wax, acute otitis media, chronic otitis media, otosclerosis, tympanic membrane injury, barotrauma, ossicular dislocation
- Sensorineural—Presbyacusis, congenital (maternal rubella, hereditary, perinatal anoxia), infective (measles, mumps, meningitis), vascular (haemorrhage, spasm or thrombosis of cochlear vessels), traumatic (head injury), toxic (streptomycin, neomycin, carbon monoxide, carbon disulphide), Meniere's disease, late otosclerosis, acoustic nerve tumours (usually unilateral)

Main requirements of the UK Noise at Work Regulations 1989

Action required where L EP,d* is likely to be:	<85 dB(A)	85 dB(A) First action level	90 dB(A) Second action level†
Empolyers' duties			
General duty to reduce risk			
Risk of hearing damage to be reduced to the lowest level reasonably practicable*	√	√	√
Assessment of noise exposure			
• Noise assessments to be made by a competent person		√	√
• Record of assessments to be kept until a new one is made		√	√
Noise reduction			
Reduce exposure to noise as far as is reasonably practicable by means other than ear protectors			√
Provision of information to workers			
• Provide adequate information, instruction, and training about risks to hearing, what employees should do to minimise risk, how they can obtain ear protectors (if they are exposed to an L EP,d between 85 and 90 dB(A)), and their obligations under the Regulations		√	√
• Mark ear protection zones with safety signs, so far as reasonably practicable			√
Ear protectors			
Ensure so far as is practicable that protectors are:			
• Provided to employees exposed to an L EP,d of 85 dB(A) or above and less than 90 dB(A), who ask for them		√	
• Provided to all exposed above the second action level			√
• Maintained and repaired		√	√
• Properly used by all exposed			√
Ensure so far as reasonably practicable that all who go into a marked ear protection zone use ear protectors			√‡
Maintenance and use of noise control equipment			
Ensure so far as is practicable that:			
• All equipment provided under the Regulations is used, except for the ear protectors provided between first action level and second action level	√	√	√
Ensure all equipment is maintained	√	√	√
Employees' duties			
Use of equipment; so far as is practicable:			
• Use ear protectors			√
• Use any other protective equipment	√	√	√
• Report any defects discovered to employer	√	√	√
Machine makers' and suppliers' duties			
Provision of information			
Provide information on the noise likely to be generated	In theory if equipment provided to comply with*	√	√

*The dB(A) action levels are values of daily personal noise exposure L EP,d.
†All the actions indicated at 90 dB(A) are also required where the peak sound pressure is at or above 200 pascals.
‡This requirement applies to all who enter the zones, even if they do not stay long enough to receive an exposure of 90 dB(A) L EP,d.

design and maintenance. Damping and enclosure of vibrating machinery can greatly reduce exposure, or people can be provided with well insulated noise refuges in otherwise noisy environments. As a last resort, people can be issued with hearing protection: ear muffs (which completely cover the ear), ear plugs (which are inserted into the auditory canal), or semi-inserts (which cover the entrance to the ear canal).

It is important to ensure not only that any ear protection offered provides adequate noise attenuation but also that it does not interfere with any other protective equipment required, and that those using it understand that even short periods of non-use will greatly reduce the protective value.

Noise hazard sign to indicate that use of hearing protection is mandatory and standard design of ear muffs

Health surveillance

Health surveillance (including audiometry), although not a legal requirement, can provide a useful adjunct to risk management and is considered good practice where the second action level (see table) is exceeded. Hearing conservation

programmes will normally include a structured interview to gather relevant health data. This should cover relevant medical history based on the differential diagnosis for noise induced hearing loss, and a history of previous noise exposure such as previous employment in noisy industries, service in the armed forces, and leisure pursuits such as shooting or regular clubbing. The ear canal and tympanic membrane should be examined. Personal protective equipment should be inspected, and workers reminded of its correct use. Audiometric testing should be undertaken in a soundproof booth, and the screening results should be fully discussed, with onward referral if required.

Such programmes aim to identify at an early stage individuals particularly susceptible to noise damage, and to reinforce hazard information together with the use of control measures. The UK Health and Safety Executive has produced comprehensive guidelines on the conduct of audiometric testing programmes, including a helpful categorisation scheme that provides a template for the management of individuals according to the degree of hearing loss identified. The five categories within the scheme and the suggested action for each, and a chart of age related hearing loss at low and high frequencies are given in the two tables.

Classification of audiograms into warning and referral levels

	Sum of hearing levels			
	0.5, 1, 2 kHz		3, 4, 6 kHz	
Age in years	Warning level	Referral level	Warning level	Referral level
20-24	45	60	45	75
25-29	45	66	45	87
30-34	45	72	45	99
35-39	48	78	54	111
40-44	51	84	60	123
45-49	54	90	66	135
50-54	57	90	75	144
55-59	60	90	87	144
60-64	65	90	100	144
65	70	90	115	144

The Health and Safety Executive categorisation scheme

Category	Symptom	Suggested action
1	Rapid change in hearing threshold has occurred (that is, a change in the sum of the hearing levels for either the low or high frequencies of 30 dB, compared with the previous audiogram, or 45 dB if the period between the tests is more than three years). This change may be due to noise exposure or disease	Referral
2	This is usually related to medical factors. Unilateral hearing loss is not normally noise induced and may indicate auditory nerve disease. Unilateral hearing loss is considered to exist if the difference in the sums of the hearing levels between the two ears exceeds 45 dB for the low frequencies, or 60 dB for the high frequencies	Referral
3	Results show a pattern that could suggest significant noise inducing hearing loss (that is, where the sum of either the low or high frequencies, or both, in either ear, exceeds the value given for the appropriate age band)	Referral
4	Hearing has deteriorated beyond the level that might be accounted for by age alone, but not to the extent that medical referral is required	Warning. Formally notify the employee of the presence of hearing damage. Employee to understand that they have suffered some hearing loss; it is essential that they comply with the employer's hearing conservation measures. Assess rate of progression of hearing loss
5	Within normal limits	None, but assess rate of progression of hearing loss

Health effects of vibration

Vibration and noise often emanate from the same source. Vibration may reach the body through a number of pathways, but consideration of adverse health effects centres on whole body vibration and hand arm vibration. As with noise, the risk of harm is a function of both the magnitude of exposure and of its duration: "doses" are therefore adjusted to a standard reference period of eight hours to allow comparison, and this figure is termed A(8). Measuring vibration is complex and should only be undertaken by those with specialist training.

Use of a vibrating tool for road breaking

Whole body vibration

Interest in the effects of whole body vibration stems from the middle of the 20th century when mechanisation, particularly of transport, became more prevalent. Vibration is transmitted either from a machine platform through the feet, or from a

seat through the buttocks. Exposure is most likely to occur with vehicle use and this includes road, off road, rail, air, and maritime use: it is estimated that as many as 9 million people in the United Kingdom are regularly exposed to whole body vibration. The disorders reported in groups exposed in this way include gastric problems, vestibular dysfunction, circulatory changes, menstrual disturbance, and psychological effects. However, the main problem associated with whole body vibration is back pain, and the UK Health and Safety Executive estimates that up to 21 000 cases may be caused by exposure, with a further 13 500-31 500 cases of exacerbation of a pre-existing condition. The evidence base for a causal link between whole body vibration and back pain nevertheless remains weak, and has recently been comprehensively reviewed.

Hand arm vibration

Vibration may be transmitted to the hands and arms by the use of hand held power tools, hand guided machinery, or by holding materials being processed by machines. Exposure is particularly common in agriculture, construction (particularly scabbling), mining, engineering, forestry, public utilities, and shipbuilding. It is estimated that about 1 million people in the United Kingdom are exposed to potentially harmful levels of hand arm vibration in their work, and as many as 300 000 may have developed adverse health effects as a result.

The health effects of exposure to hand arm vibration have been recognised for many years and have been ascribed a variety of labels. There is now general consensus on the use of the term "hand arm vibration syndrome" to describe the vascular (sometimes also known as vibration white finger), neurological, and musculoskeletal symptoms that can result. Acute vibration exposure causes vasoconstriction of the blood vessels supplying the fingers and, if prolonged, it may damage the endothelium and stimulate smooth muscle proliferation so that the lumen of the vessels gradually narrows. Damage also occurs to the peripheral nerves, with acute oedema and chronic demyelination. Muscular weakness in the hand is common, carpal tunnel syndrome is recognised in some cases, and there is evidence to indicate that premature osteoarthrosis of the wrist and elbow may occur. The precise relation between these elements of the syndrome remains a matter for debate, but there is no doubt that the vascular and neurological components can occur separately.

In the early stages of vibration injury the only symptom may be a tingling in the fingers, most noticeable at the end of the working day. This may be associated with a loss of sensation and periodic blanching of the tips of the fingers when exposed to cold. As the condition progresses the blanching extends to the root of the fingers, although the thumbs are rarely affected. In more severe cases there is considerable pain, with a loss of grip strength and dexterity, and attacks may occur even in warm surroundings. Rarely the condition can progress to the extent that circulation is permanently impaired and the fingers become cyanosed—exceptionally, cases of vibration induced gangrene have been reported.

Risk management

Assessment of risk is based on the type of vibrating equipment employed and its pattern of use. In the United Kingdom the action level for introducing preventative measures is if exposure regularly exceeds an A(8) of 2.8 m/s^2 (dominant axis). It is important to recognise that this is not a "safe" level: some individuals are likely to develop hand arm vibrations with prolonged use even if this threshold is not exceeded. A new European Vibration Directive has recently been adapted (to be transferred into UK law in 2005), which sets a limit value

Vibration induced disorders in the United Kingdom

- A UK survey on behalf of the Health and Safety Executive gave an estimate for the national prevalence estimate of vibration white finger (VWF) of 288 000
- The industry with the highest annual average rate of new assessments of disability at 1% in 1999-2000 was extraction, energy, and water supply, because of the relatively high number of claims made by current or former coal miners. Of the new assessments made in other industries, 3% were in shipbuilding, repair, or breaking; 5% were in other manufacturing industry; and 4% in construction
- In 1999-2000, coal mining accounted for 46% of cases for carpal tunnel syndrome, construction for 12%, and shipbuilding, repair, or breaking for 4%

	No of cases*
Raynaud's phenomenon or hand arm vibration or vibration white finger	935
RIDDOR† (2000-1 provisional)	
Carpal tunnel syndrome	119
Hand arm vibration	905
Prescribed diseases (1999-2000)	
Vibration white finger	3212
Carpal tunnel syndrome	475

*Musculoskeletal occupational surveillance scheme (MOSS), reporting by rheumatologists or occupational physicians reporting activities (OPRA), estimated for 2000.
†RIDDOR, Reporting of Injuries, Diseases and Dangerous Occurrence Regulations 1995 (adapted from Health and Safety Executive statistics 2000-1).

Differential diagnosis

Vascular conditions

- Connective tissue disease—scleroderma, mixed connective tissue disease, systemic lupus erythematosus, rheumatoid arthritis, dermatomyositis, polyarteritis nodosa, Sjogren's disease
- Traumatic—after injury or surgery, hand transmitted vibration, frostbite, thoracic outlet syndrome
- Arterial disease—thromboangitis obliterans, thromboembolism, arteriosclerosis
- Toxins and drugs—vinyl chloride, ergot, β blockers, clonidine
- Dysglobulinaemia—cryoglobulinaemia
- Neurogenic—poliomyelitis, syringomyelia, hemiplegia

Neurological conditions

- Peripheral nerve entrapment—carpal tunnel syndrome, ulnar nerve entrapment at elbow or wrist, thoracic outlet syndrome
- Central nervous system disorders—compression myelopathy (spondylosis or spinal cord tumor), subacute combined degeneration of the cord, multiple sclerosis
- Peripheral neuropathy—diabetic, alcoholic, toxic (for example, organophosphates, thallium, acrylamide, carbon disulphide, *n*-hexane, methyl butyl ketone, diethyl thiocarbamate, lead)
- Drug induced (for example, chloramphenicol, isoniazid, streptomycin, polymyxin, ethambutol, nitrofurantoin, metronidazole, gold, indomethacin, vincristine, perhexiline, phenytoin)

on exposure of 5 m/s^2 (sum of three axes) and an action value of 2.5 m/s^2 (sum of three axes).

Manufacturers of vibrating tools may be able to provide useful data on levels under standard conditions, but care must be taken because actual levels in field use can differ substantially from those generated in a controlled environment. Similarly, field measurements can vary widely depending on mode of use and the materials being worked. In practice it is therefore usual to institute a preventive programme wherever there is prolonged use of tools likely to be hazardous.

Prevention programmes aim to eliminate or substitute the hazardous process where possible. Where this is not possible, the procurement of low vibration machinery, fitting of vibration reducing adaptations (such as vibration reducing handles), regular maintenance and re-engineering of processes to avoid the need for prolonged tight gripping of high vibration parts will reduce exposure. Keeping the hands and body warm helps to maintain a good blood supply to the fingers and thereby reduces the risk of injury. Vibration reducing gloves are available but their efficacy is limited. A key element in a preventive programme is the provision of training and information about the hazard and the means of reducing risk.

Health surveillance

Health surveillance aims to identify those who develop early symptoms so that progression can be avoided and it is appropriate if exposure levels are likely to trigger a prevention programme. Pre-employment screening is helpful in identifying individuals with conditions such as Raynaud's disease that are a contraindication to work with vibrating tools, in establishing baseline measurements, and in educating workers about measures to minimise risk—not least the avoidance of smoking. It is good practice to repeat the assessment for newly exposed workers to identify those who may be particularly susceptible. Thereafter, annual review is recommended, with any symptoms being reported to a designated person as soon as they occur.

Assessment should comprise a structured history and relevant clinical examination that will identify early hand arm vibration syndrome and assist with differential diagnosis, as a number of constitutional conditions give rise to similar symptoms. Guidelines from the UK Health and Safety Executive (see Further reading) give a sample questionnaire and guidance on tests that may be helpful for examination. Various methods of grading signs and symptoms have been devised and those of Taylor and Pelmear, and Griffin have been widely used. However, the most commonly used system of classification for hand arm vibration syndrome is currently the Stockholm Workshop scale, which grades the vascular and sensorineural components by severity. This scale, and the speed of progression along it, can helpfully be used to guide the management of affected workers. No effective treatment is available for this condition: management relies on adjustments to work, and limitation of vibration exposure. Cessation of vibration exposure may well compromise an individual's continuing employment, and great care is therefore required before making any such recommendation. A number of additional test measurements (detailed Lindsell CJ and Griffin MJ, 1988) can be carried out by specialist centres to help confirm the degree of incapacity, and referral should be considered in such circumstances.

The photographs showing the range of instruments for measuring noise and vibration and showing a vibrating tool for road breaking are courtesy of Castle Instruments. The figure showing how vibration is measured is adapted from HS(G)88.

Vibration induced gangrene

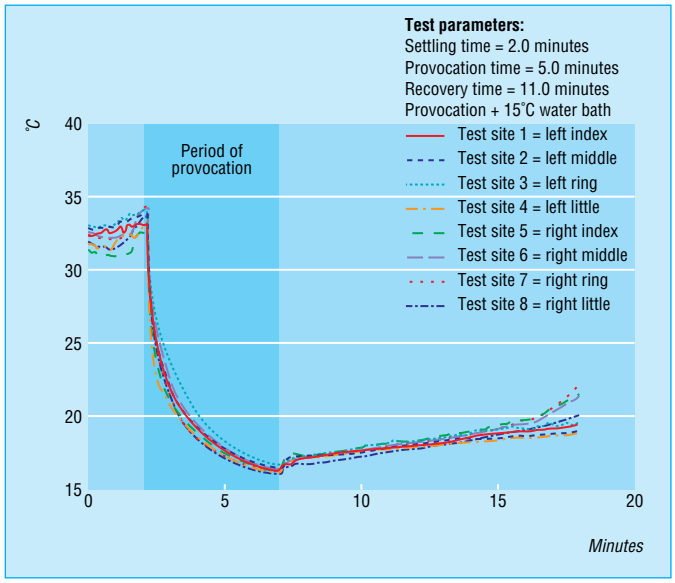
Stockholm workshop classification

Vascular component

Stage	Grade	Description
0		No attacks
1V	Mild	Occasional attacks affecting only the tips of one or more fingers
2V	Moderate	Occasional attacks affecting distal and middle (rarely also proximal) phalanges of one or more fingers
3V	Severe	Frequent attacks affecting all phalanges of most fingers
4V	Very severe	As in stage 3 with trophic changes in the fingertips

Sensorineural component

Stage	Description
0SN	Vibration-exposed but no symptoms
1SN	Intermittent numbness with or without tingling
2SN	Intermittent or persistent numbness, reduced sensory perception
3SN	Intermittent or persistent numbness, reduced tactile discrimination or manipulative dexterity or both



Results of cold provocation showing an abnormal response as found in vascular damage from hand arm vibration syndrome

Further reading

- Butler MP, Graveling RA, Pilkington A, Boyle AL. *Non-auditory effects of noise at work: A critical review of the literature post 1988. HSE Contract Research Report 241/1999.* Sudbury: HSE Books, 1999. *A review of the literature relating to the non-auditory effects of noise since 1988, updating earlier work by Smith and Broadbent, considering behavioural and psychological elements as well as physiology and health*
- Health and Safety Executive. *A guide to audiometric testing programmes. Guidance Note MS 26.* Sudbury: HSE Books, 1995. *Practical guidance on the conduct of occupational audiometry*
- Stayner RM. *Whole body vibration and shock: A literature review.* Sudbury: HSE Books, 2001. (HSE Contract Research Report 333/2001) *Review of the effects of whole body vibration, comparing the state of knowledge with noise induced hearing loss and hand transmitted vibration; concentrates on the relationship with low back pain*
- Health and Safety Executive. *Hand-arm vibration. HS(G)88.* Sudbury: HSE Books, 1994. *Practical guidance on the hazards, assessment methods and controls for hand transmitted vibration*
- Lindsell CJ, Griffin MJ. *Standardised diagnostic methods for assessing components of the hand-arm vibration syndrome.* Sudbury: HSE Books, 1988 (HSE Contract Research Report 197/1988) *Lindsell and Griffin define a standardised battery of tests for detecting the various components of hand arm vibration syndrome*
- Faculty of Occupational Medicine of the Royal College of Physicians of London. *Hand-transmitted vibration: clinical effects and pathophysiology.* London: Faculty of Occupational Medicine of the Royal College of Physicians of London, 1993. *Part one summarises the evidence relating to hand arm vibration syndrome and recommends assessment methodologies; part two outlines in some detail the evidence base for the report. Currently being revised; publication is planned for 2004*
- OHSA. *Noise and Hearing Conservation.* Occupational Safety and Health Administration. US Department of Labor. Revised 15 February 2002. http://www.osha-slc.gov/SLTC/noisehearingconservation/. *The OSHA site provides links to a wide range of US Government documents relating to noise and hearing conservation*
- Palmer KT, Coggon D, Griffith MJ, Haward BM. *Hand-transmitted vibration: occupational exposure and their health effects in Great Britain.* Sudbury: HSE Books, 1999 (HSE Contract Research Report 232/1999)

13 Respiratory diseases

Ira Madan

The pattern of occupational lung disease is changing in industrialised countries. A reduction in manufacturing industries and stricter health and safety legislation during the past 50 years have resulted in a sharp decline in the incidence of silicosis, asbestosis, and other pneumoconioses. Asthma is now the most common occupational respiratory disorder in these countries. By contrast, the traditional occupational lung diseases are commonly seen in developing countries, and occupational asthma is reported less often. However, the true prevalence of asthma attributable to occupation in these countries remains unknown.

Since 1989, the understanding of the epidemiology of occupational lung disease in the United Kingdom has been greatly enhanced by the Surveillance of Work related and Occupational Respiratory Disease (SWORD) and Occupational Physicians Reporting Activity (OPRA) projects. Occupational and respiratory physicians systematically report new cases of occupational lung diseases, together with the suspected agent, industry, and occupation. The projects have provided an estimate of the incidence and pattern of occupational lung disease in the United Kingdom.

Estimated number of cases of work related and occupational respiratory disease reported to SWORD/OPRA by diagnostic category, 1998-2000

Diagnostic category	1998	1999	2000
Benign pleural disease	625	1243	1080
Asthma	807	1129	797
Malignant mesothelioma	701	1018	964
Pneumoconiosis	225	320	292
Other diagnosis	187	239	218
Inhalation accidents	178	154	119
Bronchitis/emphysema	58	29	144
Lung cancer	112	81	126
Infectious disease	87	63	77
Allergic alveolitis	29	42	37
Total number of diagnoses	3009	4418	3854
Total number of individuals*	2934	4298	3787

*Individuals may have more than one diagnosis.

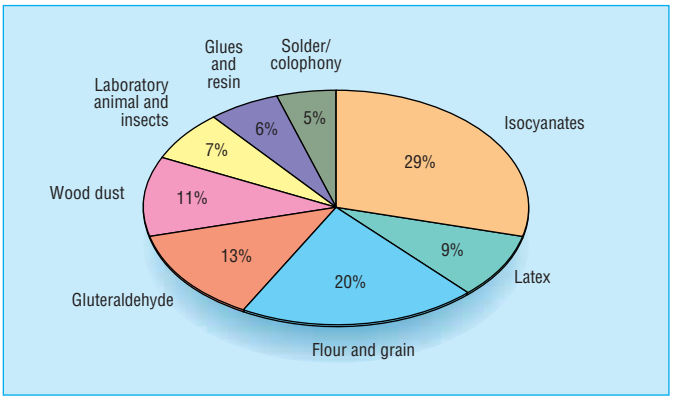
Occupational asthma

Occupational asthma is a disease characterised by variable airflow limitation and airway hyper-responsiveness caused by specific agents inhaled in the workplace. It does not include activation of pre-existing asthma or airway hyper-responsiveness induced by non-toxic irritants or physical stimuli such as cold air.

Two types of occupational asthma are recognised: immunological asthma appears after a latent period of occupational exposure; non-immunological occupational asthma develops without a period of latency and is associated with exposure to high concentrations of irritants. This latter type is referred to as reactive airways disease and is discussed separately. To date, more than 250 agents capable of causing immunological occupational asthma have been reported. In some jobs, such as hairdressing and farming, workers are exposed to many potential sensitisers and sensitisation may occur through interaction of several agents.

Substances that induce occupational asthma are classified as either high (>5 kDa) or low molecular weight allergens. High molecular weight substances are usually protein derived allergens such as natural rubber latex and flour. It is thought that some low molecular weight chemicals, such as diisocyanates, act as haptens and combine with a body protein to form a complete antigen.

Atopic individuals seem to be at increased risk of developing occupational asthma from some agents that induce specific immunoglobulin E (IgE)—for example, rat urinary proteins, and protease enzymes derived from *Bacillus subtilis* (detergent workers). However, atopic workers who are exposed to other agents—for example, isocyanates and plicatic acid (Western red cedar) seem to be at no more risk than non-atopic workers. Tobacco smokers are at greater risk of developing asthma after occupational exposure to several agents such as platinum salts, acid anhydride, and green coffee bean; the mechanism of this modifying effect is unknown.



Top eight suspected causative agents for occupational asthma cases reported to SWORD/OPRA 1998-2000

Farmers are at particular risk of developing occupational asthma because they are often exposed simultaneously to an array of potential sensitisers, such as animal derived allergens, arthropods, moulds, plants, and fungicides

Tobacco smoking and atopy are common among the working population. If these risk factors are found at pre-employment assessment the individual should not automatically be excluded from working with a respiratory sensitiser

Diagnosis

Between 5% and 10% of adult asthma is attributable to occupational factors. A detailed history of past and present occupational exposures is therefore essential in the assessment of a patient with adult onset asthma. Coughing at work or at the end of a shift is often the first symptom and precedes wheezing. Concurrent nasal congestion, lacrimation, and conjunctivitis may be associated with exposure to high molecular weight substances. The symptoms generally improve at weekends and holidays, but at advanced stages the respiratory symptoms may persist. Where possible, advice should be sought from the patient's employer's occupational health service, as they will have information on the substances that the employee is exposed to and will know if other workers have developed similar respiratory symptoms.

Investigations

Patients should record the best of three measurements of peak expiratory flow made every two hours from waking to sleeping over a period of one month (charts are available from Clement Clarke International Limited). Ideally, this period should include one or two weeks away from work. A drop in peak expiratory flow or substantial diurnal variability on working days but not on days away from work supports a diagnosis of occupational asthma. If there is any doubt, the patient should be referred to a specialist centre for further investigation.

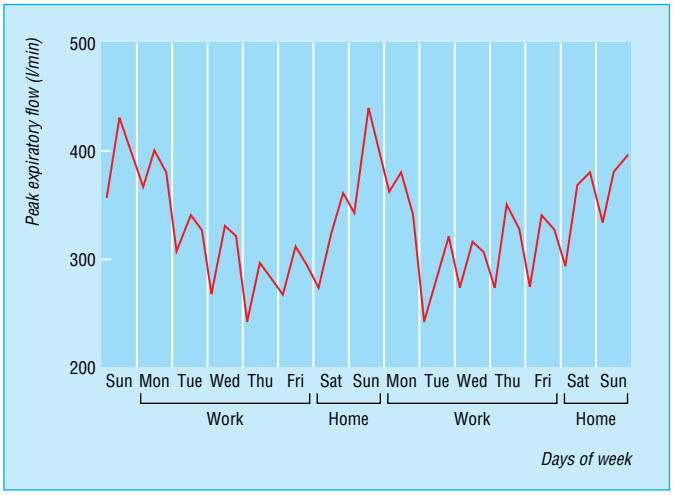
A bronchial provocation test (inhalation test) with the suspected agent may be required to give the patient advice about future employment. The test may precipitate severe bronchospasm, so the procedure must be undertaken in a specialist hospital unit with inpatient facilities. The individual is exposed to the suspected sensitiser in circumstances that most closely resemble their exposure at work. Forced expiratory volume in one second (FEV1), forced vital capacity, and responsiveness to histamine or methacholine are measured serially and then compared with serial measurements taken during a control challenge test performed on a separate day. An increase in airway hyper-responsiveness, particularly a late response, caused by the putative agent in concentrations that occur at work is taken as evidence of an allergic response. Although bronchial provocation testing is considered the gold standard test for the diagnosis of occupational asthma, false negatives can arise if the testing is conducted with the wrong material or if the concentration of the suspected agent is too low.

Management

Treatment of acute occupational asthma is the same as for asthma generally, but it is important to remove the sensitised individual from exposure to the substance causing their asthma, as subsequent exposure to even minimal quantities of the sensitising agent may precipitate severe bronchospasm. If their job entails working with the causative agent, relocation to another area will need to be considered. The employer's occupational physician will be able to advise on suitable areas for redeployment and will be in a position to liaise with the employee's manager. The employer should review their statutory risk assessments and control measures in the area

Examples of high and low molecular weight substances that may cause occupational asthma

Chemicals (low molecular weight)	Occupational group at risk/industrial use
Toluene di-isocyanate	Car or coach paint spray
Colophony (pine resin)	Electronics industry
Complex platinum salts	Platinum refinery workers
Proteins (high molecular weight)	
Flour or grain	Bakers
Rodent urinary proteins	Laboratory workers
Salmon proteins	Fish processing plant workers
Natural rubber latex	Healthcare professionals



Self recorded peak expiratory flow measurements showing a classic pattern of occupational asthma

Specialist investigation of occupational asthma

(a) Identification of atopy: skin prick tests with common allergens—for example, grass pollen, *Dermatophagoides pteronyssinus*, and cat fur
(b) Skin prick tests with specific extracts of putative sensitising agent
(c) Serology: radioallergosorbent tests (RAST) to identify specific IgE antibody
(d) Bronchial provocation test with the suspected causative agent

A worker who develops occupational asthma should avoid further exposure to the causative agent. As this often means relocation or loss of current employment, it is essential that the specific cause is identified accurately

where the affected employee was working to prevent other workers being similarly affected.

Reactive airways disease

Exposure to gases

Although fatalities from exposure to gases in the workplace are now rare in industrialised countries, inhalation accidents still occur relatively often. Accidental inhalation of gas (most commonly chlorine), fume, or vapour with irritant properties can lead to reactive airways disease. Frequently, individuals complain of a burning sensation in their nose and throat within minutes of exposure. The symptoms of asthma develop within 24 hours. The airway irritability usually resolves spontaneously but can persist indefinitely, and it may be provoked by a range of irritants or other provoking factors—for example, cold. The key to preventing the syndrome is good health and safety management.

On a wider scale, industrial accidents involving the release of a toxic irritant gas may cause pulmonary injury or even death in the surrounding population. The release of methylisocyanate from the Union Carbide pesticide plant in Bhopal, India, in 1984 resulted in many deaths from acute pulmonary oedema. Survivors still have chronic respiratory ill health.

Diagnostic criteria for reactive airways disease syndrome

- History of inhalation of gas, fume, or vapour with irritant properties
- Rapid onset of asthma like symptoms after exposure
- Bronchial hyper-responsiveness on methcholine challenge test
- Individual previously free from respiratory symptoms

The Bhopal disaster in India highlighted the need for rapid access to expert advice in the event of a chemical disaster

Byssinosis

The symptoms of byssinosis occur as a result of hypersensitive airways and an acute reduction in FEV1 in susceptible individuals when they are exposed to dusts of cotton, sisal, hemp, or flax. It occurs most commonly in cotton mill workers and is probably a response to inhaled organic contaminants of the cotton boll, such as cotton bract (leaves at the base of the cotton flower that become hard and brittle during harvesting and comprise a major constituent of cotton dust in the mill). Smokers are at increased risk of developing the disease, but the pathogenic mechanisms underlying the disease remain obscure.

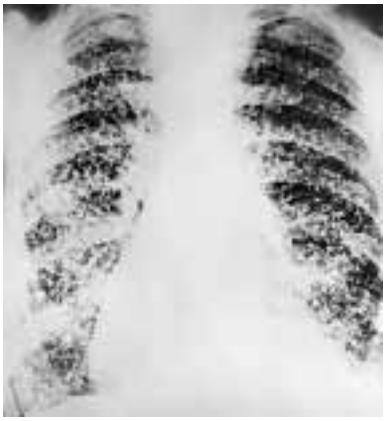
Characteristically, individuals experience acute dyspnoea with cough and chest tightness on the first day of the working week, three to four hours after the start of a work shift. The symptoms improve on subsequent working days, despite continued exposure to the sensitising agent. As the disease progresses the symptoms recur on subsequent days of the week, and eventually even occur at weekends and during holidays. Exposure of textile workers to cotton and flax dust per se does not seem to cause a significant loss of lung function. However, if the subset of workers who develop byssinosis are not removed from further exposure, they go on to develop long term respiratory impairment and subsequently have an excess risk of mortality from respiratory disease.

Harvested cotton consists of leaves, bracts, stems, bacteria, fungi, and other contaminants. Steaming or washing it before processing can reduce the biological activity of cotton

Pneumoconiosis

Pneumoconiosis is the generic term for the inhalation of mineral dust and the resultant diffuse, usually fibrotic, reaction in the acinar part of the lung. The term excludes asthma, neoplasia, and emphysema.

Silicosis is the commonest type of pneumoconiosis worldwide. It is caused by inhalation of crystalline silicon dioxide, and may affect people working in quarrying, mining, stone cutting and polishing, sandblasting, and fettling. Silicosis

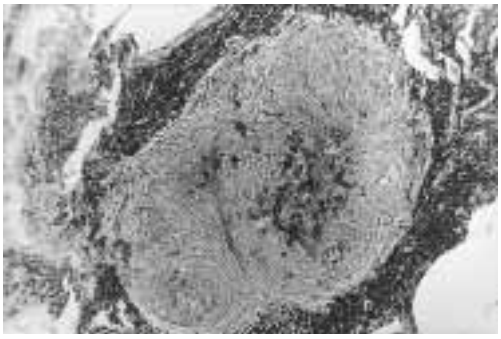


Chest radiograph of quarry worker showing extensive simple silicosis

occurs in several different forms depending on the level and duration of exposure.

Simple nodular silicosis is the most common form, and is similar clinically and radiographically to coal worker's pneumoconiosis. Chronic silicosis presents with increasing dyspnoea over several years and chest radiography shows upper lobe fibrosis or calcified nodules. Acute silicosis results from a brief but heavy exposure: patients become intensely breathless and may die within months. Chest radiographs show an appearance resembling pulmonary oedema. Accelerated silicosis occurs as the result of less heavy exposure and presents as slowly progressive dyspnoea caused by upper lobe fibrosis.

Coal worker's pneumoconiosis is caused by inhalation of coal dust, which is a complex mixture of coal, kaolin, mica, silica, and other minerals. Simple coal worker's pneumoconiosis usually produces no symptoms or physical signs apart from exertional dyspnoea. The diagnosis is made by a history of exposure and the presence of characteristic opacities on chest radiographs. A small proportion of individuals with simple coal worker's pneumoconiosis go on to develop progressive massive fibrosis which, when sufficiently advanced, causes dyspnoea, cor pulmonale, and ultimately death. Coal worker's pneumoconiosis is disappearing in developed countries as mines close and working conditions improve; however, it remains widespread in China and India.

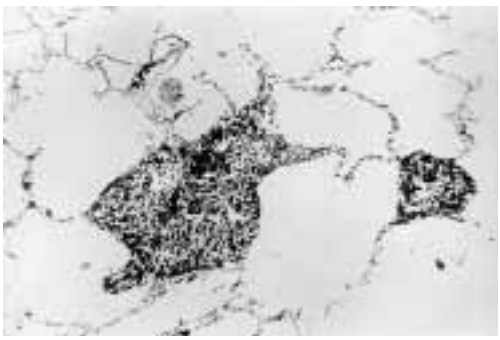


Silicotic nodule

Classification of radiographs for pneumoconiosis is based on the 1980 International Labour Office (ILO) system. This is a method of describing the pattern and severity of the change in groups of workers. The classification has been used worldwide for epidemiological research, surveillance, and medical checks of dust exposed workers

Chronic obstructive pulmonary disease and mining

The relationship between occupational exposure to coal dust and loss of ventilatory function is well established. However, after accounting for the effects of smoking and dust exposure, some miners still develop a severe decline in FEV1; the reasons for this are not fully understood. In the United Kingdom, chronic obstructive pulmonary disease due to coal dust is a prescribed industrial disease in those who have worked underground for at least 20 years and whose FEV1 is at least 1 litre below the predicted value.



Coal miners' pneumoconiosis

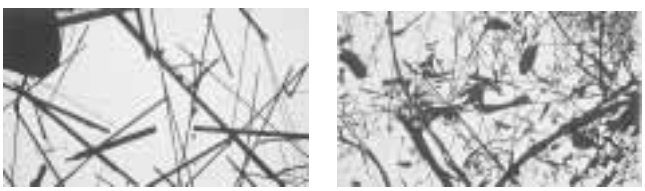
Asbestos related diseases

Exposure to asbestos causes several separate pleuropulmonary disorders, including pleural plaques, diffuse thickening of the pleura, benign pleural effusions, asbestosis, bronchial cancer, and malignant mesothelioma. Bronchial cancer and malignant mesothelioma are discussed in chapter 15.

Asbestosis is a diffuse interstitial pulmonary fibrosis caused by exposure to fibres of asbestos, and its diagnosis is aided by obtaining a history of regular exposure to any form of airborne asbestos. The presence of calcified pleural plaques on a chest radiograph indicates exposure to asbestos and helps to distinguish the condition from other causes of pulmonary fibrosis. Once the diagnosis is made, workers should be removed from further exposure. As there may be a synergistic effect between smoking and asbestosis in the development of lung cancer, workers should be encouraged to stop smoking.

Occupational groups at greatest risk of developing asbestos related diseases

- Carpenters and electricians
- Builders
- Gas fitters
- Roofers
- Demolition workers
- Shipyard and rail workers
- Insulation workers
- Asbestos factory workers



Blue asbestos fibres (left); white asbestos fibres (right)

Extrinsic allergic alveolitis

Extrinsic allergic alveolitis is a granulomatous inflammatory reaction caused by an immunological response to certain inhaled organic dusts and some low molecular weight chemicals. Farmer's lung and bird fancier's lung remain the most prevalent forms of the disease.

Acute extrinsic allergic alveolitis usually occurs after exposure to a high concentration of the causative agent. After a sensitising period, which may vary from weeks to years, the individual develops flu-like symptoms after exposure to the sensitising antigen. Prolonged illness may be associated with considerable weight loss, but symptoms usually improve within 48 hours of removal from the causative agent.

Chronic extrinsic allergic alveolitis is caused either by chronic exposure to low doses of the causative antigen, or as a consequence of repeated attacks of acute alveolitis over many years. It results in irreversible pulmonary fibrosis, and the dominant symptom is exertional dyspnoea. Weight loss may be considerable but other systemic symptoms are usually absent.

Diagnosis principally depends on a history of relevant exposure and on identification of a potential sensitising agent at home or at work. Inspiratory crackles may be heard on examination of the chest, and chest radiography in acute extrinsic allergic alveolitis may show a ground glass pattern or micronodular shadows. In chronic extrinsic allergic alveolitis lung shrinkage in the upper lobes is usually apparent. The diagnosis is confirmed by detailed pulmonary investigations and the demonstration of precipitating antibodies (precipitins) to the causal antigen in the serum.

Some causes of extrinsic allergic alveolitis

Disease	Source of antigen	Antigen
Farmer's lung	Mouldy hay and straw	*Micropolyspora faeni* *Thermoactinomyces vulgaris*
Bird fancier's lung	Bird excreta and bloom	Bird serum proteins
Bagassosis	Mouldy sugar cane	*Thermoactinomyces sacchari*
Ventilation pneumonitis	Contaminated air conditioning systems	*Thermophilic actinomycetes*
Malt worker's lung	Mouldy barley	*Aspergillus clavatus*
Mushroom worker's lung	Spores released during spawning	*Thermophilic actinomycetes*
Cheese washers' lung	Mould dust	*Penicillium casei*
Animal handler's lung	Dander, dried rodent urine	Serum and urine proteins
Chemical extrinsic allergic alveolitis	Polyurethane foam manufacture and spray painting	Toluene (TDI) and diphenylmethane di-isocyanate (MDI)

Farmers and pigeon fanciers often deny a relation between causative exposure and symptoms for fear of compromising their livelihood or hobby

Characteristic abnormalities of lung function in extrinsic allergic alveolitis

- Total lung capacity—reduced
- Residual volume—reduced
- Vital capacity—reduced
- Forced expiratory volume in one second (FEV1)—reduced
- FEV1/forced vital capacity—normal or increased
- Transfer factor for carbon monoxide—reduced*
- Gas transfer coefficient—reduced

*Sensitive indicator of the disease

Further reading

- Meyer JD, Holt DL, Chen Y, Cherry NM, McDonald JC. SWORD 1999. Surveillance of work-related and occupational respiratory disease in the UK. *Occ Med* 2001;51:204-8. *This paper reports on the 1999 SWORD results and findings*
- Mapp CE. Agents, old and new, causing occupational asthma. *Occup Environ Med* 2001;58:354-9. *An up to date review of the causative agents of occupational asthma, including detailed discussion on isocyanates, latex, flour, enzymes, glutaradehyde, and acrylates. The paper concludes with an extensive reference list for further reading*
- Kogevinas M, Anto JM, Sunyer J, Tobias A, Kromhard H, Burney P. Occupational asthma in Europe and other industrialised areas: a population based study. *Lancet* 1999;353:1750-4. *The results of a study of 15 637 young adults in Western European and other industrialised countries. The aim was to verify which occupations carry a high risk of asthma, and to estimate the proportion of asthma cases in the general population attributable to occupational exposures*
- Health and Safety Executive. Proposals for reducing the incidence of occupational asthma, including an Approved Code of Practice: Control of substances that cause occupational asthma. Sudbury: HSE Books, 2002. *This publication details the Health and Safety Executive's current strategy for reducing the incidence of occupational asthma in the United Kingdom*
- Baxter PJ, Adams PH, Aw TC, Cockcroft A, Harrington JM. *Hunter's diseases of occupations.* London: Edward Arnold, 2000. *A multiauthor textbook that contains several chapters on occupational lung disease written by leading experts in the field*

The picture of victims of the Bhopal disaster is reproduced with permission of Rex Features. The table showing the estimated number of cases of work related and occupational respiratory disease is adapted from Health and Safety Executive Statistics 2000-1

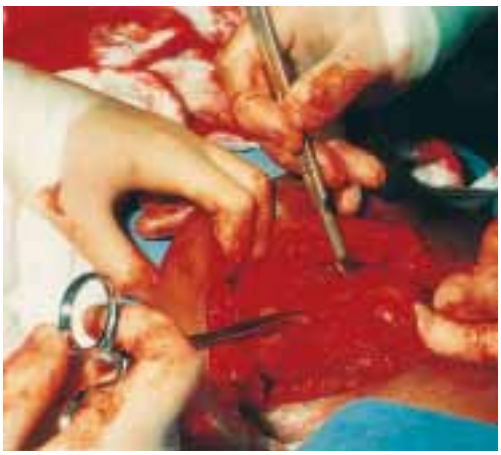
The photograph of a harvester is reproduced with permission from Jeremy Walker/Science Photo Library. The photograph of cotton is with permission from Bill Barksdale/Agstrct/Science Photo Library. The photograph of the Bhopal disaster is reproduced with permission from Rex Features Ltd.

14 Occupational infections

Dipti Patel

The pattern of infectious hazards at work changes constantly. Occupational infections, although not common, can be serious and easy to miss unless there is a high index of suspicion combined with an understanding of infectious disease. Furthermore, infections that are predominantly of historic interest in the developed world continue to pose a considerable problem in the developing world, and the changing pattern of travel means that those who visit or work overseas remain exposed. Drug resistance, the resurgence of certain diseases, and the emergence of new or previously unrecognised organisms further complicate matters, as does an increasing number of immunocompromised individuals. A detailed occupational history is therefore essential, as this will often point to the diagnosis of unusual illnesses caused by infectious hazards.

Occupational infections may be work specific or may be common in the general population, but they occur more often in those with occupational exposure. Like all occupational diseases, they are mostly preventable.



Healthcare workers are at risk acquiring infections from human sources such as bloodborne viruses

The traditional model of infectious disease causation

The epidemiological triangle

- An external agent—the organism that produces the infection
- A susceptible host—attributes that influence an individual's susceptibility or response to the agent—for example, age, sex, lifestyle
- Environmental factors that bring the host and agent together—factors that affect the agent and opportunity for exposure—for example, climate, physical surrounding, occupation, crowding

Basic concepts in infectious disease

- The *infectivity* of an agent is the proportion of exposed people who become infected (attack rate)
- The *pathogenicity* is the proportion of people exposed who develop clinical disease
- The *virulence* is the proportion of people with clinical disease who become severely ill or die
- The *infectious dose* is the number of organisms that are necessary to produce infection in the host, and this will vary according to the route of transmission and susceptibility of the host

Occurrence

- An infectious disease is *endemic* if there is a persistent low to moderate level of occurrence
- It is *sporadic* if the pattern of occurrence is irregular with occasional cases
- When the level of disease rises above the expected level for a period of time, it is referred to as an *epidemic*
- An *outbreak* is two or more cases of illness that are considered to be linked in time and place

Reservoir

This is any person, animal, arthropod, soil, etc. in which the infectious agent normally resides

Mode of transmission

This is the mechanism by which an infectious agent is spread from source or reservoir to a susceptible person—that is, direct (touching, biting, eating, droplet spread during sneezing, etc.), indirect (inanimate objects, fomites, vector borne) transmission, or airborne spread (dissemination of microbial aerosol to a suitable port of entry, usually the respiratory tract)

Main occupational groups at risk of infection

The three main categories of occupational infections are zoonoses, infections from human sources, and infections from environmental sources

Zoonotic infections

About 300 000 workers are at risk in the United Kingdom. Zoonotic infections include anthrax, leptospirosis, Q fever, Lyme disease, orf, and psittacosis. Workers at risk:

- Farmers and other agricultural workers
- Veterinary surgeons
- Poultry workers
- Butchers and fishmongers
- Abattoir workers and slaughtermen
- Forestry workers
- Researchers and laboratory workers—that is, animal handlers
- Sewage workers
- Tanners
- Military staff
- Overseas workers

Infections from human sources

About 2 million people are employed in the health service sector in the United Kingdom. Infections in this category include tuberculosis, erythema infectosum, scabies, bloodborne viruses, and rubella. Workers at risk:

- Healthcare workers
- Social care workers
- Sewage workers
- Laboratory workers
- Overseas workers
- Archaeologists (during exhumations)

Infections from environmental sources

Examples include legionellosis and tetanus. Workers at risk:

- Construction workers
- Archaeologists
- Engineering workers
- Military staff
- Overseas workers

The European Union has introduced the Biological Agents Directive (ongoing with updates since 1993), which is designed to ensure that the risk to workers from biological agents in the workplace is prevented or adequately controlled. In the United Kingdom this directive has been implemented via the Control of Substances Hazardous to Health (COSHH) Regulations 2002

Assessment of health risks of an infectious hazard, and its prevention or control should include:

- Details of the hazard group the agent belongs to
- The diseases it may cause
- How the agent is transmitted
- The likelihood of exposure and consequent disease (including the identification of those who may be particularly susceptible—for example, asplenic individuals, those with generalised immune deficiency, pregnant staff), taking into account the epidemiology of the infection within the workplace
- Whether exposure to the hazard can be prevented
- Control measures that may be necessary
- Monitoring procedures
- Need for health surveillance, which may include assessment of worker's immunity before and after immunisation

Hazard classification

In the United Kingdom biological agents are classified into four hazard groups according to their ability to cause infection

- *Group 1*—unlikely to cause human disease—for example, *Bacillus subtilis*
- *Group 2*—can cause human disease and may be a hazard to employees; it is unlikely to spread to the community, and there is usually effective prophylaxis or treatment available—for example, *Borrelia burgdorferi*
- *Group 3*—can cause severe human disease and may be a serious hazard to employees; it may spread to the community, but there is usually effective prophylaxis or treatment available—for example, *Bacillus anthracis*
- *Group 4*—causes severe human disease and is a serious hazard to employees; it is likely to spread to the community and there is usually no effective prophylaxis or treatment available—for example, *Ebola* virus

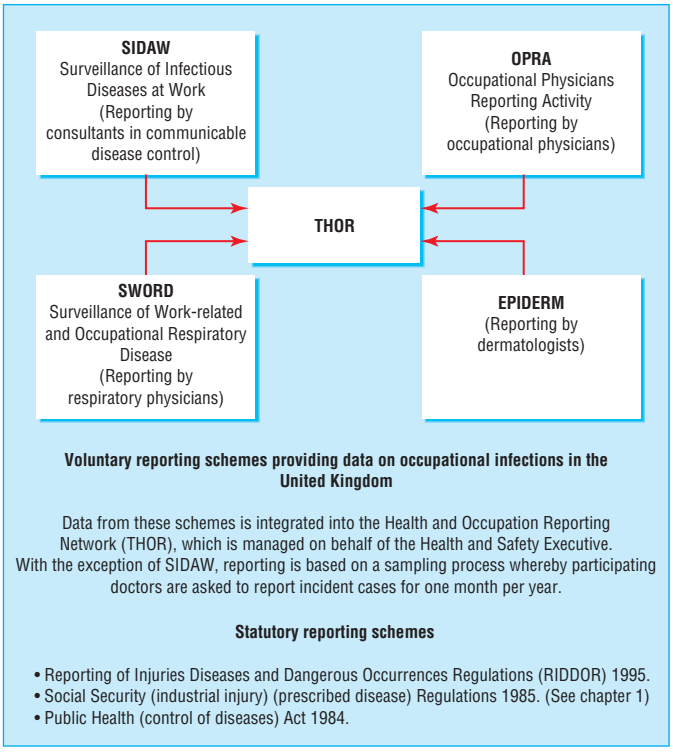
When a biological agent does not have an approved classification, the COSHH Regulations 2002 contain guidance on how biological agents should be classified. If in doubt, a higher classification should be assigned

Epidemiology

As with all occupational ill health statistics, no single source of information provides comprehensive data on occupationally acquired infections. In the United Kingdom, the principal data sources, although useful, underestimate the true incidence of occupational infections.

Data from UK reporting schemes. The industry with the highest estimated rates of infection per 100 000 workers per year for 1998-2000 was health and social care, followed by fishing, and agriculture and forestry. Diarrhoeal illnesses were the most frequently reported conditions

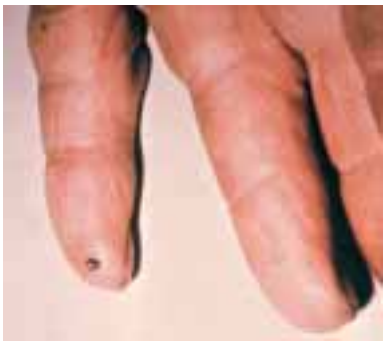
Disease	No of cases of infectious disease SIDAW 2000 (estimated)
Diarrhoeal illness	367
Hepatitis	–
Legionellosis	4
Leptospirosis	7
Ornithosis	4
Pulmonary tuberculosis	4
Q fever	–
Other (for example, scabies)	175
SIDAW total	**561**
SWORD/OPRA 2000	**77**
EPIDERM/OPRA 2000	**88**
RIDDOR (2000-1 provisional)	**93** Anthrax (1), chlamydiosis (2), hepatitis (4), legionellosis (14), leptospirosis (12), Lyme disease (3), Q fever (1), tuberculosis (15), others (41)
Prescribed diseases (1999-2000)	**7** Leptospirosis (1), tuberculosis (4), viral hepatitis (2)



Main information sources for occupational infections in the United Kingdom

Zoonoses

These are infections that are naturally transmissible from vertebrate animals to man. The World Health Organization estimates that there are over 200 zoonoses worldwide, and around 40 occur in the United Kingdom.



Typical painless blister of Orff

Protection of workers exposed to zoonotic infections relies on a number of control measures

Control of the disease in the animal reservoir
- Stock certification and vaccination (for example, anthrax or brucellosis)
- Quarantine measures (for example, for psittacine birds)
- Infection free feeds (for example, *Salmonella* free feed for poultry)
- Avoidance of contamination of animal drinking water
- Test and slaughter policies (for example, for bovine tuberculosis)
- Good standards of hygiene in stock housing
- Regular stock health checks by vets
- Meat inspection

Safe work practices
- Safe handling of animals or animal products (for all zoonotic infections)
- Safe disposal of carcasses and animal waste (for example, hydatid disease)
- Avoidance of equipment likely to cause cuts, abrasions, and grazes

Strict personal hygiene
- Covering existing wounds with waterproof dressings before work
- Prompt cleaning of any cuts or grazes that occur while handling animals
- Regular and correct hand washing, and avoidance of contact between unwashed hands and the mouth, eyes, or face

Personal protective equipment
- Waterproof aprons or parturition gowns
- Obstetric gauntlets for lambing or calving
- Face protection if there is a risk of splashing from urine or placental fluids
- Plastic or synthetic rubber gloves for oral or rectal examinations
- Gloves, overalls, and face protection for slaughtering animals or dressing carcasses
- Chainmail gloves for butchers

Other measures
- Immunisation of at risk worker (anthrax, Q fever)
- Provision of health warning cards (leptospirosis)

NB
For protection of laboratory workers advice on control measures has been provided by the Advisory Group on Dangerous Pathogens

(Adapted from Health and Safety information sheet "Common zoonoses in agriculture")

Anthrax

Also known as malignant pustule, Woolsorter's disease, and Ragpicker's disease, anthrax is a notifiable disease, a prescribed disease, and RIDDOR 1995 reportable. It is an acute infection caused by *Bacillus anthracis* (a spore forming Gram positive bacterium), and the normal animal reservoirs are grazing mammals such as sheep, cattle, and goats. Human anthrax is primarily an occupational hazard for workers who process hides, hair, wool, bone, and bone products, but it also occurs in vets and agricultural workers who handle infected animals. It is rare in the United Kingdom (only three cases of anthrax were reported in England and Wales between 1998 and 2001), occurring in those who work directly or indirectly with infected animal products from epizootic areas. Most cases of anthrax occur in Africa, the Middle East, and the former Soviet Union.

Cutaneous anthrax accounts for 95-8% of cases, and occurs when the organism enters a cut or an abrasion. After an incubation period of one to seven days, a small papule develops at this site. Over 24-48 hours, it enlarges, eventually forming a characteristic black ulcer (eschar). If not treated, cutaneous anthrax may progress to bacteraemia, meningitis, and death.

Pulmonary and gastrointestinal anthrax occur infrequently, and are the result of inhalation and ingestion of anthrax spores, respectively. In pulmonary anthrax (Woolsorter's disease) non-specific upper respiratory tract symptoms follow an incubation period of one to six days. Rapid deterioration in respiratory function and death generally follow unless treatment is started promptly. Gastrointestinal anthrax is characterised by severe abdominal pain, watery or bloody diarrhoea, and vomiting. Progression to bacteraemia is usually two to three days. Case fatality in both these forms of anthrax is high.

Most naturally occurring strains of anthrax are susceptible to penicillin, although doxycycline and ciprofloxacin have been used recently. Immunisation is also available for at risk workers, and oral antibiotics (ciprofloxacin and doxycycline) have been used as prophylaxis for individuals exposed to anthrax spores.

Patient with cutaneous anthrax

Anthrax has recently received attention because of its potential for use in bioterrorism. Other potential bioterrorism organisms include:

Organism (disease)	Potential source	Ability to cause disease
Brucella (Brucellosis)	Aerosol or food	High
Clostridium botulinum toxin (Botulism)	Food, water, or aerosol	High
Coxiella burnetii (Q fever)	Aerosol or food supply	High
Variola virus (Smallpox)	Aerosol	High
Vibrio cholerae (Cholera)	Food, water, or aerosol	Low

Leptospirosis

Leptospirosis is also known as Weil's disease, canicola fever, haemorrhagic jaundice, mud fever and swineherd disease. It is a notifiable disease, prescribed disease, and RIDDOR 1995 reportable. Leptospirosis is a rare cause of septicaemia caused by pathogenic leptospires belonging to the genus

Leptospira interrogans (*Li*). The genus has over 200 serovars; of most importance in humans are *Li hardjo* (cattle associated leptospirosis), *Li icterohaemorrhagiae* (Weil's disease), and *Li canicola*. The principal animal reservoirs are cattle, rats, and dogs, respectively.

The distribution of human disease depends on the local prevalence of animal infection and local environmental conditions. At risk occupational groups include agricultural workers, farmers, vets, miners, abattoir workers, and sewer and canal workers. In 2001 there were 25 notified cases of leptospirosis in England and Wales (predominantly caused by *Li hardjo*).

Leptospirosis is usually acquired by direct contact with infected animals or their urine, contaminated soil, food, or water (a hazard for those indulging in watersports). The incubation period is usually five to 14 days, and symptoms, which are not serotype specific, typically consist of fever, flu-like symptoms, headache, myalgia, photophobia, and conjunctival injection. In severe cases (often associated with *Li icterohaemorrhagiae*) haemorrhage into skin and mucous membranes, vomiting, jaundice, and hepatorenal failure may occur.

Mild infection is often self limiting, but penicillin, erythromycin, and doxycycline are all effective treatments. For more severe disease, intensive and specialised therapy is necessary.

Immunisation of animals is possible for certain serovars, and in some countries (Japan, Italy, Spain) immunisation of at risk workers against certain serovars is available. In the United Kingdom, workers who may be exposed to leptospires usually carry an alert card provided by their employer to warn their doctors should they develop such symptoms.

Farm workers are at increased risk of catching animal borne diseases

Transmissible spongiform encephalopathies (TSEs)

Prion disease

These are a group of progressive and fatal neurological disorders occurring in humans and certain animal species. TSEs are thought to be caused by infectious proteins (prions) that are unusually resistant to conventional chemical and physical decontamination. They do not seem to be highly infectious and, with the exception of scrapie, do not seem to spread through casual contact.

Bovine spongiform encephalopathy (BSE) was first recognised in British cattle in 1986. Its origin is still uncertain, but it probably originated in the early 1970s, developing into an epidemic because of changing practices in rendering cattle offal to produce animal protein in the form of meat and bonemeal, which was included in compound cattle feed. This resulted in the recycling and wide distribution of BSE. In 1996, a previously unrecognised form of Creutzfeldt-Jakob disease (CJD) occurring in younger patients (range 14-53 years, mean 28 years), with a different symptom profile and different postmortem changes in the brain tissue, was identified in the United Kingdom.

The Government's Spongiform Encephalopathy Advisory Committee concluded that the most likely explanation for the emergence of this variant CJD (vCJD) was that it had been transmitted to humans through exposure to BSE as a result of consumption of contaminated bovine food products.

A major concern now is the risk of transmission in a healthcare setting. Although there have been no reported cases of nosocomial transmission of vCJD, an expert group has been established by the UK government to advise on prevention and management of possible exposures.

Transmissible spongiform encephalopathies (TSEs)

Human TSEs
- Creutzfeldt-Jakob disease (CJD)
- Variant CJD (vCJD)
- Gerstmann Sträussler Scheinker syndrome
- Kuru
- Fatal familial insomnia

Animal TSEs
- Scrapie (sheep and goats)
- Bovine spongiform encephalopathy (BSE) (cattle)
- Transmissible mink encephalopathy (farmed mink)
- Chronic wasting disease (deer)
- Feline spongiform encephalopathy (domestic cats and captive exotic felines)
- Spongiform encephalopathy (captive exotic ungulates)

A number of measures have been taken to minimise disease transmission among animals and humans, and although there is no clear evidence of occupational risk, advice on safe working practices has been provided by the Advisory Committee on Dangerous Pathogens. Those potentially at risk include workers in abattoirs, slaughterhouses and rendering plants; farmers; neurosurgeons; pathologists; and mortuary technicians

The incubation period of vCJD is unknown, but is likely to be several years. The infectious dose is also unknown, and is likely to be dependent on the route of exposure. However, by February 2002, a total of 114 individuals throughout the United Kingdom (106 dead and 8 alive) were considered to have had definite or probable vCJD. There is currently no evidence to link any cases of vCJD with surgical procedures or with transmission by blood, but the possibility cannot be ruled out.

It is already current practice to dispose of instruments used on anyone showing symptoms of vCJD. A problem, however, occurs in those who have presymptomatic disease; precautions to avoid this theoretical risk of transmission are therefore essential. Any assessment of risk of transmission from instruments must consider a wide range of scenarios, and precautionary measures should be taken against risks that might occur, even if the level of risk is not known. The key message in reducing any risk of vCJD transmission is the rigorous implementation of washing, decontamination, and general hygiene procedures.

In January 2001, a recommendation that single use instruments were used for tonsillectomy and adenoidectomy surgery was made. This recommendation was withdrawn in December 2001 because adverse incidents (mainly haemorrhage, but also one death) were reported after the introduction of single use instruments. It was felt that on balance, the single use instruments represented an actual risk to patients, whereas the concerns regarding vCJD transmission were only theoretical. Further information is available at http://www.open.gov.uk/doh/coinh.htm.

Although it is likely that most of the UK population has been exposed to BSE, the true number of individuals who have been infected is not known

Reasons that vCJD might be spread from person to person in healthcare settings

- Classical CJD has been transmitted from person to person by a range of medical procedures including surgery, grafts or transplants, and treatment with pituitary extracts, and about 1% of classical CJD cases in the past are considered to have been iatrogenic
- Abnormal prion protein has been shown in the lymphoreticular tissue (tonsils, spleen, and lymph nodes) of patients with established vCJD
- Abnormal prion protein has been shown in the appendix of a patient who subsequently developed vCJD
- Although, to date, the transmissible agent has not been shown in blood, it is possible that abnormal prion protein, at concentrations not detectable with current techniques, may be associated with circulating B lymphocytes and with other cells of the immune and circulatory systems
- Abnormal prion protein has been shown to be highly tenacious and may not be inactivated by conventional sterilisation and decontamination procedures

Infections from human sources

These infections are of most relevance to healthcare workers. They are important because healthcare workers are at high risk of acquiring infections occupationally and they are also a potential source of infection to their patients, particularly those who are immunologically impaired.

Bloodborne viruses

Occupational exposure to blood or body fluids poses a small risk of transmission of bloodborne pathogens. Those presenting the greatest crossinfection hazard are HIV, and hepatitis B and hepatitis C viruses. Although healthcare staff are at greatest risk, other occupational groups (for example, police officers) may also be exposed.

The risk of infection depends on the type and severity of the exposure, the infectivity of the source patient, the immune status of the exposed healthcare worker, and the availability of treatment after exposure.

Prevention entails minimising exposure to blood or body fluids, and consists of strict infection control, adherence to universal precautions, immunisation against hepatitis B, and prompt management of any occupational exposure.

Healthcare workers infected with bloodborne viruses can potentially transmit infection to their patients, and although the risk is small, guidelines exist in many countries to reduce this risk further. In the United Kingdom, all healthcare workers who perform exposure prone procedures are required to provide evidence that they are immune to hepatitis B as a result of immunisation, or that they are not HBe antigen (HBeAg) positive. Because of transmissions of HBV associated with codon 28 precore mutations, those who are HB surface antigen positive, but HBeAg negative, must now be tested for hepatitis B virus DNA; they may perform exposure prone work provided that their HB viral load is below 10^3 genome equivalents per millilitre, and this is subject to annual testing.

Exposure prone procedures are invasive procedures where there is risk that injury to the worker may result in the exposure of a patient's open tissues to the blood of the worker. These incude procedures where a healthcare worker's gloved hand may be in contact with sharp instruments or tissues inside a patient's open body cavity, wound, or confined anatomical space where the hands may not be completely visible at all times

Risk of transmission of bloodborne viruses

High risk body fluids
- Cerebrospinal fluid
- Peritoneal, pleural, pericardial fluid
- Synovial fluid
- Amniotic fluid
- Breast milk
- Vaginal secretions
- Semen
- Saliva associated with dentistry
- All visibly blood stained fluid
- Unfixed organs or tissues

Low risk body fluids (unless visibly bloodstained)
- Urine
- Faeces
- Vomit

Significant exposures
- Percutaneous injury
- Exposure of broken skin
- Exposure of mucous membrane

Estimated risk of seroconversion after percutaneous exposure

HIV: 0.32% (based on data of 6202 healthcare workers). Risk of mucous membrane exposure is 0.09%, and there have been no transmissions associated with exposure of intact skin. In the United Kingdom there have to date (March 2002) been five definite occupationally acquired transmissions of HIV. Worldwide by 1999, 102 definite and 217 possible cases of occupationally acquired HIV had been reported

The risk of percutaneous exposure is increased if the injury is deep, the device is visibly blood stained, the injury is from a needle placed in artery or vein, or the source patient has terminal HIV infection

Hepatitis C virus: 1.8%
Hepatitis B virus: 37-66% from HBeAg positive source; 23-37% from HBsAg positive source
Effective immunisation is available for hepatitis B virus and 80-90% of individuals mount an adequate response

Post-exposure prophylaxis (PEP)

HIV

Most countries now recommend a four week course of zidovudine with lamivudine, and many recommend the addition of a protease inhibitor. The choice of drugs, doses, route of administration, and the length of PEP are somewhat empirical. However, because most studies indicate a time limited response to PEP, the need for timely and early therapy is vital. In the United Kingdom, HIV PEP generally consists of zidovudine and lamivudine (Combivir) with nelfinavir, indinavir, or soft gel saquinavir (March 2002)

Hepatitis B virus

Hepatitis B virus immunoglobulin (HBIG) is available for passive protection and is normally used in combination with hepatitis B vaccine to confer passive-active immunity to susceptible individuals after exposure. The post-exposure efficacy of combination HBIG and hepatitis B vaccine has not been evaluated in the occupational setting, but increased efficacy (85-95%) has been observed perinatally. Although HBIG may not completely inhibit virus multiplication, it may prevent severe illness and the development of a chronic carrier state

Hepatitis C virus

No effective PEP exists. Recommendations for post-exposure management are therefore intended to achieve early identification of infection, with appropriate specialist referral. Although consistent data are lacking, one uncontrolled trial has shown a substantially better response rate of early treatment using interferon compared with treatment of patients with chronic disease

Individuals who, as a result of testing, are found to be hepatitis C RNA positive should not perform exposure prone procedures. However, hepatitis C infected workers who have been successfully treated with antiviral therapy and remain hepatitis C virus RNA negative six months after finishing treatment should be able to resume exposure prone procedures or start professional training for a career that relies on the performance of exposure prone procedures.

HIV testing is not compulsory for healthcare workers in the United Kingdom and many other countries. In the United Kingdom, professional regulatory bodies state that workers who may have been exposed to HIV have an ethical responsibility to be tested. If found to be HIV infected, exposure prone work is prohibited. The UK Department of Health is currently reviewing their policy.

Recent UK guidance for hepatitis C
- Healthcare workers who carry out exposure prone procedures and already know themselves to be infected with hepatitis C should be tested for hepatitis C virus RNA (if not already done)
- All healthcare professionals intending to undertake professional training for a career that relies on the performance of exposure prone procedures should be tested for hepatitis C infection
- Those who perform exposure prone procedures and believe that they may have been exposed to hepatitis C should seek and follow confidential and professional advice on whether they should be tested for hepatitis C

Tuberculosis

Tuberculosis is a notifiable disease, a prescribed disease, and RIDDOR 1995 reportable.

Mycobacterium tuberculosis continues to be the leading cause of adult death from any single infectious agent worldwide.

Bloodborne viruses and risk to patients

HIV

- In the United Kingdom, the Expert Advisory Group on AIDS (EAGA) provides guidance on look-back procedures for HIV. As UK studies of over 30 000 patients after look-back exercises have shown no evidence of transmission of HIV to patients, it is likely that look-back procedures for HIV in the United Kingdom will stop
- Two incidents of transmission from a healthcare worker to a patient have been reported: a Florida dentist who infected six patients, and a French orthopaedic surgeon who infected one patient
- Dr Patrick Ngosa, an HIV positive obstetrician, was removed from the UK General Medical Council's Register in 1997 when it was discovered that he had refused to have an HIV test and continued to perform exposure prone procedures after learning that a former sexual partner was HIV positive. A total of 1750 women on whom he had operated were sent letters informing them that there was a possibility that they had been exposed to HIV

Hepatitis B virus

- A number of look-back studies involving surgical staff from 1975 to 1990 have identified transmission risks of 0.9-20%
- The most recent look-back exercise for hepatitis B virus in the United Kingdom was in 2001. About 350 patients were contacted in Fife, when infection in two patients was traced back to one healthcare worker
- A surgeon infected with hepatitis B (Dr Gaud) who lied about his infectivity was convicted and jailed for the common law charge of "public nuisance" after knowingly operating on patients and putting them at risk of infection

Hepatitis C virus

- In the United Kingdom there have been five patient notification exercises after investigations of hospital acquired hepatitis C infection. Since 1994 there have been 15 documented transmissions of hepatitis C virus to patients from infected healthcare workers during exposure prone procedures

The emergence of multidrug resistant tuberculosis (MDRTB), with its high case fatality, its prolonged sputum positivity (and consequently, high transmission risk), and its complex treatment has re-emphasised the importance of tuberculosis control.

Tuberculosis remains a hazard in the healthcare setting, and incidence in healthcare workers parallels (but is higher than) that in the community; a study in the mid-1990s found about a twofold increased risk of tuberculosis among healthcare workers in England and Wales. Healthcare workers should therefore be protected against infection, and measures should be taken to detect tuberculosis in new or existing staff in order to protect their patients and colleagues. Protection begins at pre-employment, and continues with strict infection control measures for nursing infected patients.

In the United Kingdom, protection of healthcare workers should follow the guidelines produced by the Joint Tuberculosis Committee of the British Thoracic Society.

Adults with non-pulmonary tuberculosis can usually be nursed on general wards, but those with pulmonary tuberculosis should initially be admitted to a single room vented to the open air until their sputum status is known. Those with smear positive sputum should be managed as infectious. In the case of known or suspected MDRTB, particular care must be taken, and patients should be admitted to a negative pressure single isolation room until MDRTB is excluded or until sputum smears have been negative on three consecutive occasions over 14 days.

Outbreaks of MDRTB in the United States and Europe have emphasised the importance of control. These outbreaks have occurred predominantly in institutional settings (prisons, residential homes, and hospitals) and have mainly been in HIV infected patients. Contributory factors in these outbreaks included lapses in respiratory isolation, inadequate ventilation in isolation rooms, and "immunocompromised convergence" (the assembling of immunocompromised HIV infected patients in institutions).

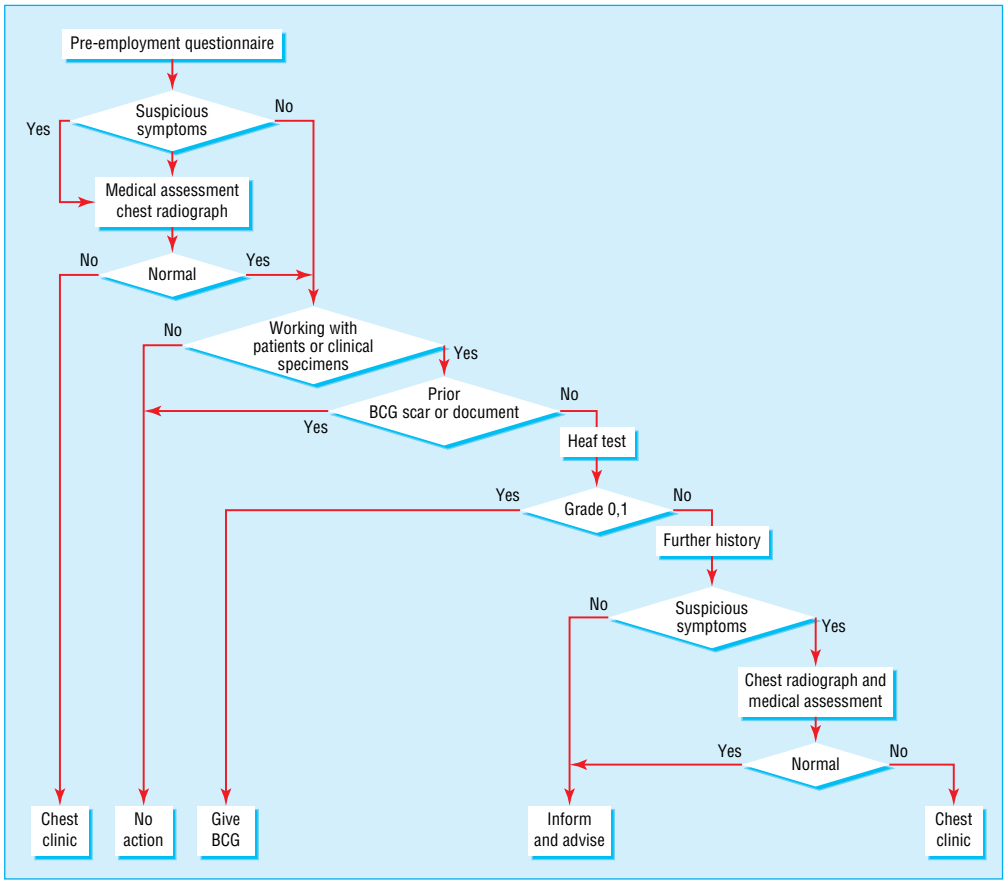
Multidrug resistant tuberculosis (MDRTB)

MDRTB is tuberculosis resistant to at least isoniazid and rifampicin

Effective control of MDRTB requires a multidisciplinary approach involving the hospital infection control team, microbiologist, tuberculosis physician, consultant in communicable disease control, engineers, and occupational health

- Visitors to patients with known or suspected MDRTB should be kept to a minimum
- The number of healthcare workers exposed to MDRTB patients should be kept as low as possible
- All who enter the rooms of MDRTB patients should wear suitable particulate masks that filter down to particles of 1 micron in diameter
- Staff should wear masks during aerosol generating procedures, such as sputum induction, bronchoscopy, and pentamidine therapy. These procedures should only be performed in suitably ventilated facilities
- Individuals who have not been checked for immunity to tuberculosis, or those with a negative skin test who have not received BCG vaccination should avoid contact with MDRTB patients, as should those who are immunocompromised

The decision to discontinue strict isolation and infection control procedures should only be made after discussion between the clinician with responsibility for the patient, the hospital infection control team, occupational health, and a consultant in communicable diseases



Protection of healthcare workers with tuberculosis

When a patient or member of staff is found to have tuberculosis, infection control and occupational health staff have to assess the need for contact tracing. In the United Kingdom, most staff are not considered to be at special risk and should be reassured and advised to report any suspicious symptoms. Those who are immunocompromised, have undertaken mouth to mouth resuscitation, prolonged high dependency care, or repeated chest physiotherapy without appropriate protection should be regarded as close contacts and followed up according to national guidelines. Similar precautions should be taken if the index case is highly infectious.

Chickenpox

Chickenpox is a systemic viral infection resulting from primary infection with varicella zoster virus. It is highly infectious and transmitted directly by personal contact or droplet spread, and indirectly through fomites. Shingles (herpes zoster) is a reactivation of dormant virus in the posterior root ganglion and can be a source of infection generally by contact with the skin lesions, but occasionally by the respiratory route in immunocompromised individuals.

Primary infection in adults can be severe, resulting in a higher frequency of complications such as pneumonia, encephalitis, and hepatitis, but its main importance is the risk to non-immune pregnant women and the immunosuppressed.

Although the prevalence of seropositivity for varicella zoster virus in healthcare workers in temperate climates is high (90-98%), nosocomial exposure to the virus is a major occupational health problem requiring non-immune healthcare workers to be excluded from patient contact from day 8 to 21 after a substantial exposure. A live attenuated vaccine is now available in many parts of the world.

Complications of varicella zoster virus

- Severe disease due to fulminating varicella pneumonia is more likely in adults, especially pregnant women, and smokers
- Pregnant women are at greatest risk late in second or early third trimester
- In the immunocompromised and neonates, disseminated or haemorrhagic varicella is more likely
- The risk to fetus and neonate from maternal infection relates to gestation at time of infection
  - *First 20 weeks*—congenital varicella syndrome (limb hypoplasia, microcephaly, hydrocephalus, cataracts, growth retardation, and skin scarring)
  - *Second and third trimester*—herpes zoster in otherwise healthy infant
  - *A week before to a week after delivery*—severe and even fatal disease in the neonate (particularly premature babies)
- Human varicella zoster immunoglobulin (VZIG) is available and can be given for post-exposure prophylaxis in individuals who fulfill the following conditions:
  - a clinical condition that increases risk of varicella infection
  - no antibodies to varicella zoster virus
  - substantial exposure to chickenpox or herpes zoster
- A substantial exposure to varicella zoster virus depends on:
  - the type of infection in the index case—for example, the risk of acquiring infection from an immunocompetent individual with non-exposed shingles is remote
  - the timing of the exposure in relation to onset of rash in the index case—the critical time for chickenpox or disseminated zoster is 48 hours before the onset of rash until crusting of lesions for varicella zoster virus, and day of onset of rash until crusting in localised zoster
  - closeness and duration of contact—contact in same room >15 minutes, or face to face contact
- The recommendation that VZIG is used for exposed non-immune pregnant women during the first 20 weeks of pregnancy is based on biological plausibility. No evidence exists showing that the risk of congenital varicella syndrome is reduced
- VZIG is not recommended for healthy healthcare workers, but in the United States, varicella vaccine is recommended for use in susceptible individuals after exposure; data from hospital and community settings suggest that it is effective in preventing illness or modifying severity if used within three days of exposure

Other infections

Other infections worth mentioning include skin infection in engineers associated with the re-use of cutting oils (which can lead to oil mists being contaminated with bacteria and fungi), pseudomonal otitis externa in deep sea divers who use saturation techniques, and legionellosis (which can occasionally be occupationally acquired). Finally, travel associated infections are becoming an important cause of occupationally acquired disease with the increase in international business travel and overseas workers (see *ABC of Healthy Travel*).

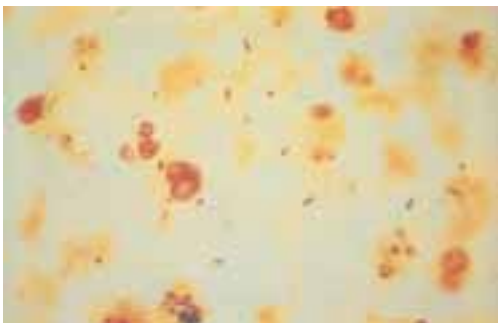
Legionellosis (Legionnaire's disease, Pontiac fever)
This infection is RIDDOR 1995 reportable. It is an acute bacterial infection caused by a Gram negative bacillus belonging to the genus *Legionella*. Two clinical presentations are recognised: Legionnaire's disease and Pontiac fever, and the majority of infections are due to *L. pneumophila*. The bacillus is an ubiquitous aquatic organism that thrives in warm environments (25-45°C, but preferably at 30-40°C), and is often isolated from natural habitats (rivers, creeks, hot springs) and from artificial equipment where the temperature is maintained at levels favouring bacterial proliferation.

Transmission of infection is from inhalation of contaminated aerosols, and both Legionnaire's disease and Pontiac fever present initially with non-specific flu-like symptoms. Pontiac fever occurs after an incubation period of 4 to 66 hours, and is a self limiting non-pneumonic form of the infection. By contrast, the incubation period for Legionnaire's disease is two to ten days. Initial symptoms of fever, malaise, anorexia, and myalgia are followed by progression to pneumonia and associated multisystem involvement, with diarrhoea, vomiting, confusion, and renal failure. Case fatality can range from 5% to 15%, but may be higher in outbreaks.

Treatment generally consists of erythromycin (although rifampicin may be used as an adjunct). If infection is confirmed, local public health authorities need to be notified as contacts may need to be identified, and the source of infection needs to be established and appropriately controlled.

Legionnaire's disease

- Travel abroad is a major risk factor for Legionnaire's disease in the United Kingdom, with nearly 50% of cases being contracted abroad
- About 15% of UK cases are linked to local outbreaks (caused by wet cooling systems or hot water systems), and roughly 2% are hospital acquired. Many cases are sporadic, or from an unidentified source
- Hospital outbreaks in particular have high case fatalities
- The highest risk of infection occurs with water systems leading to the aerosolisation of water that is stored at temperatures of 25-45°C. This includes:
  - wet cooling systems (for example, cooling towers and evaporative condensers)
  - hot water systems (especially showers)
  - whirlpool spas
  - indoor and outdoor fountain and sprinkler systems
  - humidifiers
  - respiratory therapy systems
  - industrial grinders
- Prevention of infection relies on ensuring that equipment and systems are kept as clean as possible, and disinfected regularly. Where possible, water temperatures should be kept above 50 °C or below 20°C. Use of biocides may also need to be considered. In the United Kingdom, the Health and Safety Executive provides guidance on the prevention and control of legionellosis



Legionella bacteria

Conclusion

The extent of occupationally acquired infections is unknown, but it is likely that they are extremely common, particularly mild infections in agricultural and healthcare workers. Preventing infection is an important aspect of occupational health practice as it will impact favourably on communicable disease in the general population. Similarly, the control of communicable disease in both the general (and animal) population will decrease the risk to certain occupational groups.

The table showing Data from UK reporting schemes depicting the industries with the highest estimated rates of infection is adapted from Health and Safety Executive Statistics 2000-1. The figures showing the protection of workers with tuberculosis are adapted from the guidelines of the Joint Tuberculosis Committee of the British Thoracic Society. Control and prevention of tuberculosis in the United Kingdom: code of practice 2000. Http://www.brit-thoracic.org.uk

Further reading

- Heponstall J, Cockcroft A, Smith R. Occupation and infectious diseases. In: Baxter P, Adams PH, Cockroft A, Harrington JM, eds. *Hunter's diseases of occupations.* London: Edward Arnold, 2000:489-517. *This chapter provides comprehensive information on occupational infections*
- Hawker J, Begg N, Blair I, Reintjes R, Weinberg J. *Communicable disease control handbook,* 1st ed. Oxford: Blackwell Science Ltd, 2001
- Chin J. *Control of communicable diseases manual, 17th ed.* Washington: American Public Health Association, 2000.
  Both these references, although aimed at public health practitioners, provide extensive detail on communicable diseases, their epidemiology, clinical features, prevention, and control
- http://www.who.int
- http://www.cdc.gov
- http://www.phls.co.uk
  These websites are excellent resources for information on infectious diseases (both occupational and non-occupational)
- http://www.open.gov.uk/doh/dhhome.htm *The UK department of health website is particularly useful for information on bloodborne viruses and BSE*
- http://www.hse.gov.uk *This site provides practical and clear information on prevention and control of a variety of infectious hazards in the workplace, and is also a source of occupational ill-health statistics*

15 Occupational cancers

John Hobson

The first report of cancer caused by occupational exposure was in 1775 by Percival Pott, a British surgeon who described scrotal cancer in boy chimney sweeps. A century later, in 1895, Rehn, a German surgeon working in Frankfurt, treated a cluster of three cases of bladder cancer in workers at a local factory producing aniline dyestuffs from coal tar.

Occupational cancer is any malignancy wholly or partly caused by exposures at the workplace or in occupation. Such exposure may be because of a particular chemical (such as β-naphthylamine), a physical agent (such as ionising radiation), a fibre like asbestos, a biological agent (such as hepatitis B virus), or an industrial process in which the specific carcinogen may elude precise definition (such as coke production).

Overall it is estimated that 4% of all cancers are caused by occupation (range 2-8%), but for bladder cancer this may be as high as 20%. In the working population as many as one in five cancers may be attributable to exposure in the workplace. In England and Wales at least 3000 men die each year from potentially preventable malignancies.

The International Agency for Research on Cancer (IARC) was set up to identify carcinogenic hazards to humans. To date, 874 chemicals, groups of chemicals, complex mixtures, occupational exposures, cultural habits, and biological and physical agents have been evaluated. The findings have been published in 79 monographs and eight supplements.

Foundry workers may be exposed to a complex mixture of carcinogenic agents in fumes

Of all the occupationally related diseases, cancer evokes particular concern and strong emotions, because of the opportunity afforded for attribution, blame, and compensation. However, occupational cancers also have unique potential for prevention

Mechanisms of cancer

Cancer is a genetic disorder of somatic cells and can be triggered by the genotoxic action of carcinogens. There are five or six independent stages of carcinogenesis, each of which is rate limiting. The best available model is colorectal cancer, which requires seven independent genetic events. The three key stages are initiation (by a mutagen), promotion (where development of tumours is enhanced by other stimuli to cell proliferation such as lung fibrosis), and progression (development of malignant tumours from benign neoplasms).

Mutations in tumour suppressor genes (for example, *p53*) are particularly important, and half of all cancers contain *p53* mutations, of which there are 6000 possible point mutations. Several environmental and occupational carcinogens are linked to *p53* mutations—for example, ultraviolet light and skin cancer, and tobacco and oral cancer. Other factors linked with *p53* include alcohol, vinyl chloride, and asbestos.

Most carcinogens are genotoxic (DNA reactive) and cause mutation. There is no threshold below which they are not carcinogenic and therefore exposure levels are set at acceptable levels. Tests for genotoxicity such as Ames and fluorescent in situ hybridization (FISH) are now well established. The Ames test is the most widely used procedure for assessing the mutagenicity of chemicals. The relative mutagenic potency of an agent is indicated by the number of bacterial colonies growing on a plate containing the toxic agent relative to those growing on a plate containing normal medium. FISH is used to assess chromosomal abnormalities.

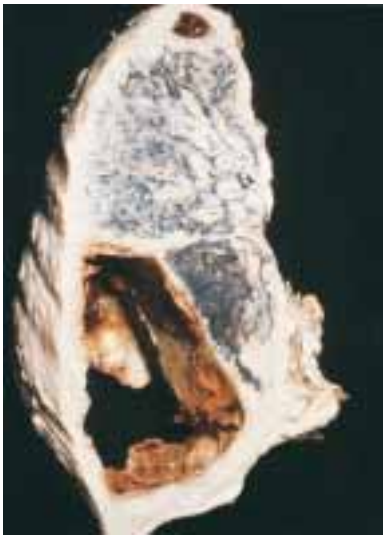
Epigenetic carcinogens (also known as non-genotoxic or cocarcinogens) act more directly on the cell itself, through

International Agency for Research on Cancer (IARC) classifications to date

Group		Number
1	Carcinogenic	87
2A	Probably carcinogenic	63
2B	Possibly carcinogenic	233
3	Unclassifiable as to carcinogenicity in humans	490
4	Probably not carcinogenic to humans	1 (caprolactam)

hormonal imbalances, immunological effects, or promoter activity, to cause abnormal cell proliferation and chromosomal aberrations that affect gene expression. These carcinogens have a threshold dose for carcinogenicity and it is possible to set exposure levels. There is probably a minimal threshold dose as well as a clear dose-response relation influencing the occurrence of cancers. For example, all workers involved in distilling β-naphthylamine eventually developed tumours of the urothelial tract, whereas only 4% of rubber mill workers who were exposed to β-naphthylamine (a contaminating antioxidant (at 0.25%) used in making tyres and inner tubes) developed bladder cancer over a 30 year follow up.

Polymorphisms are different responses to the same factor such as a drug. Slow acetylators who are heavy smokers are 1.5 times more likely to get bladder cancer if exposed to carcinogens. Certain polymorphisms increase the risk of mesothelioma 7.8 times. It will be possible in the future to rapidly and cheaply test individuals for polymorphisms and genotypes.



Thick walled mesothelioma of pleura with haemorrhagic cavitation in a former insulation worker

Sites of cancers

Carcinogens are organotropic. In the United Kingdom the most commonly affected sites are the lung (mesothelium) (75%), bladder (10%), and skin (1%). Other sites affected are the haemopoietic system, nasal cavities, larynx, and liver.

Natural course of cancers

Occupationally related cancers are characterised by a long latent period (that is, the time between first exposure to the causative agent and presentation of the tumour). This latency is not usually less than 10 to 15 years and can be much longer (40-50 years in the case of some asbestos related mesotheliomas): presentation can therefore be in retirement rather than while still at work. However, susceptibility to occupational carcinogens is greater when the exposure occurs at younger ages. An occupationally related tumour does not differ substantially, either pathologically or clinically, from its "naturally occurring" counterpart.

Diagnosis of work related cancer

- Detailed lifelong occupational history
- Comparison with a checklist of recognised causal associations
- Confirmation of requisite exposure
- Search for additional clues: shift to a younger age, presence of signal tumours, other cases and "clusters," long latency, absence of anticipated aetiologies, unusual histology or site

Recognition and diagnosis

For a group of workers, occupational cancer is evidenced by a clear excess of cancers over what would normally be expected. Some common malignancies that can be work related also have a well recognised and predominant aetiology related to other agents, diet, or lifestyle (for example, lung cancer from smoking). There are, however, some features that may help to distinguish occupational cancers from those not related to work.

History taking

Taking a patient's occupational history is paramount. It should be defined in detail and sequentially. For example, a holiday job in a factory that lasted only a few months could easily be overlooked, but it may have included delagging a boiler or handling sacks of asbestos waste.

Signal tumours

Several uncommon cancers are associated with particular occupations. Thus, an angiosarcoma of the liver may indicate past exposure to vinyl chloride monomer in the production of polyvinyl chloride, although there have been no cases in workers exposed since 1969. A worldwide registry of all exposed workers exists.

Rubber workers in mill room

Age
A younger age at presentation with cancer may suggest an occupational influence. For example, a tumour of the urothelial tract presenting in anyone under the age of 50 years should always arouse suspicion.

Patients' information
Patients may speak of a "cluster" of cancer cases at work, or they may have worked in an industry or job for which a warning leaflet has been issued.

Cystoscopic view of papillary carcinoma of the bladder in a 47 year old rubber worker

Prevention

Primary prevention seeks to prevent the onset of a disease. Secondary prevention aims to halt the progression of a disease once it is established. Tertiary prevention is concerned with the rehabilitation of people with an established disease to minimise residual disabilities and complications or improve the quality of life if the disease itself cannot be cured.

Levels of prevention

	Stages			Outcomes		
	Health	**Asymptomatic**	**Symptomatic**	**Disability**	**Recovery**	**Death**
Intervention strategies	Health education, immunisation, environmental measures and social policy	Presymptomatic screening	Early diagnosis and prompt effective treatment	Rehabilitation		
Level of prevention	Primary	Secondary		Tertiary		

Adapted from Donaldson and Donaldson, 1999.

Primary prevention of occupationally related cancers depends essentially on educating employers and employees; firstly about recognising that there is a risk, and then about the practical steps that can be taken to eliminate or reduce exposure and to protect workers. Modern risk based legislation now directs these educational and practical measures.

Secondary prevention
Screening procedures may enable earlier diagnosis, but there is little evidence to suggest that most screening makes a difference to outcome. Screening is of proven benefit in cutaneous cancers of occupational origin, mainly because of the excellent prognosis afforded by treatment. Routine skin inspections should be initiated where there is exposure to known skin carcinogens. Routine urine cytology has been carried out in many industries where there has been previous exposure to known carcinogens. It is possibly of benefit but this has not been proven. *β*-Naphthylamine was withdrawn from use by 1950, but many former workers continue to participate in urine cytology screening programmes. Once commenced, surveillance should be lifelong. In the United Kingdom it is recommended workers exposed to 4,4-methylene-bis-(2-chloroaniline) (MbOCA) should have their urinary levels of MbOCA and its *N*-acetyl metabolites checked, but periodic urine cytology for those exposed remains controversial. Screening for lung and liver cancer is not of benefit.

Action for primary prevention of occupational cancers

- Recognition of presence of hazards and risks
- Education of management and workforce
- Elimination of exposure by substitution and automation
- Reduction of exposure by engineering controls (such as local exhaust ventilation and enclosure, changes in handling, and altering physical form in processing)
- Monitoring of exposure and maintaining plant
- Protection of workers with personal protective equipment
- Limiting access
- Provision of adequate facilities for showering, washing, and changing
- Legislative provisions

Criteria for screening

- Is the condition an important health problem?
- Is there a recognisable early stage?
- Is treatment more beneficial at an early stage than at a later stage?
- Is there a suitable test?
- Is the test acceptable to the population?
- Are there adequate facilities for diagnosis and treatment?
- What are the costs and benefits?
- Which subgroups should be screened?
- How often should screening take place?

Legislation and statutory compensation

Essential legislative provisions in the United Kingdom and the European Union are comprehensive. Ten types of cancer are prescribed diseases and are eligible for industrial injuries benefit. Some cancers are reportable under the Reporting of Injuries, Diseases, and Dangerous Occurrence Regulations 1995 (RIDDOR), although many occur in those who have retired. Most occupational cancers recorded or eligible for benefit are mesotheliomas. In 2000, 652 people received benefits in the United Kingdom for mesothelioma, which is less than half the number of deaths recorded as caused by this disease (1595 deaths in 1999). About 80 people with other occupational cancers receive benefits each year, these being split between bladder cancer and asbestos related lung cancer. Bladder cancers have slowly increased over the past decade, whereas lung cancers have decreased. The figure for asbestos related lung cancers substantially under-represents the true number.

Benefits and disadvantages of screening

Benefits
- Improved prognosis for some cases detected by screening
- Less radical treatment for some early cases
- Reassurance for those with negative test results

Disadvantages
- Longer morbidity for cases whose prognosis is unaltered
- Over treatment of questionable abnormalities
- False reassurance for those with false negative results
- Anxiety and sometimes morbidity for those with false positive results
- Unnecessary medical intervention for those with false positive results
- Hazard of screening test
- Resource costs: diversion of scarce resources to screening programme

Specific carcinogens

Metals and metalliferous compounds

Arsenic, beryllium, cadmium, chromium(VI), nickel, and iron are considered to be proven human carcinogens, either as the metal itself or as a derivative. The risk from iron is related only to mining the base ore and is caused by coincidental exposure to radon gas. In foundries, where there is concomitant exposure to several agents in a complex mix of emanating fume, the responsible agents are not clearly defined.

With all the metallic carcinogens, the lung is the main target organ, but other potential sites are shown in the table. The main occupational exposures occur in the mining, smelting, founding, and refining of these metals, and less commonly in secondary industrial use.

Main legislative provisions in the United Kingdom

- Control of Substances Hazardous to Health (COSHH) Regulations 2002 and associated approved code of practice on the Control of Carcinogens
- European Commission Carcinogens Directive (90/934/EEC)
- Chemical Agents Directive (98/24/EC)
- Chemicals (Hazard Information and Packaging) Regulations 1999 (CHIP)
- Ionising Radiations Regulations (1999)
- Control of Asbestos at Work Regulations (1998)
- Reporting of Injuries, Diseases, and Dangerous Occurrences Regulations (RIDDOR) 1995

Aromatic amines

Aromatic amines are among the best known and most studied of chemical carcinogens. The bladder is the main target organ, but any site on the urothelial tract composed of transitional cell epithelium can be affected—that is, from the renal pelvis to the prostatic urethra. Tumours of the upper urothelial tract (renal pelvis or ureter) are very rare, and a cluster of these signal tumours usually heralds an underlying risk of occupational cancer. The carcinogenic potential of aromatic amines lies not in the parent compound but in a metabolite formed in the liver and excreted through the urinary system.

The occupations classically associated with risk from these chemicals were in the industries' manufacturing chemicals and dyestuffs. In the early 1950s an investigation of bladder cancers in workers in British chemical industries showed that individuals exposed to benzidine and 2-napthylamine had a 30 times greater risk of developing bladder cancer than the general population. Occupational bladder cancer became a prescribed disease in 1953.

Antioxidants contaminated with β-Naphthylamine were used in the rubber and cable making industries until the end of 1949 (when they were universally withdrawn), and they caused an excess of bladder cancer. The level of contamination was only about 0.25%, yet it almost doubled the risk for the workforce so exposed. People who started work in the rubber industry after 1951 seem to have no excess risk.

There is now increasing evidence that some polycyclic aromatic hydrocarbons can act as urinary tract carcinogens.

Metalliferous carcinogens

Agent	Target organ
Arsenic	Lung and skin
Beryllium	Lung
Cadmium	Lung, prostate gland
Chromium (hexavalent)	Lung
Nickel	Lung, nasal sinuses
Iron in:	
Haematite mining (radon)	Lung
Iron and steel founding	Lung, digestive tract

Aromatic amine carcinogens

Proved
- 4-Aminobiphenyl (xenylamine) and its nitro derivatives
- β-Naphthylamine
- Benzidine
- Auramine and magenta

Probable
- Polycyclic aromatic hydrocarbons

Possible
- The hardener MbOCA (4,4-methylene-bis-(2-chloroaniline))

This is reflected in excesses seen in aluminium refiners and in painters exposed to solvents.

Occupations causally associated with urothelial tract cancers

- Dyestuffs and pigment manufacture
- Rubber workers (in tyre, tube, and cable making before 1950)
- Textile dyeing and printing
- Manufacture of some chemicals (such as 4,4-methylene-bis-(2-chloroaniline) (MbOCA))
- Gas workers (in old vertical retort houses)
- Laboratory and testing work (using chromogens)
- Rodent controllers (formally using (alpha)-naphthylthiourea)
- Painters
- Leather workers
- Manufacture of patent fuel (such as coke) and firelighters
- Tar and pitch workers (roofing and road maintenance)
- Aluminium refining

Asbestos

Few natural materials used in industry have been the subject of more epidemiological and pathological research than the fibrous mineral, asbestos. Lung cancer because of asbestos was first reported in the 1930s and its association was confirmed in the 1950s. In 1960, Wagner and colleagues reported 33 cases of the "rare" tumour mesothelioma in workers exposed to asbestos in South Africa.

In asbestos workers who have developed asbestosis the risk of lung cancer is increased at least five times. For chrysotile there is a linear relationship between exposure and risk of lung cancer. Each additional fibre exposure (every ml a year) is equivalent to a 1% increase in the standardised mortality ratio.

Between 0.6% and 40% of lung cancers have been attributed to occupation, depending on place and time. Chlormethylesters, used in ion exchange resins, increase the risk of lung cancer 20 times and have a short latent period of 10-15 years. The type of cancer is small cell, also caused by uranium and beryllium (which also causes adenocarcinoma). Painters have a 30-100% increase in lung cancer. This may be caused by heavy metal salts or chromates, organic solvents, or exposure to silica and asbestos.

Asbestos related cancers

- Lung
- Malignant mesothelioma—most commonly of pleura, occasionally peritoneal, and rarely of pericardium
- Larynx
- Possibly gastrointestinal tract

Over 40% of people with asbestosis die of lung cancer, and 10% die of mesothelioma. Mesotheliomas, which are predominantly of the pleura (ratio of 8:1 with peritoneum), have usually been growing for 10 to 12 years before becoming clinically evident. This latency can be very long—often 30 years and sometimes up to 50 years. However, median survival from the time of initial diagnosis is usually short—three to 12 months.

The amphibole fibres in crocidolite (blue asbestos) and amosite (brown asbestos) carry the greatest risk of causing mesothelioma, but the serpentine fibres in chrysotile (white asbestos) can also do so, especially if they contain tremolite. In about 90% of patients with mesothelioma, close questioning will usually show some earlier exposure to asbestos. The possible risk to neighbourhoods outside asbestos factories from discharged asbestos dust or contaminated clothing brought home should not be forgotten.

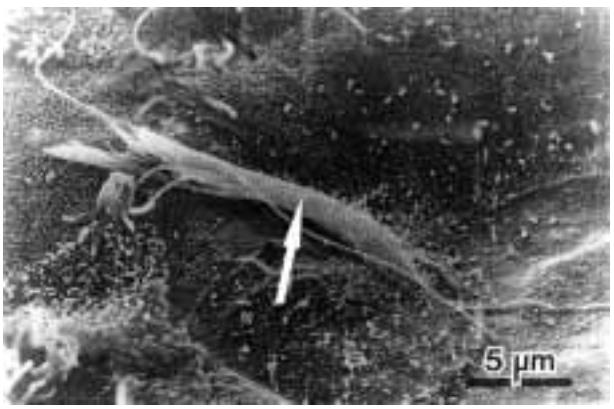
The annual number of deaths from mesothelioma has increased rapidly from 153 in 1968 to 1595 in 1999. The latest projections suggest that male deaths from mesothelioma may peak in about 2011, at about 1700 deaths every year. Occupations with the highest risk of mesothelioma for men include metal plate workers (including shipyard workers), vehicle body builders (including rail vehicles), plumbers and gas fitters, carpenters, and electricians.

Smoking and asbestos

Asbestos	Tobacco	Lung cancer rate per 100 000
−	−	11
+	−	58
−	+	123
+	+	590

Smoking with concomitant exposure to asbestos greatly increases the risk of developing lung cancer: compared with non-smokers not exposed to asbestos, a smoker exposed to asbestos has a 75-100 times greater risk if exposure was sufficient to cause asbestosis, otherwise the risk is about 30-50 times higher. This multiplicative theory on effects of asbestos exposure and smoking, however, has recently been disputed

Blue asbestos



White asbestos

Asbestos legislation

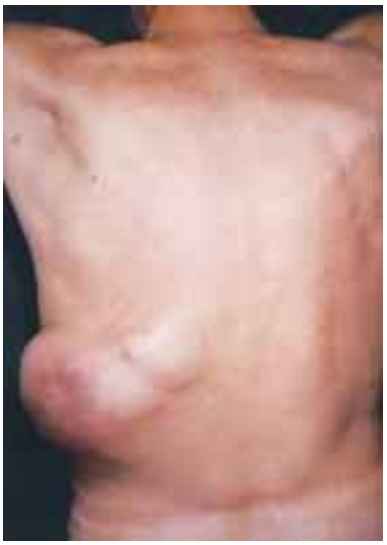
Asbestos is controlled in the United Kingdom by three complementary sets of regulations:

- **The Asbestos (Licensing) Regulations** (amended 1998) require work with the most dangerous types of asbestos (coating, lagging, and asbestos insulating board) to be carried out only by contractors who have a licence issued by the Health and Safety Executive
- **The Control of Asbestos at Work Regulations** (amended 1998) lay down the practices that must be followed for all work with asbestos, including that which requires a licence. Employers must prevent the exposure of employees to asbestos or, where this is not reasonably practicable, reduce exposure to a level that is as low as possible
- **The Asbestos (Prohibitions) Regulations** (amended 1999) prohibit the importation into the United Kingdom, and the supply and use within Great Britain, of amphibole asbestos—crocidolite (blue) asbestos and amosite (brown) asbestos—and, since 1999, of chrysotile (white) asbestos

The supply and fitting of vehicle brake linings containing asbestos is prohibited in **The Road Vehicles (Brake Linings Safety) Regulations 1999**, and the European Union has amended the **Marketing and Use Directive (76/769/EEC)**, which prohibits the marketing and use of chrysotile asbestos throughout the EU after 1 January 2005, with one derogation for diaphragms for the chlor-alkali process

Forthcoming legislation will require employers to manage the risk from asbestos in non-domestic premises

The latest amendments to the Control of Asbestos at Work Regulations 1987 (which came into force in 1999) target workers who come across asbestos accidentally, such as electricians, plumbers, other maintenance workers, and demolition workers. The Amendment Regulations also tighten the law on control of exposure to asbestos by lowering the action level and the control limit for chrysotile



Mesothelioma extending through needle biopsy tract

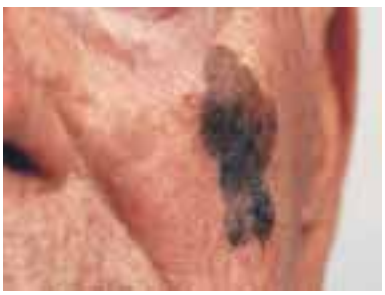
Occupations involving exposure to asbestos

- Manufacture of asbestos products
- Thermal and fire insulation (lagging, delagging)
- Construction and demolition work
- Shipbuilding and repair (welders, metal plate workers)
- Building maintenance and repair
- Manufacture of gas masks (in second world war)
- Plumbers and gasfitters
- Vehicle body builders
- Electricians, carpenters, and upholsterers
- Armed forces (historical)

Tyndall beam photography showing asbestos fibres released by handling of asbestos boards (left), emphasising the need for proper protection when dealing with asbestos (right)

Ultraviolet radiation

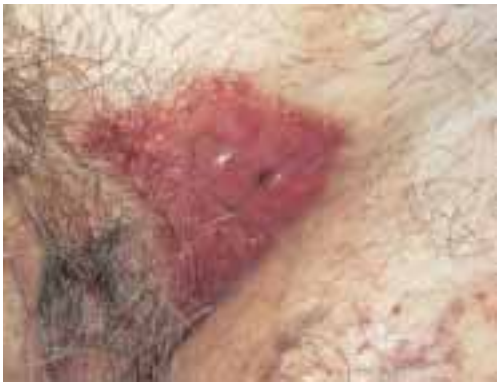
Ultraviolet radiation from exposure to sunlight causes both melanotic and non-melanotic skin cancers (basal cell and squamous cell carcinomas), but an excess of skin cancers in outdoor workers is seen only in those with fair skin. Initial presentation may be that of solar keratoses or a premalignant state. Immunosuppression can increase the risk; other possible additive factors are trauma, heat, and chronic irritation or infection.



Premalignant melanosis (lentigo maligna) in a man who retired after a lifetime of working outdoors

Mineral oils

The classic epithelioma of the scrotum or groin caused by contact with mineral oil is rarely seen today, but these tumours can appear at other sites (such as arms and hands) if contamination with oil persists.



Epithelioma of groin caused by past exposure to mineral oil

Other occupational carcinogens

Ionising radiation is a carcinogen at low doses (0.2 gray or a dose rate of 0.05 mSv per min). Cancer or hereditary defects are known as stochastic effects and can only be minimised. Cataract, sterility, and skin disorders are deterministic effects and can be prevented by keeping exposure below threshold. The recommended dose limit is 20 mSv a year averaged over five years for occupational exposures and 1 mSv for the public.

Proven human carcinogens

Miscellaneous proved human carcinogens	Site of cancer
Aluminium production	Lung, bladder, skin
Benzene in petroleum associated industries	Haemopoietic
Bis-(chloromethyl)-ether in production of ion exchange resin	Lung
Benzene and leather dust in boot and shoe making and repair	Haemopoietic, nasal
Polycyclic aromatic hydrocarbons and aromatic amines in coal gasification and coke production	Lung, bladder, skin
Coal tars and pitch in roofing and road maintenance	Lung, bladder, skin
Ethylene oxide as medical steriliser and chemical intermediary	Lung, bladder, skin
Formaldehyde and hardwood dust in furniture and cabinet making	Nasal, paranasal
Isopropyl alcohol manufacture	Nasal, paranasal
Mineral and shale oils in engineering and metal machining, past exposure to mule spinning in cotton industry and jute processing	Skin, scrotum
Solvents and pigments in painting and decorating	Lung, bladder, stomach, oesophagus
Mists of strong inorganic acid (sulphuric acid) in acid pickling and soap making	Nasal, larynx
Radon in underground mines	Lung
Soots from chimney sweeping and flue maintenance	Lung, skin
Antineoplastic agents	Bladder, haemopoietic

Frieben documented the first case of skin cancer on the hand of an *x* ray tube factory worker in 1902. Cancer risk estimates on nuclear workers are still not conclusive, and the Gardener hypothesis that the children of radiation workers have an increased risk of leukaemia has not been supported. However, incidence may be increased in emergency workers. The epidemiological evidence from studies concerning airline crew who may receive the equivalent of 100 mSv over a 20 year period from cosmic radiation are inconclusive. No excess cancer has been reported among therapeutic or diagnostic radiologists.

All studies on electromagnetic radiation show inconsistencies and seldom indicate dose-response trends. This may mean that there is no association between electromagnetic fields and cancer, or that there is a risk but studies have not been able to show it. Particular aspects studied so far have been leukaemia, brain cancer, male breast cancer, electrical workers, and welders, but a broader research hypothesis is needed.

Studies of manmade mineral fibres have looked only at small exposures in terms of fibres and years of exposure. An increased risk of lung cancer was found in rock wool workers but it was not possible to conclude that it was caused by manmade mineral fibres. No risk was found in glass wool or glass filament workers. Five deaths from mesothelioma have

Further reading

- McDonald C, ed. Occupational cancer. In: *Epidemiology of work-related diseases*, 2nd ed. London: BMJ Books, 2000. *Excellent review of occupational cancer epidemiology and the evidence for it*
- HSE. *Health and Safety Statistics 2000/01 Part 2: Occupational ill-health statistics*. Sudbury: HSE Books, 2001. *Comprehensive summary of UK occupational cancer statistics*
- Venitt S, Harrington JM, Boffetta P, Saracci R. Occupational cancer. In: Baxter BJ, Adams PH, Aw TC, Cockcroft A, Harrington JM, eds. *Hunter's diseases of occupations*, 9th ed. London: Edward Arnold, 2000:623-88. *Definitive in depth text on occupational cancer*
- IARC. Monographs on the evaluation of carcinogenic risks to humans. Volumes 1-79. Lyons: International Agency for Research on Cancer, 1972-2001. http://monographs.iarc.fr/ *Comprehensive highly detailed studies of chemicals and processes thought to cause cancer*
- Wilson JMJ, Jungner G. Principles and practice of screening for disease. WHO Public Health Paper 1968;34
- Donaldson LJ, Donaldson RJ. *The promotion of health in essential public health*, 2nd ed. Newbury: Petroc Press, 1999. *Public health textbook*
- Peckham MJ. *Oxford textbook of oncology*. Oxford: Oxford University Press, 2002. *Up to date oncology bible*

been found in various cohorts, but at least three of these workers may have previously been exposed to asbestos.

There is sufficient evidence for the carcinogenicity of inhaled crystalline silica in the form of quartz or cristobalite. Studies show the Bradford-Hill criteria of temporality, consistency, exposure-response gradients, and convergence with experimental and clinical evidence. Measures to prevent silicosis are likely to reduce lung cancer risk.

The box showing criteria for screening is adapted from Wilson JM, Jungner G. *Principles and practice of screening for disease.* WHO Public Health Paper 1968;34. The table showing levels of prevention is adapted from Donaldson LJ and Donnaldson RJ. *The promotion of health in essential public health,* 2nd ed. Newbury: Petroc Press, 1999.

16 Occupational dermatoses

Ian R White

Skin disorders are among the most often encountered problems in the occupational health setting, and although there are many dermatoses that have occupational relevance, the overwhelming majority are dermatitic. In the United Kingdom, in the period 1998-2000, of the estimated 4540 workers each year with work related skin disease seen by specialist physicians, about 80% were diagnosed as having contact dermatitis. Occupations considered to be at greatest risk are hairdressers and barbers, grinding machine setters and operators, galvanisers, rubber process operatives, and printers.

Contact dermatitis

In current terminology the term "dermatitis" is used synonymously with "eczema" to describe inflammatory reactions in the skin with a spectrum of clinical and histopathological characteristics.

A dermatitis may be entirely endogenous (constitutional) or entirely exogenous (contact). The latter consists of irritant and allergic contact reactions. Commonly, a dermatitis has a multifactoral aetiology and may be aggravated by the presence of pathogens (for example, *Staphylococcus aureus*). Assessment of the relative importance (contribution) of the possible factors can be difficult and subjective. Atopic hand dermatitis and vesicular hand dermatitis are examples of endogenous conditions.

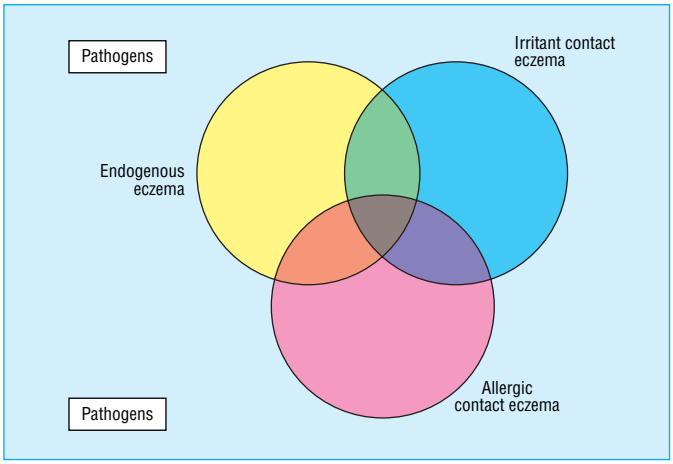
An occupational dermatitis is one where the inflammatory reaction is caused entirely by occupational contact factors or where such agents contribute to the reaction on a compromised skin—that is, they are partially responsible.

In most cases, an occupationally related dermatitis will affect the hands alone or there may be spread onto the forearms. Occasionally, the face may be the prime site of dermatitis (for example, with airborne agents); other sites may be affected.

Irritant contact dermatitis

This is initiated by direct chemical or physical damage to the skin. All individuals are susceptible to the development of an irritant contact dermatitis if exposure to the irritant (toxic) agent or agents is sufficient. It occurs particularly where the stratum corneum is thinnest. Hence, it is often seen in the finger webs and on the backs of the hands, rather than on the palms. Irritant contact dermatitis is of two principal types: acute and chronic. The former is caused by exposure to an agent(s) causing early impairment in stratum corneum function followed by an inflammatory reaction. The latter is caused by repeated exposure to the same or different factors, resulting in "cumulative" damage until an inflammatory reaction ensues and persists for a prolonged period, even after further exposure is stopped. Those with a history of atopic eczema, and especially atopic hand eczema, are at particular risk of developing chronic irritant contact dermatitis. Chronic irritant contact dermatitis is particularly seen in "wet work."

Wet work, solvents, detergents, soluble coolants, vegetable juices, wet cement, low relative humidity, and occlusive gloves are all examples of common irritants.



Dermatitis may be endogenous or exogenous, or a combination of these, and may be aggravated by pathogens

Indications for occupational cause of dermatitis

- A dermatitis first occurred while employed
- There is a history of aggravation by work
- There may be, at least initially, improvement (or clearance) when not at work
- There is exposure to irritant factors or potential allergens
- Work is in an "at risk" occupation

Irritant contact dermatitis

Acute

- Severity of reaction depends on "dose" of irritant agent
- "Chapping" can be considered a minor form, with a "chemical burn" (for example, cement burn) being an extreme event.
- Intermediate eczematous reactions are common; minor reactions are very common
- May occur on the face—for example, low humidity occupational dermatosis, airborne irritant vapours
- Once the irritant factor(s) has been removed, resolution is usually spontaneous without important sequelae

Chronic

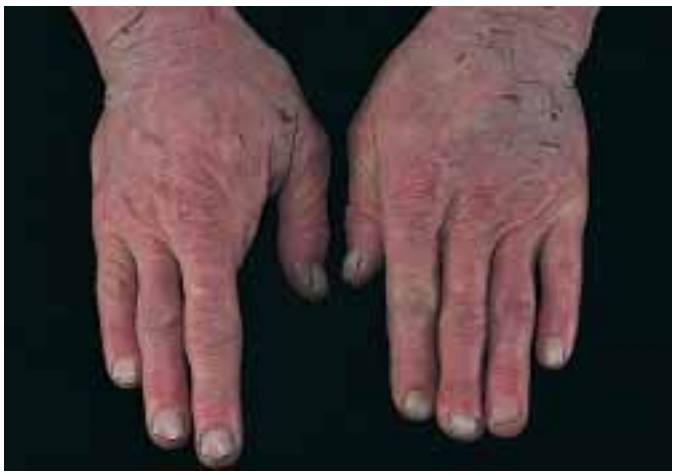
- A persistent dermatitis and the most common cause of continued disability from occupational skin disease
- Problem continues for long periods even with avoidance of aggravating factors
- Re-exposure to even minor irritant factors may cause a rapid flare
- Even after apparent healing there may be an indefinitely increased susceptibility to recurrence of a dermatitis after irritant exposure

Examples of common occupational allergens

- Biocides—for example, formaldehyde, methyldibromo-glutaronitrile, methylchloroisothiazolinone
- Hairdressing chemicals—for example, *p*-phenylenediamine
- Chromate (leather, cement)
- Rubber accelerating chemicals—for example, thiurams, carbamates, mercaptobenzothiazole
- Epoxy resin monomers (plastics manufacturing, electrical manufacture)
- Plant allergens—for example, sesquiterpene lactones (horticulture)

Listing of ingredients

All cosmetic (skin care) products in Europe have full ingredient listing on the product packaging, with uniform nomenclature. Skin cleansers, barrier creams, and after work creams are legally cosmetic products. Labelling permits tracing of sources of exposure to allergens. However, there is a lack of meaningful ingredient labelling on other types of consumer and industrial products.



Lichenified and fissured eczema on hands of bricklayer resulting from chronic irritant dermatitis. Patch tests were negative, and he was not sensitive to chromate

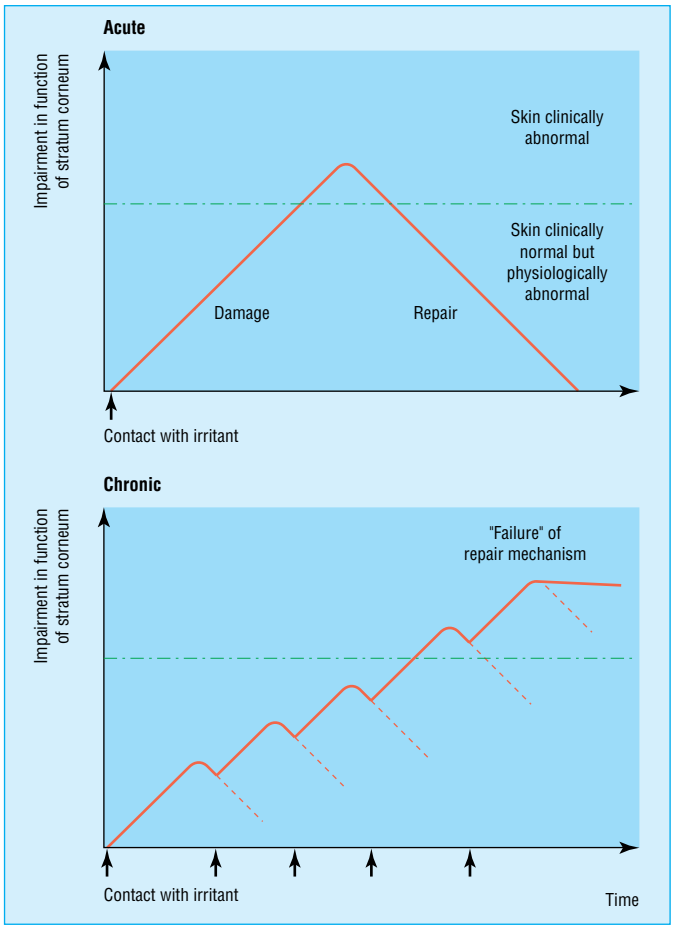
Allergic contact dermatitis

This is a manifestation of a type IV (delayed) hypersensitivity reaction. An allergic contact dermatitis will develop at the site of skin contact with the allergen, but secondary spread may occur. Contaminated hands may spread the allergen to "non-exposed" sites. Trivial or occult contact with an allergen may result in the persistence of a dermatitis; some allergens are "ubiquitous."

Presentation of an allergic reaction has two phases: induction and elicitation. Even with potent experimental allergens there is a minimum period of about 10 days from the first exposure to the immunological acquisition of hypersensitivity. The probability of developing hypersensitivity depends on the sensitising capacity of the chemical and exposure to it. Exposure is assessed in terms of dose every unit area applied to the skin. Most potential allergens on the consumer and industrial market have a low intrinsic allergenic potential, but there are important exceptions, including some biocides (preservatives). Contact allergens tend to be low molecular weight (< 600) and capable of forming covalent bonds with carrier proteins in the skin. It is not possible to determine an individual's susceptibility to the development of contact allergy. Hypersensitivity is specific to a particular molecule or to molecules bearing similar allergenic sites. Although hypersensitivity may be lost over a long time, once acquired it should be considered to last indefinitely.

Management of occupational dermatitis

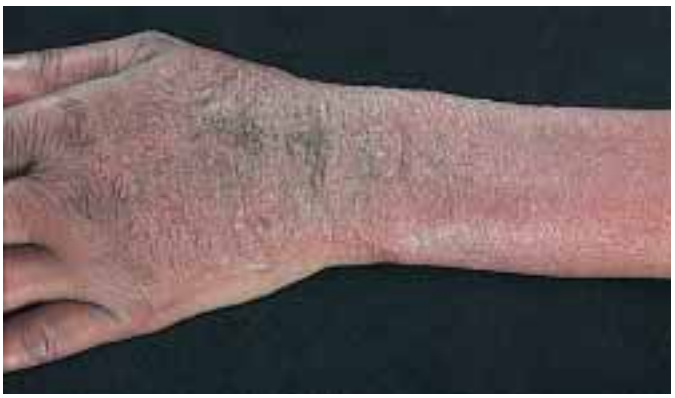
- An understanding of the patient's job is essential. A job title is not sufficient for this understanding: the question to be asked is not "what do you do?" but "what do you actually do and how do you do it?" The title "engineer" carries a multiple of descriptions ranging from the desk bound professional to the lathe worker exposed to soluble coolants.



Development of acute (top) and chronic (bottom) irritant contact dermatitis

Relevance of contact allergens

- **C**urrent: exposure is causing dermatitis or is aggravating it
- **O**ld: past history of exposure but no current exposure
- **D**on't know: unknown whether there is current exposure to the allergen or whether exposure is important to the dermatitis
- **E**xposure: must have occurred but no history of it



Sheeted eczema over the dorsal aspect of the hand and up to the forearm, resulting from allergic contact dermatitis to a carbamate accelerator in protective rubber gloves

From the job description, it will be possible to estimate sources of excessive contact with potentially irritant contact factors or allergens. The provision of material safety data sheets may be helpful in this evaluation, although the information that they contain is often superficial, generic, and is that required for regulatory requirements. A site visit to watch the worker working may be necessary.

- The history and anatomical distribution of the dermatitis may provide clues as to the aetiology.
- Irritant contact dermatitis may occur as "epidemics" in a workplace if hygiene has failed. Allergic contact dermatitis is usually sporadic in a workplace.
- The evaluation of irritant factors is always subjective. Evaluation of allergic contact factors is objective and provided only by diagnostic patch test investigations. Properly performed, patch tests will show the presence or absence of relevant allergens.
- Patch testing is the only method for the objective evaluation of a dermatitis; however, there are major pitfalls in the use of this essential tool, so adequate training and experience is necessary if it is to be used properly. The ability to assess relevance of allergens is central.
- A competent assessment requires all of the above followed by recommendations on reducing or stopping exposure to the offending agent(s) and similar ones.
- The diagnosis of an occupational dermatitis should describe thoroughly the nature of the condition with due regard to any endogenous or aggravating factors. A general practitioner's entry in a medical record of "Works in a factory, contact dermatitis. 2/52" is inadequate as a description of an important disease process, and it can have profound implications on the patient's concept of his problem and employment.
- Delays in diagnosis resulting in continued exposure to relevant irritants or allergens can adversely affect the prognosis.
- Early referral to an appropriate dermatology department is necessary for a comprehensive assessment of a suspected occupational dermatitis; improper assessment can have devastating effects on future employment prospects for the individual, with important medicolegal implications. If in doubt, the patient should be referred.

The primary prevention of occupational dermatitis is aimed at providing appropriate information and protection

- Awareness by employer and employee of the potential risks of exposure
- Education on the necessity of good occupational hygiene precautions
- Adequate provision of suitable and effective means of reducing exposure
- Awareness of the limitations of personal protection devices

It is not possible to be definitive about aetiology from the distribution and morphology of a dermatitis on the hands. For example, vesicular hand dermatitis with a "classical" endogenous distribution may be mimicked by an allergic contact dermatitis to isothiazolinone biocides or chromate sensitivity. ***It is a major error to rely on patterns of hand dermatitis in making a diagnosis***

Patch testing

- Properly performed requires expertise, time, and proper facilities
- Difficult to undertake adequately in the workplace. There are no short cuts
- Primarily a hospital based procedure
- Should be performed only by those with appropriate training who can prescribe an appropriately comprehensive screen, know what not to test, know what to dilute for testing, can competently read the reactions, and can give authoritative advice after interpretation of the reactions
- Anyone can patch test; few do it well. If you don't know how to do it, don't do it

Rubber latex protein sensitivity

Of continuing concern is the immediate type 1 hypersensitivity reaction to proteins present in natural rubber latex used to make gloves and other items. This should be differentiated from irritant contact dermatitis and allergic contact dermatitis, which can also be attributed to chemical agents used in latex products, particularly gloves.

Prevalence and incidence of sensitivity to rubber latex proteins remain unquantified. The prevalence in the general population is thought to be less than 1%, but is likely to be higher within certain risk groups. A prevalence of 2.8-17% has been reported in healthcare workers, and in other occupations where workers are regularly exposed to rubber latex (hairdressers, greenhouse workers, housekeeping staff, and glove factory workers) the frequency of allergy has been reported as ranging from 5-11%.

At particular risk are people with spina bifida (prevalence reported to be 18-65%), atopic individuals, and individuals with certain food allergies (for example, to avocado, chestnut or banana, and kiwi fruit).

Rubber latex protein sensitivity can result in reactions including urticaria, rhinitis, and asthma. The latter is more

Immediate contact reaction to latex proteins in examination gloves. Type 1 hypersensitivity reactions to latex proteins are of growing concern

common when starch powdered gloves have been used. In the healthcare setting, gloves made from rubber latex are likely to be the main cause of sensitisation in staff, as well as the main cause of symptoms in those who are allergic. Therefore, prevention includes clear policy regarding the type of glove used. If latex gloves are to be used, powdered rubber latex gloves should not be used, and extractable protein levels in latex gloves must be as low as possible, as should the level of allergenic protein residues. Staff with known sensitivity should be provided with non-latex alternatives.

A definitive demonstration of hypersensitivity can be made by skin prick testing with the water soluble proteins. Commercial preparations are available; the proteins can also be eluted from a suspect rubber item. Radioallergosorbent tests are less sensitive.

Other occupational dermatoses

- Contact urticaria—type I hypersensitivity reaction—for example, natural rubber latex protein
- Chloracne (halogenacne)—acneiform eruption caused by certain halogenated aromatic hydrocarbons; a symptom of systemic absorption
- Oil folliculitis (oil acne)—irritant effect of neat petroleum oils localised to hair follicles
- Depigmentation (leukoderma)—caused by hydroquinone and phenol derivatives
- Hyperpigmentation—caused by mineral oils, halogenated hydrocarbons, photodynamic actions of psoralenes and tar products
- Skin cancer (see chapter 15)
- Skin infections (see chapter 14)

Further reading

- Beach J. The problem with material safety data sheets. *Occup Med* 2002;52:67-8
- Wakelin SH, White IR. Natural rubber latex allergy. *Clin Exp Dermatol* 1999;24:245-8
- Smith HR, Armstrong DK, Wakelin SH, Rycroft RJ, White IR, McFadden JP. Descriptive epidemiology of hand dermatitis at the St John's contact dermatitis clinic 1983-97. *Br J Dermatol* 2000;142:284-7
- Robinson MK, Gerberick GF, Ryan CA, McNamee P, White IR, Basketter DA. The importance of exposure estimation in the assessment of skin sensitization risk. *Contact Dermatitis* 2000;42:251-9
- Rycroft RJG, Menne T, Frosch PJ, Lepoittevin J-P, eds. *Textbook of contact dermatitis*, 3rd ed. Berlin: Springer-Verlag, 2001
- Kanerva L, Elsner P, Wahlberg JE, Maibach HI, eds. *Handbook of Occupational Dermatology*. Springer, 2000
- The monthly journal *Contact Dermatitis* (Blackwell Science Ltd) publishes papers and case reports on matters relevant to occupational dermatology

17 Work, genetics, and reproduction

Nicola Cherry

The sequencing of the human genome, and the intense interest that accompanied this achievement, again raised issues surrounding the interaction of genetic and environmental exposures in causing disease. Even where exposures at work and in the environment have been clearly shown to be related to specific pathologies, and preventive measures initiated, the question of genetic susceptibility remains: why do only some of those exposed develop the disease? Understanding of such susceptibility will seldom exclude workers, but it may improve understanding of the mechanisms by which disease occurs, and suggest approaches to prevention or treatment.

There is also concern that occupational or environmental exposures may affect subsequent generations through changes to stem cells; by this means an infant born to such a parent may be at greater risk of disease even if exposure of the parent ceased long before the child was conceived. Evidence from human studies of occupational exposure affecting the genetic blueprint of the next generation is sparse and controversial, but pressing questions remain about whether such exposures can cause infertility, affect the outcome of pregnancy, or influence the development of the infant in later life.

In all these areas—genetic susceptibility, genetic alteration, and reproductive health—exposures in the working population may be of particular concern because exposures tend to be greater than in the general population, a large proportion of those exposed are of reproductive age, and any exposure effects that are found are, in principle, preventable. Environmental exposures to the general public (including the very young, and pregnant or nursing mothers) through contaminants in food, water, and air may also be suspected of affecting reproductive health. For example, even trace amounts of chemicals affecting the endocrine system of pregnant women may be responsible for the increased rate of testicular cancer seen in many societies.

Work and genetics

Why should occupational health professionals be concerned with the genetic make up of people in the workforce or who seek to join it? Firstly, in some, genetic inheritance, even in the absence of a specific occupational exposure, will lead to disease that will put at risk themselves, their fellow workers, or the general public. For example, a worker genetically programmed to develop Huntington's disease, if employed as a driver of a high speed train, may put the public at risk in the early stages of the disease before a diagnosis can be made that permits redeployment or retirement on medical grounds. Secondly, some genetic conditions render a person unable to tolerate work environments that can be tolerated by other workers. For example, deep sea diving may induce a crisis in a worker carrying the gene for sickle cell disease and, as a result, the worker and others may be put at peril. Thirdly, a particular genetic variant (or polymorphism) or a combination of variants may carry a risk of ill health if a worker is exposed to a chemical that is detoxified by the enzyme produced by the gene. The case of slow acetylators is a well known example. Where such a disease is a serious threat to quality of life or life

Genetic testing

- Can identify a predisposition to illness—for example, thalassaemia, Huntington's disease, sickle cell disease
- Could be used for genetic monitoring—for example, exposure to radiation or polycyclic aromatic hydrocarbons
- Has been used for estimation of fitness to work (exclusion or protection)—for example, exclusion of those with sickle cell trait from flying, diving and compressed air work, and exclusion of those with glucose-6-phosphate dehydrogenase deficiency from work involving naphthalene or trinitrotoluene, and cultivation or processing of broad beans
- Has been proposed as a way to:
  - predict the likelihood of common diseases (diabetes, schizophrenia—for example) that might raise sickness absence rates or medical costs to employers
  - identify resistant individuals (rapid acetylators—for example) who could, in theory, be exposed without harm to higher concentrations of toxic chemicals

Genetic testing or screening in an occupational context is clearly beset with problems of ethics, effectiveness, and practicality

Genetic information

- Is a unique identifier
- Can be done on a small sample
- Can be done covertly, without consent
- Can be used for prediction
- Is of interest to employers, insurance companies, and relatives
- Has potential commercial value (patents)
- Can outlast the source
- Can define susceptible groups
- Can be used for purposes other than those for which it was collected

Statement of the Nuffield Council on Bioethics 1993

Genetic screening of employees for increased occupational risks ought only to be contemplated where:

(i) Strong evidence exists of a clear connection between the working environment and the development of the condition for which genetic testing can be conducted
(ii) The condition in question is one that seriously endangers the health of the employee or is one in which an affected employee is likely to present a serious danger to third parties
(iii) The condition is one for which the dangers cannot be eliminated or significantly reduced by reasonable measures taken by the employer to modify or respond to the environmental risks

itself, it may be tempting to consider introducing screening to monitor such workers and exclude them from exposure. The balance of opinion, however, is that such screening for genetic susceptibility is seldom, if ever, justified from either an ethical or a practical standpoint.

A further issue is whether the potential for a substance to cause mutation (and ultimately cancer) can be monitored through the formation of "adducts" when a chemical binds to DNA after exposure (for example, to polycyclic aromatic hydrocarbons) and can be measured in cells obtained from a routine blood sample. Those with the highest number of adducts may be thought to be at the greatest risk of developing cancer, either because their exposures have indeed been higher (perhaps because of poor environmental controls) or because a finding of a particularly high level of adducts may in itself be an indication of an inherent inability to detoxify a particular mutagen (or to repair damage when it occurs). Given the wide variation in adducts in the same individual measured on separate occasions and the uncertainty in interpreting such measures in assessing the risk of cancer in later years, the routine use of DNA adducts as exposure effect markers for individual workers may not be defensible. However, occupational health professionals need to understand the potential importance of such measures, as evidence of a relevant mechanism (that is, the formation of adducts) is already being used by the International Agency for Research on Cancer in the designation of chemicals as carcinogens (for example, ethylene oxide).

Genetic analysis may also have a place in the attribution of causality after disease has occurred. Mutations in the suppressor gene *p53* have been found in most types of cancer; in individual cases, it may be helpful to consider whether the mutation observed is one that occurs more often in tumours associated with one type of exposure, increasing the post facto probability that this is the exposure that was responsible. In epidemiological studies, where an excess of ill health is observed but the importance of exposure is uncertain, showing that those with a genetic susceptibility to the exposure are more likely to develop the disease may again shift the balance towards acceptance of causality.

Genetics and reproduction

The time window for genetic damage in reproductive stem cells differs markedly between men and women. In women, ova that will be available for fertilisation in adulthood go through most phases of development while the fetus is in utero. The implication is that genetic changes to the ovum that will affect children born to the woman will be caused by exposures to the grandmother while she was pregnant. In practice, there is little evidence that this does occur (at least for occupational exposures). In men, by contrast, damage to stem cells that may affect the genetic complement of the resulting child can happen at any time, from in utero exposure to the point at which production of the sperm occurs. There is then a further three month window for adverse environmental effects as the sperm that will eventually fertilise the ovum moves through its final stages of development. Although the time period of opportunity for damage is much greater for men, and the protection from external influences is less stringent than in utero, the evidence of such effects from occupational exposure in humans is sparse.

Finally, environmental exposures in utero have been suspected to cause childhood cancers, although the best evidence for this again comes from pharmaceutical products—mothers' use of diethylstilboestrol—for example,

Statement of the Human Genetics Advisory Committee 1995

(i) An individual should not be required to take a genetic test for employment purposes—an individual's "right not to know" their genetic constitution should be upheld

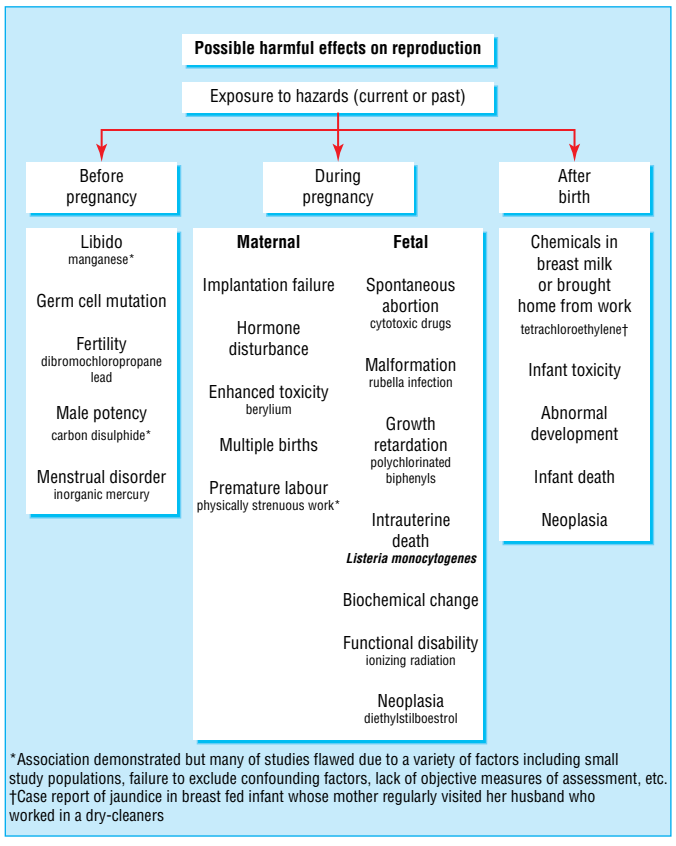
(ii) An individual should not be required to disclose the results of a previous genetic test unless there is clear evidence that the information it provides is needed to assess either current ability to perform a job safely or susceptibility to harm from doing a certain job

(iii) Employers should offer a genetic test (where available) if it is known that a specific working environment or practice, while meeting health and safety requirements, might pose specific risks to individuals with particular genetic variations. For certain jobs where issues of public safety arise, an employer should be able to refuse to employ a person who refuses to take a relevant genetic test

(iv) Any genetic test used for employment purposes must be subject to assured levels of accuracy and reliability, reflecting best practice. We recommend that any use of genetic testing should be evidence based and consensual. Results of any tests undertaken should always be communicated to the person tested and professional advice should be available. Information about and resulting from the taking of any test should be treated in accordance with Data Protection principles

Furthermore, test results should be carefully interpreted, taking account of how they might be affected by working conditions, and

(v) If multiple genetic tests were to be performed simultaneously, then each test should meet the standards set out in (ii), (iii), and (iv)



Possible harmful effects on reproduction

was responsible for vaginal cancers in female offspring as they reached adolescence and beyond. It is likely that such somatic mutations are more common than those to stem cells, which would perpetuate genetically mediated disorders through the generations.

Work and reproductive health

Although the fertility of both men and women can be adversely affected by exposure to chemical compounds (particularly certain pesticides and solvents), metals, the physical environment (heat, radiation), and other factors at work, evidence suggests that the range of exposures with such adverse reproductive effects is fairly limited. Once the fertilised ovum is implanted and begins to develop, the risk seems much greater, with exposure to chemicals, infective agents, and radiation having the capacity to interrupt fetal development during the period of organogenesis (as happened with thalidomide), to interfere with the development of the nervous system (with effects on hearing or eyesight—for example, and possibly on rates of spina bifida), or to result in retardation (not evident at birth) in the infant as it develops. Importantly, there is good epidemiological evidence that heavy physical demands at work are related to fetal death and prematurity. Few occupational cohort studies have been able to follow the offspring of workers into childhood to determine subtle effects on development that may result from exposure in utero, but if community studies of environmental exposures are correct in their interpretation, similar effects of occupational exposure would be anticipated.

Environment and reproductive health

Many of the concerns about effects on the developing infant have arisen from interpretation of community studies of the relation between exposure to lead (from flaking paint or gasoline), household pesticides (used repeatedly in poor quality housing in hot climates), and neurotoxic substances (for example, organic mercury) from diet (fish, game) or water and infant development. Of particular interest in recent years has been the suggestion that endocrine modulators from water, diet (phytoestrogens such as soya), or exposures to—for example, plasticisers such as phthalates, have effects in utero on the male fetus, leading to congenital malformations (hypospadias), low sperm count, and testicular cancer. Results of research into such effects in humans are just becoming available and are not wholly supportive of this overarching hypothesis, but the impetus arising from this elegant synthesis has pushed environmental (and occupational) reproductive health into the focus of regulators throughout the western world.

The figure showing possible harmful effects on reproduction is adapted from Barlow SM, Dayan AD, Stabile IK. Workplace exposures and reproductive effects. In: Baxter PJ, Adams PH, Aw TC, *et al.*, eds. *Hunter's diseases of occupations.* London: Edward Arnold, 2000.

Agents associated with risk to male fertility

Chemical
- Carbaryl: abnormal sperm morphology
- Carbon disulphide: oligospermia, abnormal morphology
- Chlordecone: oligospermia, reduced sperm motility, abnormal sperm morphology
- Dibromochloropropane: oligospermia/azoospermia
- Lead: oligospermia, reduced sperm motility, abnormal sperm morphology

Physical
- Heat: oligospermia
- Ionising radiation: oligospermia or azoospermia

Biological
- Mumps: oligospermia or azoospermia

Some hazards associated with adverse pregnancy outcome

Chemical
- Anaesthetic gases: spontaneous abortion, growth retardation, intrauterine death*
- Organic solvents: spontaneous abortion*
- Lead: spontaneous abortion, intrauterine death, prematurity
- Polychlorinated biphenyls (PCBs): congenital PCB syndrome

Physical
- Ionising radiation: spontaneous abortion, growth retardation, malformation of central nervous system, childhood cancer
- Heavy physical demands, shift work, extremes of temperature: spontaneous abortion, prematurity, growth retardation, intrauterine death*

Biological
- Rubella: spontaneous abortion, intrauterine death, congenital rubella syndrome
- Varicella zoster infection: neonatal infection, congenital varicella syndrome
- Parvovirus B19: hydrops fetalis, fetal loss

*The epidemiological evidence is conflicting for some of these hazards

Further reading

- McDonald JC, ed. *Epidemiology of work related disorders,* 2nd ed. London: BMJ Publishing 2000. *Of particular interest are chapters on work and pregnancy; occupation and infertility; and molecular assessment of exposure, effect, and effect modification*
- Rawbone RG. Future impact of genetic screening in occupational and environmental medicine. *Occup Environ Med* 1999;56:721-4
- Sharpe RM, Skakkebaek NE. Are oestrogens involved in falling sperm counts and disorders of the male reproductive tract? *Lancet* 1993;341:1392-5
- Cherry N, Mackness M, Durrington P, Povey A, Dipnall M, Smith T, et al. Paraoxonase (PON1) polymorphisms in farmers attributing ill health to sheep dip. *Lancet* 2002;359:763-4
- Information on genetic testing can be found at www.hgc.gov.uk

18 Pollution

Robert Maynard

Pollution of the air, soil, and water is a major problem in many parts of the world. In developed countries the worst excesses of industrial pollution are coming under control but have been replaced by pollutants generated by motor vehicles. In developing countries the rapid increase in industrialisation combined with the increased use of motor vehicles is producing conditions as bad, if not worse, than those seen in developed countries a century ago. Dense chemical smog is common in megacities such as Mexico City and São Paulo and is an increasing problem in many of the cities of China and India. Photochemical air pollution is a problem in the Mediterranean area; in fact, only the dense and damp smogs so characteristic of London until the 1950s seem to have disappeared. The combination of a damp, foggy climate and intensive use of soft coal in inefficient household fireplaces does not seem to have been repeated on such a scale elsewhere, although similar conditions may have occurred in Eastern European countries and in Istanbul. High concentrations of coal smoke and sulphur dioxide do occur in some Chinese cities, and forest fires have, over recent years, caused significant "haze" conditions in South East Asia.

Air pollution is not solely an outdoor problem: in many countries indoor pollution produced by the use of biomass as a fuel damages health, especially that of women and young children who may be exposed for much of a 24 hour day. The seemingly inevitable link between poverty and poor environmental conditions persists, and efforts to resolve this and instil a sense of environmental justice are only now beginning.

Air pollution is a major problem but so is pollution of water. Attention has been drawn to the contamination of drinking water with arsenic leached from soil in West Bengal. High levels of lead, nitrates, and pesticides have also been detected in drinking water in various countries. A recent problem in California has been the seepage of methyl *tert* butyl ether (MTBE) into drinking water: an ironic problem as MTBE was added to petrol as an oxygenating agent designed to reduce the production of air pollutants.

London street scene from 1923. The figure shows a classic London "smog"

Mixture of water vapour and smoke being emitted from an industrial site

Air pollution

Air pollution is a worldwide problem. A recent publication by WHO estimated that of 17 major cities, nine had serious problems with suspended particulate matter—the WHO guideline was exceeded by more than a factor of two. The impact of air pollution on health is large: some three million deaths each year are attributed by WHO to air pollution. Of these, 2.8 million result from indoor exposure (1.9 million occurring in developing countries) and only 0.2 million occur as a result of outdoor exposure. Of these 0.2 million deaths, only 14 000 are thought to occur in developed countries. These figures are not easy to interpret. In the United Kingdom, airborne particles (PM_{10}) are thought to be associated with about 10 000 extra deaths every year. Those affected experience by far the greatest part of this exposure indoors. It is salutary to consider how much effort is put into controlling outdoor concentrations of air pollutants compared with indoor concentrations.

Particulate air pollution

Until recently it was believed that airborne concentrations of particles in countries like the United Kingdom had fallen to such levels that effects on health had essentially disappeared. This is now known to be untrue.

An increase in PM_{10} of 10 μg for every cubic metre is associated with about a 1% increase in deaths, although recent studies suggest that a lower percentage increase, perhaps 0.7%, might be more accurate. The effect on non-accidental hospital admissions is of the same degree. In a small country like the United Kingdom this leads to a large impact on health: 8100 deaths brought forward (all causes), and 10 500 hospital admissions (respiratory) either advanced in time (that is, the admission would have occurred but occurs earlier as a result of exposure to pollution) or caused de novo.

It has been argued that the extra deaths calculated in this way are merely deaths advanced by just a few days in those who are already seriously ill. This does not seem to be true, however: recent work by Schwartz has suggested that at least some of the deaths may be advanced by months. Studies in the United States have shown that living in a city with a comparatively high level of particles leads to a reduction in life expectancy.

Calculating the extent of the impact at an individual level is impossible because we do not know how many in a population are affected. If all people were affected equally, then at levels of particles found in the United Kingdom, the individual impact would be small, some days only. If, however (as is much more likely), the effect is unevenly distributed across the population, some would lose months, or even years, of life.

If this is the case in the relatively unpolluted United Kingdom then the effect in much more polluted developing countries must be large indeed. Predicting the size of the effect in developing countries is not easy as it will, in part, depend on the background prevalence of disease. Note that cardiovascular disease is increasing in some developing countries.

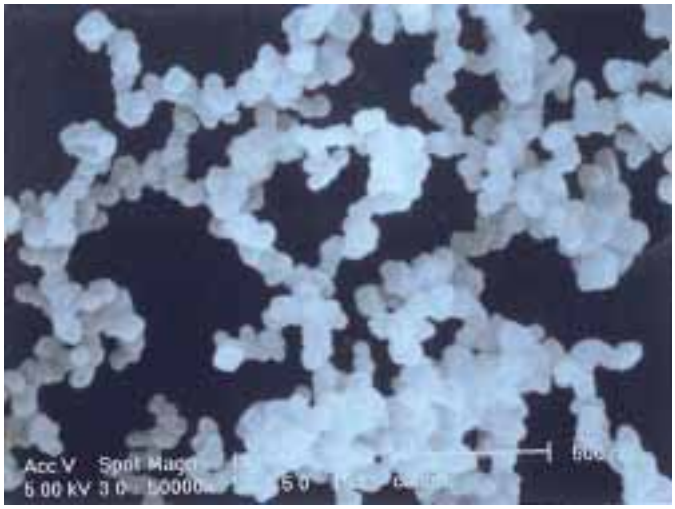
These calculations of the impact of particles on health have produced a revolution in thinking in inhalation toxicology. Some, being unable to understand the exact mechanism of effect, have argued that the associations are not causal. Others have, rather more usefully, set out to find the mechanism of effects, and research has flourished.

Ultrafine particles (less than 100 nm in diameter) have been suggested to play an important role. These particles contribute little to the mass concentration of the ambient aerosol but a great deal to its number concentration. The idea that the number of particles in every cubic metre of air may be more important than the mass per cubic metre has gained ground in recent years. More recently, the idea that total particle surface area per unit volume of air may be important has been discussed. If this is true then air quality standards dependent on mass measurements will need revision. The unusual and unexpected toxicological properties of ultrafine particles have been recently reviewed (see Further reading).

Photochemical air pollution

Concern about secondary pollutants generated from primary emitted pollutants by photochemical reactions began in Los Angeles in the late 1940s. Ozone is the best known photochemical air pollutant produced from nitrogen dioxide (see box) particles; other chemical species, including peroxy radicals derived from volatile organic compounds, are also important. Ozone is the classic example of a secondary air pollutant: essentially no ozone is emitted by sources of outdoor air pollution.

Ozone is a strong oxidising agent and at concentrations above 100 parts per billion (200 $\mu g/m^3$) produces inflammation of



Electron micrograph of diesel particles. Individual particles are about 25 nm in diameter. Photograph kindly provided by Professor RJ Richards, Cardiff University

Modern epidemiological techniques employing time series analysis have shown that day to day variations in outdoor mass concentrations of particles are related to daily counts of events including deaths, hospital admissions, general practitioner consultations, and days of restricted activity

New trends in research on particulate air pollution

- Effects are not limited to the respiratory system; effects on the cardiovascular system are likely to be more important
- Small particles (less than 2.5 μm in diameter) are likely to play an important role
- The production of free radicals, perhaps as a result of metals acting as catalysts, is likely to be important
- Changes in the control of the heart's beat to beat interval and in the production of clotting factors may be important

Ozone production reactions

$$NO_2 + h\nu \rightarrow NO + O^{\bullet}$$
$$O^{\bullet} + O_2 \rightarrow O_3$$
$$RO_2 + NO \rightarrow NO_2 + RO$$

It will be appreciated that as long as sunlight (represented by $h\nu$), oxygen, nitrogen dioxide, and peroxy radicals (RO_2, produced from volatile organic compounds emitted by motor vehicles) are present, ozone production will continue. The reactions stop at night and levels of ozone fall, to build up again the next day. Ozone is thus a problem in cities with heavy traffic and bright sunlight: Athens, Los Angeles, and Mexico City are examples. In the United Kingdom ozone is a greater problem in rural than in urban areas, the formative reactions taking place in polluted air masses drifting from the city to the countryside

the respiratory tract. This is reflected in a reduction in the forced expiratory volume in one second and peak expiratory flow rate. Pain on deep inspiration occurs and these effects lead, unsurprisingly, to a reduction in athletic performance. Interestingly, the effect is short lived, and daily exposure studies have shown that the effect is much reduced by about the fourth or fifth day. Epidemiological studies show that daily deaths and hospital admissions for asthma and other respiratory diseases are related to daily ozone concentrations. Discussion about a possible threshold of effect remains unresolved. If no threshold is assumed, the effects in the United Kingdom are large.

Combinations of air pollutants

Chemical air pollutants never occur alone. There is always a mixture, and it is likely that effects on health are caused by the mixture and might vary with the composition of that mixture. Separating out the more important pollutants has proved to be difficult, and recent studies have shown that the effects of one pollutant may be modified by co-pollutants. This seems to be the case in co-exposures to ozone and nitrogen dioxide. Much more work is needed in this area.

Carbon monoxide (a pollutant that is well known to produce lethal effects at high concentrations) has recently been shown by epidemiological studies to be associated with heart attacks and heart failure at current outdoor concentrations—a remarkable finding. Carbon monoxide may be acting as a marker for other pollutants in the ambient mixture, or at low concentrations it may have unexpected effects in sensitive subjects. Recent studies in volunteers who had angina have shown that carboxyhaemoglobin concentrations as low as 2% are associated with a reduction in "time to pain" on exercise.

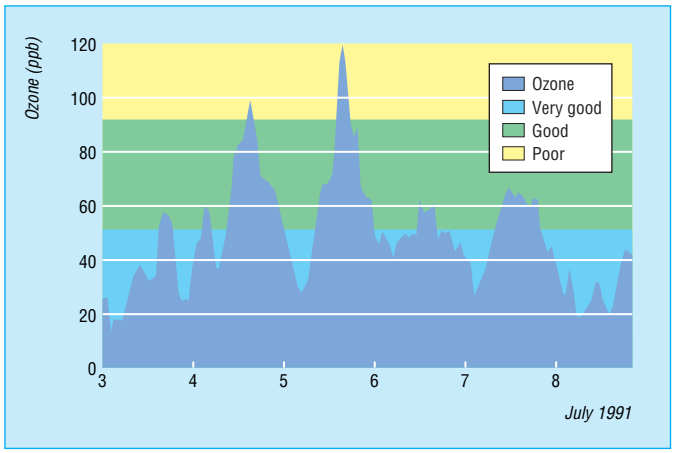
Carcinogenic air pollutants

Many well recognised human carcinogens occur in ambient air, both outdoors and, often to a greater extent, indoors. Studies in UK homes have shown—for example, that concentrations of benzene indoors may exceed those outdoors. Motor vehicles generate benzene, 1,3-butadiene, and polycyclic aromatic hydrocarbons. High levels of arsenic may occur near metal smelting works. These carcinogens are genotoxic and thus at all levels of exposure no guarantee of safety can be provided.

All estimates of increased risk of this sort are based on mathematical extrapolation from studies, in animals or man, of measured increases in risk on exposure to high concentrations. The process is unlikely to be precise and the accuracy of the predictions cannot be ascertained. This has led UK regulators to adopt a pragmatic approach and to set standards for ambient concentrations at levels at which the risk is judged to be very small and not to attempt quantification of the effects. Thus, in the case of benzene, a standard of five parts per billion (15.6 $\mu g/m^3$) expressed as an annual average concentration has been adopted.

Indoor air pollution

All the pollutants discussed above, with the exception of ozone (which reacts rapidly with furnishings and fittings and disappears), occur indoors. Indoor concentrations are, in part, driven by outdoor sources as well as by specific indoor sources. Carbon monoxide and nitrogen dioxide may be produced by fires and by cooking—peak levels in kitchens can be higher than those commonly found outdoors. Recent work has led to concern about an association between nitrogen dioxide and respiratory infections, worsening of lung function in women with asthma, and increased sensitisation and response to allergens. Long term exposure to low levels of carbon



Daily variations in ozone concentrations.

Numbers of deaths and hospital admissions for respiratory diseases per year caused by ozone in both urban and rural areas of Great Britain (GB) during summer only

	GB, threshold (in parts per billion)	
	50	**0**
Deaths (all causes)	700	12 500
Hospital admissions for respiratory disease	500	9900

The WHO has published "unit risk factors" that allow the risk to be estimated (expressed as an increase in risk of getting a specified cancer as a result of lifetime exposure to a unit concentration of the carcinogen). For example, lifetime exposure to benzene of 17 $\mu g/m^3$ is estimated to be associated with an increase of risk of 1 in 10 000. The unit risk at 1 $\mu g/m^3$ is estimated as 6×10^{-6}

monoxide that produce only mild symptoms may lead to lasting neurological effects.

Regulating indoor pollutant concentrations is difficult: fewer countries have produced standards for indoor air quality than for outdoor air quality. The need for regular maintenance of devices that can produce pollutants indoors, for smoke alarms, and for constant vigilance on the part of doctors dealing with potentially poisoned patients is obvious.

Water and soil pollution

In developed countries the quality of drinking water is often accepted unthinkingly as high: we assume that the water is safe to drink. In many countries, however, such an assumption may be unwise because of microbiological and chemical contamination. The former causes more disease than the latter but will not be considered here. Accidental contamination of water supplies occurs from time to time in all countries: in the United Kingdom the accidental contamination of water with aluminium sulphate in Camelford (Cornwall) in 1988 (see chapter 20) led to widespread complaints that are still the subject of investigation. The quality of water supplies is improving in many countries, but the rate of improvement is uneven. WHO reported that in the period 1990-4 the number of people without a satisfactory water supply increased in Africa, Latin America, and the Caribbean. In some countries, including developed countries such as the United Kingdom, concern has been expressed about the possible impacts on health of so called endocrine disrupting chemicals.

Conclusion

Pollution of air, soil, and water remains a problem in nearly all parts of the world. In developed countries air pollution tends to attract the greatest attention, and considerable efforts to control outdoor sources of air pollutants have been made. In developing countries both air and water pollution remain important problems, and a large effort will be needed before these are removed.

Compounds that are of proved concerns as water and soil pollutants

Arsenic
Arsenic is found in high concentrations in many countries including Argentina, Canada, Chile, China, Japan, Mexico, the Philippines, and the United States. The recent discovery of arsenic concentrations at 70 times the national standard of 0.05 mg/l in West Bengal has highlighted this pollutant. Poisoning via water leads to evidence of chronic toxicity including melanosis, hyperkeratosis, and skin cancer. In West Bengal 200 000 people are reported to be suffering from arsenical skin lesions

Nitrates
Nitrates leached from agricultural land may enter drinking water. The use of infant food prepared with such water can lead to poisoning, methaemoglobin being produced by interaction between nitrite ions (produced from nitrate ions) and haemoglobin. The reaction is an oxidative one (ferrous iron in haemoglobin being converted to ferric iron in methaemoglobin) but the exact mechanism is unclear. In very young children cyanosis may occur. In 15 European countries 0.5-10% of the population may be exposed to nitrate levels in excess of the WHO standard of 50 mg/l

Lead
Lead can be mobilised from pipes and solder joints, especially in areas with acidic water supplies ("soft water" areas). Lead is accumulated in the body and can damage the central nervous system. A number of studies have linked lead intake and a decreased intelligence quotient. Mercury and cadmium are examples of other metals that contaminate water supplies

Fluoride
Fluoride is added to water in some countries to provide protection against tooth decay: effective protection is provided at levels of 0.5-1.0 mg/l. The margin between protective and toxic effects is unfortunately narrow, and effects ranging from dental fluorosis (mottling of enamel) to skeletal fluorosis occur in some areas. High levels of fluoride are found in parts of the Middle East, Africa, and North and South America

Based on recent work by the WHO

Further reading

- Holgate ST, Samet JM, Koren HS, Maynard RL, eds. *Air Pollution and Health*. London, New York: Academic Press, 1999. *A comprehensive review of all aspects of air pollution*
- World Health Organization. *Air Quality Guidelines for Europe*, 2nd ed. WHO Regional Publications, European Series, No 91. Copenhagen: WHO, 2000. *An update of the original 1987 edition providing guidelines for 35 air pollutants*
- World Health Organization. *Health and Environment in Sustainable Development. Five years after the Earth Summit*. Geneva: WHO, 1997. *A useful and wide ranging report*
- Maynard RL, Howard CV, eds. *Particulate matter: properties and effects upon health*. Oxford: BIOS Scientific Publishers Ltd, 1999. *Detailed collection of papers by leading research workers*
- Schwartz J. Harvesting and long term exposure effects in the relation between air pollution and mortality. *Am J Epidemiol* 2000;151:440-8. *An important paper by the leading exponent of time series studies of the effects of air pollutants on health*
- Department of Health. *Committee on the medical effects of air pollutants. Non-biological particles and health*. London: HMSO, 1995. *A detailed report containing much useful information but becoming a little dated*
- Department of Health. Advisory Group on the Medical Aspects of Air Pollution Episodes. First Report. *Ozone*. London: HMSO, 1991. *A detailed report containing much useful information but now dated*
- Department of Health. Committee on the Medical Effects of Air Pollutants. *Quantification of the effects of air pollution on health in the United Kingdom*. London: HMSO, 1998. *A report providing guidance on how to estimate the effect of short term exposure to air pollutants*
- Department of Health. Committee on the Medical Effects of Air Pollutants. *Statement and Report on Long term effects of particles on mortality*. Available from http://www.doh.gov.uk/comeap/statementsreports/longtermeffects.pdf. *A report providing a novel approach to estimating the effects on life expectancy of long term exposure to air pollutants*
- Department of the Environment. Expert Panel on Air Quality Standards. *Benzene*. London: HMSO, 1994. *Provides an example of a pragmatic approach to setting standards for carcinogenic air pollutants*
- World Health Organization. *International programme on chemical safety. Guidelines for drinking-water quality*, 2nd ed. Vol. 1. *Recommendations*. Geneva: WHO, 1993
- World Health Organization. *International programme on chemical safety. Guidelines for drinking-water quality*, 2nd ed. Vol 2. *Health criteria and other supporting information*. Geneva: WHO, 1997. *These two reports contain detailed and invaluable accounts providing background material to the guidelines*
- Brunekreef B, Holgate ST. Air pollution and health. *Lancet* 2002;360:1233-42. *A comprehensive review article with extensive references*
- *Health effects of climate change*. www.doh.gov.uk/airpollution/climatechange02/index.htm. *An up-to-date and comprehensive overview of all aspects of this problem*

19 Global issues

Tony Fletcher

Environmental effects on health can be direct, as in the effect of lead pollution on the development of children, or indirect—for example, the eventual health impacts of the loss of biodiversity among plant and animal species. Many widespread and global environmental issues impact on human health. A selection of important areas follow along with some of the policy instruments that seek to mitigate these global risks.

Volcanic eruption, Montserrat—pollution on a grand scale and a social disaster

Carrying capacity

Man's activities of production and consumption affect not only our local environment but the environment of whole regions and the entire planet. Given the large scale of such activities in an increasingly globalised world, certain polluting or resource depleting activities that the carrying capacity of the local environment used to absorb now result in overload or contamination of global proportions. One of the best recognised examples in the 20th century has been the devastating effect of acid rain on natural ecosystems whose ability to absorb and eliminate sulphuric acid was overwhelmed.

The carrying capacity amounts to some 10 hectares for every person for the richest countries compared with only 2.5 hectares per person on a global average. Thus, on this and other measures there is not enough land to support the world's population at the level of consumption enjoyed by the most industrialised countries. Of course, there is no widespread enthusiasm to reduce levels of consumption. On the contrary, there are widespread aspirations to increase industrial production and employment and to reduce, or eliminate, poverty.

Various attempts have been made to estimate how many people can comfortably and sustainably live on this planet, based on some reasonable compromise between the (low) current average standard of living and the high average in the richest countries. Realistic estimates based on food production, water usage, energy consumption, and the integrated footprints fall mostly in the range of three to five billion people. With a world population of six billion and projected increases to at least 10 billion before any prospect of levelling off, the sustainable carrying capacity of the planet is already being exceeded.

Silver mine—the social price paid for precious metal

Biodiversity

As pressure on land has increased in the past 100 years, the rate of extinction has accelerated. It is estimated that 20-50% of species present 100 years ago will have become extinct by 2100, with the rate of loss accelerating from now until then.

Many species are lost as biodiverse tropical rainforests are depleted by clearance and burning. This has practical consequences on human health by affecting food and drugs. Medicines have been identified and developed from tropical plants, and pharmacological possibilities for numerous species have not been explored. In the case of food, we have in the past relied on the cross breeding of food crops with wild strains to maintain productivity and resistance to pests, and will no doubt need to continue to do this, whatever achievements arise from genetic modification in laboratories.

The concept of *carrying capacity* for people derives from the *ecological footprint*: "the area of productive land and water required on a continuous basis to produce all the resources consumed, and to assimilate all the wastes produced"

The rate of loss of species is estimated to have increased from 10 000 a year in 1900 to some 50 000 a year in 2000

Arsenic in ground water

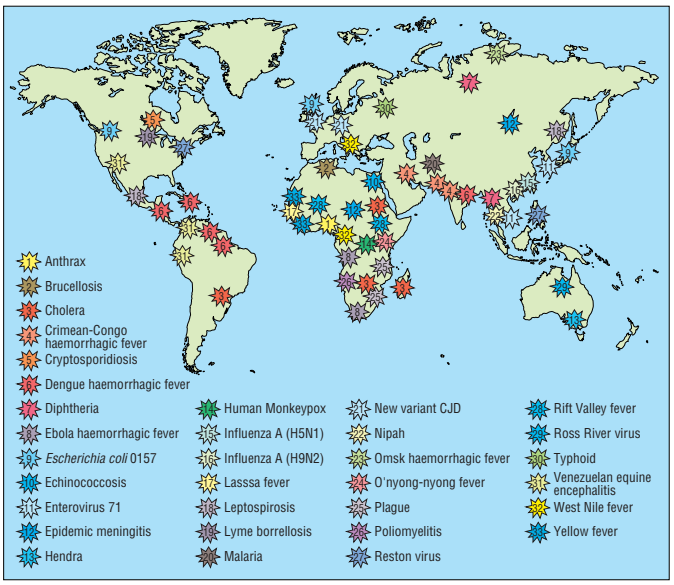
Arsenic is a widespread mineral in the Earth's crust and occurs at rather high concentrations in some ground waters that get exploited as sources of drinking water. In Europe few populations have been exposed and signs of arsenic toxicity, including melanosis and keratosis of the skin, peripheral neuropathy, and vascular and respiratory ill health, are rare. Consequently, attention has shifted to cancer risk, estimated by extrapolation from more exposed populations. Skin and internal cancers (bladder and lung in particular) have been clearly associated with higher levels of arsenic exposure and concern about risks at relatively low concentrations has prompted most countries to adopt a standard of 10 μg/l in drinking water. However, in some developing countries, most notably Bangladesh, it has recently been discovered that more than 20 million people are being exposed to over 50 μg/l and levels over 1000 μg/l have been found. Thus, many thousands of cases of arsenic poisoning have already occurred, and this is likely to be followed by many fatal cancers, especially if remediation measures are not put into place rapidly.

Global climate change

A significant upward trend in mean temperature is now in little doubt, and anthropogenic sources are substantially contributing, especially in relation to carbon dioxide. The Intergovernmental Panel on Climate Change (IPCC) in 2001 predicted an increase in 2-3°C over the course of the 21st century. The authoritative IPCC, among others, has highlighted a number of serious consequences if this were to happen. One consequence, the increase in frequency of extreme weather events, already seems to be underway. Melting of the polar ice caps, accompanied by a rise in sea levels, would lead to disastrous flooding of low lying coastal regions. Shifts in climate patterns would lead to changes in agricultural productivity and, given the time and investment it takes to adapt to such changes, most likely famine and conflict. The distribution of infectious diseases is expected to change, with malaria migrating north as temperatures rise. Other vector borne diseases threatening to spread include dengue fever, viral encephalitis, schistosomiasis, leishmaniasis, onchocerciasis, and yellow fever.

Stratospheric ozone depletion

Few have not heard of the ozone hole, with the accompanying increase in ground level exposure to ultraviolet radiation and the consequent increase of sunburn and skin cancer. Damage to the ozone layer is the result of chemical reactions between stratospheric ozone and certain chemicals, most notably the chlorofluorocarbons used as aerosol propellants and refrigerants. Although chlorofluorocarbons are inert at ground level temperatures, in the very cold stratospheric environment over the poles, sunlight breaks them down into reactive intermediates which in turn destroy the ozone present there. As the ozone is depleted, more harmful solar radiation gets through. In recognition of this problem, chlorofluorocarbons were banned under the Montreal Protocol of 1987. This was not implemented in every country, however, and in the meantime other ozone destroying chemicals have been identified. Because it takes some time for the ozone to build up again, ozone depletion is not expected to recover until the middle of the 21st century. Ground level ultraviolet light is predicted to rise by 12-15% relative to 1970s levels.



Unexpected outbreaks—examples of emerging and re-emerging infectious diseases, 1994–9

Emerging and re-emerging infectious diseases

Emerging infectious diseases are those that have been recently discovered, have increased in humans over the past two decades, or threaten to increase in the future

Re-emerging infections are infectious diseases which have increased (previously having diminished in incidence) because of ecological changes, public health decline, or development of drug resistance

Six major factors have contributed to their emergence or re-emergence:

1. Changes in human demography and behavior (for example, immunosuppression, aging population, migration, risky behaviours)
2. Advances in technology and changes in industry practices (for example, air conditioning cooling towers, changes in food processing, changes in rendering.)
3. Economic development and changes in land use patterns (encroachment on the tropical rainforests, conservation efforts, climate changes)
4. Dramatic increases in volume and speed of international travel and commerce of people, animals, and foodstuffs
5. Microbial adaptation and changes
6. Breakdown of public health capacity for infectious diseases

In most instances, the emergence of a specific agent results from a complex interaction of several factors

Examples of emerging and re-emerging infections include:

HIV
Legionnaire's disease
Hantavirus
E.Coli O157
Vancomycin resistant enterococci
Severe acute respiratory syndrome (SARS)

Lead in the environment

Lead has multiple toxic health effects—haematological, renal, and neurological—although at typical levels of exposure in the environment, neuropsychological impacts are the main concern, especially for developing children. Aside from local contamination or pollution, exposure to lead has been quite widespread from dissolution into drinking water from lead piping, use of lead in paint in old houses, and airborne exposure from leaded petrol. In addition, people have been exposed via their food from the use of lead solder for sealing cans, although this has now been completely phased out. However, the other sources still lead to exposure. Although the use of lead additive in petrol has virtually ceased, there is still much dust on roadsides from past use and this is resuspended or picked up by children; lead present in paint in older houses remains an important source as it is chipped off through normal wear and tear; again in older dwellings, lead pipes in the home or connecting with the main water supply can be a source, with solubility depending on the chemistry and pH of the water supply. Research into the effects of lead exposure on children's neurological development measures their intelligence quotient and emotional and behavioural development.

In children, a doubling of body lead burden 10–20 mcg/dl is associated with a deficit of 1–2 full scale IQ points

Lead smelter—the starting point of dissemination of a toxic metal

Polychlorinated biphenyls (PCBs)

PCBs, along with dioxins and chlorinated pesticides such as DDT, exhibit a particular persistence. They are not broken down in the environment and indeed are accumulated as they are taken up by plants, herbivores, and carnivores, with concentrations increasing at each level. Mass production of PCBs started in 1929 and expanded enormously until environmental damage was recognised in the 1960s. The first evidence of adverse health effect came in 1968 with the mass poisoning of Japanese who ate rice oil contaminated with PCBs. The resulting Yusho disease (named after the place where this incident took place), entailed disfiguring pigmentation of the skin, sweating, conjunctivitis, headaches, weakness, cough, and liver damage then later an increased incidence of cancer. For pregnant women, exposure led to malformed children. The effect of bioaccumulation on wildlife was subsequently established—for example, reduced fertility in various seabirds feeding on fish with accumulated PCBs. Production and use was, after some delay, reduced, with bans being introduced in the 1970s in Western Europe and the 1980s in Eastern Europe. Environmental levels of PCBs are slowly reducing again.

Smokestack industry—global relocation to the poorest countries

Transport and health

In recent years attention has been increasingly focused on the links between transport and health. The traditional dominant concern—death and injury from collision—has been extended to embrace transport related emissions. Pollution from industry and domestic fireplaces has fallen, but as a result of the phenomenal increase in mobility and car ownership, motor vehicles are now the main source of emissions in the United Kingdom (directly or through atmospheric reactions) of particulates, oxides of nitrogen, and ozone.

Added to these more obvious health risks are other concerns relating to quality of life and health—for example, trade-offs between different psychosocial impacts. Greater mobility has the potential to increase more distant social contacts and potentially to provide protection against

vulnerability to assault. Set against this is the community severance and loss of social networking and support within communities that is a consequence of busy roads passing through residential areas. Regular physical exercise is reduced if walking and cycling are cut in favour of car journeys; this has been shown dramatically in the case of children's journeys to school. Not only is this unhealthy per se, but it ingrains the habit of car dependence, leading to the cumulative effect of insufficient exercise.

Policy responses

This range of examples has been selected to illustrate the wide range of environmental impacts on health. Our choice of mode of transport has both local and global impacts. Our energy consumption has an impact via its contribution to global warming. Toxic chemicals in the environment may be local problems or bioaccumulating and thus contribute to very distant increases of risk. These may be natural as in the case of the arsenic contamination of drinking water, or manmade as in lead pollution.

In a parallel manner, the range of policy instruments for preventing adverse environmental impacts operates on various levels. At a global level international conventions play a major role, although the important and potentially expensive ones are the most difficult to get all parties to agree to and ratify. The Kyoto agreement on limiting climate change gases will remain in limbo as long as the major polluters refuse to ratify it.

Regulation at national and, for the European Union, European level is embodied in directives, regulations, and policies, such as, in the United Kingdom, NAAQS.

Local initiatives prompted by the meeting on the environment and development in Rio de Janeiro have become an important focus for both local authority initiatives and the involvement of civil society. The so-called La21, or Local Agenda 21, developed as an idea intended to catalyse local environmental initiatives.

Finally, industrial undertakings by their very size can have large environmental impacts, or make products with significant environmental impacts. Responsible corporate and product stewardship can be implemented to seek to reduce adverse environmental (or environmental health) impacts. This may be represented by adherence within the worksite to quality standards such as the Eco-Management and Audit Scheme or ISO 14 000 environmental quality schemes, or the adoption of "cradle to grave" product stewardship initiatives, ensuring that raw materials such as wood are derived from sustainable sources, recycling is maximised, and products are designed so that they can be recycled.

Toxic waste

Many of the materials we use are useful yet have an inherent toxicity, or to produce them entails generating toxic waste, which in any case needs careful disposal to avoid or at least minimise human exposure. This has frequently not been the case, with a legacy of contaminated land or poorly documented waste sites in the United Kingdom, or waste being exported to other less well regulated countries. This latter practice has been somewhat restricted through international agreements such as the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal (1989), although not all countries are signatories to this convention.

War pollution—burning oil wells in Kuwait

Further reading

- Fletcher T, McMichael AJ, eds. *Health at the Crossroads: Transport Policy and Urban Health*. Wiley, 1997
- Intergovernmental Panel on Climate Change. *Third assessment report*. Cambridge: Cambridge University Press, 2001
- Koppe JG, Keys J. *PCBs and the precautionary principle in late lessons from early warnings: the precautionary principle 1896-2000*. Copenhagen: European Environment Agency, 2001
- Lippmann M, ed. *Environmental toxicants: human exposures and their health effects*, 2nd ed. Wiley, 2000
- McMichael T. *Human frontiers, environments and disease; past patterns, uncertain futures*. Cambridge: Cambridge University Press, 2001
- Patz JA, Kovats RS. Hotspots in climate change and health. *BMJ* 2002;325:1094-8. *A readable account of the key issues*
- Smith AH, Lingas EO, Rahman M. Contamination of drinking water by arsenic in Bangladesh: a public health emergency. *Bulletin of the World Health Organization* 2000;78:1093-1103
- United Nations Environment Programme. www.unep.org. *Up-to-date information on world environmental issues such as POPs (persistent organic pollutants), post-conflict problems, sustainable construction, safe construction, safe technology transfer, fair trade, and disaster management, as well as the topics highlighted in this chapter*
- World Development Report 1992: Development and Environment. Oxford: Oxford University Press, 1992

The figure showing examples of emerging and re-emerging infectious diseases is adapted from the World Health organization infectious disease report (http://www.int/int/infectious-disease-report)

20 Occupational and environmental disease of uncertain aetiology

Andy Slovak

Occupational and environmental conditions, by their nature, invite and create contention. This is particularly so where causality is uncertain. The diagram seeks to explain why this might be. Individual and group beliefs, behaviours, and so on, and their social modulation seem to play as substantial a part in the experience of symptoms as does exposure to the range of putative causal agents.

The issues of causality, attitude, and perception that affect approaches to these conditions are discussed first, before discussing specific syndromes. From a practical point of view, there is an obvious dichotomy between the support it is proper to give to patients and the more detached objectivity one would wish to bring to understanding their condition scientifically. This is particularly so when, as is often the case, health professionals are invited to make a commitment to a particular belief system related to the causality of the disease under discussion. At the same time, those health professionals are all members of the public and as such are susceptible to prevalent, popular, belief systems.

The box lists a selection of medical syndromes whose nature and aetiology are at present uncertain. They are an apparently disparate grouping, but as far as broader circumstances are concerned, they tend to reflect some common themes:

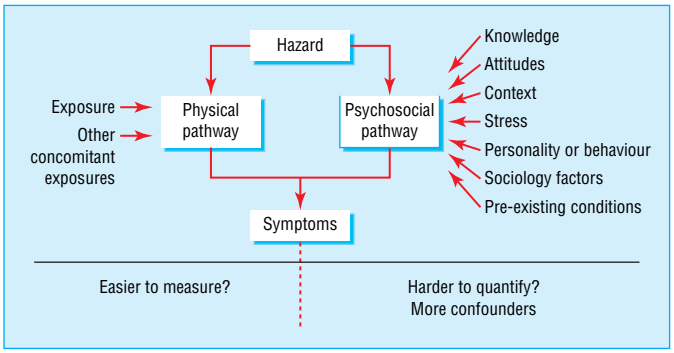
- Multifactoriality (both symptoms and putative causes)
- Lack of control ("involuntary" exposure)
- Marked variation in susceptibility
- Tendency to ascribe to external causes.

As such, there are resonances between these conditions and others considered elsewhere in this book or that are beyond its scope. These conditions include non-specific upper limb disorders, regional pain syndromes, fibromyalgia, and stress.

It is also perhaps worthwhile deconstructing the nature of the multifactoriality a little. In the cases of sick building syndromes, multiple chemical sensitivity, and war syndromes, the factors implicated are truly extensive and highly varied, whereas in situational syndromes (for example, after the Braer disaster and Camelford incidents) they are defined by the event, although still difficult to pinpoint. The debates about electromagnetic fields and nuclear installations are even more unusual because they are biphasic with more or less distinct occupational and environmental modes.

Landfill sites present another situational pattern, having a wide but segmented range of attributed, putative effects (in time and space), and an even wider range of possible hazards and attendant risks. In contrast to all of the foregoing, the putative causal agents cited for conditions attributed to sheep dipping are highly specific—namely, organophosphate pesticides. Acute effects are well known and well characterised, but the controversy continues as to how they might be implicated in longer term effects.

More dramatic examples are provided by disasters and their health consequences, which may vary according to factors other than easily measurable ones. The health deficit produced by the Chernobyl incident in 1986 has been, with the exception of



Medical syndromes with uncertain nature and aetiology

- Long term conditions claimed to be associated with proximity to electromagnetic fields (for example, power lines) and nuclear installations
- Gulf war syndromes
- Multiple chemical sensitivity
- Situational syndromes: the Braer disaster and the Camelford incident
- Sick building syndromes
- Conditions claimed to be associated with proximity to landfill sites
- Long term conditions claimed to be associated with pesticides used for sheep dipping

"Natural" disasters such as volcanic eruptions are less likely to have prolonged health effects (other than obvious immediate ones or secondary effects due to—for example, evacuation) than manmade disasters that usually last longer (food contamination), take time to come to light (bovine spongiform encephalopathy or variant Creutzfeldt-Jakob disease), can have a serious impact on Government or medical credibility, and create many cases of "illness" among the worried well

a few thousand cases of childhood thyroid cancer, wholly psychological and socioeconomic.

Electromagnetic fields and nuclear installations

Studies, mainly epidemiological, relating to power lines have been pursued for about 30 years and for nuclear installations, for about 20 years. Their conclusions are still hotly debated.

The suspect agent in the electrical sector has been electromagnetic fields. The issues of interest occupationally have been leukaemia and brain cancer in electrical workers. Environmental concerns have focused on childhood leukaemia and, to a lesser extent, other childhood cancers.

The outcomes of extensive, painstaking epidemiological work on electromagnetic fields in both the occupational and environmental sectors have been tantalising.

As an example, the table shows the amalgamated risk data from a number of recent Nordic studies of electromagnetic fields "exposure" and childhood leukaemia.

Numbers are sparse and exposure criteria poorly defined. This situation of finding difficulty in differentiating "effect from background noise" is typical of these sorts of long running debates. The rationale for continuing is nevertheless powerful because of the universality of exposure, the likelihood that exposure will increase in the future, and for reasons of risk perception. Studies on those occupationally exposed are even weaker.

With regard to nuclear installations, concerns have also centred around childhood leukaemia and other childhood cancers. In the United Kingdom the debate was initiated by a single television programme in 1983. The putative risk factor at that time was assumed to be installation discharges of radioactive materials. However, such discharges produce doses to the general public that are very small (by several orders of magnitude) when compared with those that might be expected to produce such effects according to robust scientific risk estimations.

Studies on electromagnetic fields have struggled to develop sufficient power to dispel the tentative concerns raised by other studies

In a letter to the *Lancet*, Ahlbom and colleagues reported: "Our results show that the three Nordic studies taken together support the hypothesis that exposure to magnetic fields of the type generated by transmission lines has some aetiological role in the development of leukaemia in children" (*Lancet* 1993;342:1295)

Amalgamated data from three principal Nordic studies concerning childhood cancer risk and proximity to power lines

	Leukaemia			Nervous system tumours			Lymphoma			Total		
Study	Cases	Relative risk	95% CI	Cases	Relative risk	95% CI	Cases	Relative risk	95% CI	Cases	Relative risk	95% CI
1	7	2.7	1.0 to 6.3	2	0.7	0.1 to 2.7	2	1.3	0.2 to 5.1	12	1.1	0.5 to 2.1
2	3	1.5	0.3 to 6.7	2	1.0	0.2 to 5.0	1	5.0	0.3 to 82.0	6	1.5	0.6 to 4.1
3	2	1.6	0.3 to 4.5	5	2.3	0.8 to 5.4	0	0.0	0.0 to 4.2	11	1.5	0.7 to 2.7
Total	13	2.1	1.1 to 4.1	9	1.5	0.7 to 3.2	3	1.0	0.3 to 3.7	29	1.3	0.9 to 2.1

* Data from Anlbom et al. *Lancet* 1993; 342: 1295.

A period of intensive research (1983-90) failed to find much support for an environmental (discharge) hypothesis, and this theory was substantively supplanted in 1990 by an occupational hypothesis based on paternal preconceptual irradiation in radiation workers. Centred on the experience of workers living in Seascale, a village near Sellafield in Cumbria, the paternal preconceptual irradiation theory did not survive when tested elsewhere in Cumbria and more generally in the United Kingdom and other countries. More recently, excess childhood leukaemia rates have been attributed to population mixing in communities that have a high proportion of incomers, raising the possibility of a viral aetiology.

Even though the scientific plausibility has subsided, these matters can still polarise scientific opinion at the extremes of construct belief.

Gulf war syndromes

In 1992, after some months of preparation, a multinational army was engaged in a brief, one sided conflict against Iraq resulting in minimal military casualties to them. Subsequently there emerged among the veterans of some nations, particularly the United States and United Kingdom, a series of symptoms now collectively known as Gulf war syndromes. These have now been extensively studied in the surprisingly large numbers of personnel involved.

A number of separate or combined entities have been suggested as possible causes, and it is easy to draw up a list of 30 or 40 factors that might plausibly come into the frame. Some of these are listed in the box.

Epidemiological studies have looked for distinct Gulf war syndromes and what exposures or experiences might be associated with them. To date, the key observations are that there is no specific syndrome, but that war theatre veterans have discernibly stronger symptoms per person than military personnel who did not serve in the Gulf. There is an association between symptom frequency and complexity of inoculation programmes (where given). Mortality, morbidity, and reproductive outcomes have been unremarkable to date, although, of course, it is early days yet. The same caution that is applied to negative data must also be applied to early positive associations now beginning to be reported (for example, amyotrophic lateral sclerosis).

Those who remember the Vietnam war recall analogous later manifestations of complaints attributed to that experience and specifically to Agent Orange, a widely deployed defoliant. Recent retrospective research suggests that increased symptom frequency may be associated with conflict experience back to the mid-19th century. This subject is set to run for some time.

Some of the most popular possible causes of Gulf war syndromes

- Inoculation programmes
- Prophylaxis against biological warfare agents
- Depleted uranium
- Insecticide spraying
- Pyrolysis due to military action or "scorched earth" action
- Involuntary dispersal of chemical or biological war agents due to military action

Gulf war soldiers in protective clothing

Multiple chemical sensitivity

Multiple chemical sensitivity is a difficult entity to position clinically. The range of symptoms observed within the scope of the condition are protean. They include chronic fatigue type syndromes, weakness, sleep disturbance, rashes, headache, chest tightness, and oppression, but these examples are far from a complete and arguable list.

It is accepted that multiple chemical sensitivity applies to a group of patients with a disabling condition with symptomatology whose severity seems to be lessened by restriction of exposure to the everyday environment, particularly by inhalation. It is inferred, therefore, that the aetiology or precipitation of the condition is derived from that environment. Underlying immunological and neurological mechanisms have been proposed but not substantiated. Diagnostic criteria are hard to define, as are objective investigative methods.

Braer and Camelford

The Braer oil tanker ran aground in Northern Scotland, releasing a cargo of light crude oil (1992-3). At Camelford in Cornwall in 1988, a specific agent, aluminium sulphate, was inadvertently introduced into the water supply. Both incidents resulted in immediate symptoms in local residents. Those associated with the Braer were primarily acute and upper

respiratory, whereas those in Camelford were more diffuse. The former did not persist; the latter did.

The persistence of a "malaise" syndrome, as at Camelford (for example, fatigue, joint pain, depression, memory loss) is a feature of a number of the conditions described in this chapter, and the contrast offered between the Braer and Camelford events may therefore be generally instructive. Toxicologically the circumstances were different, and indeed it is possible to argue that local exposures after the Braer disaster were minimal or non-existent because of vigorous weather dispersal of pollution.

Others have argued that the difference lay in contrasting approaches to the situation by the responsible public authorities. In a perhaps oversimplistic way, it has therefore been inferred that a strong social and sociopsychological modulation of response can be obtained in situational events by authoritative communication and action.

Aspects of societal action that have been perceived as positive in these circumstances are given in the box.

Aspects of societal action perceived as positive following situational disasters such as Braer and Camelford

- Timely communication of an action plan
- Timely communication of hazard information
- Effective dialogue with the population at risk
- Feedback to the population at risk

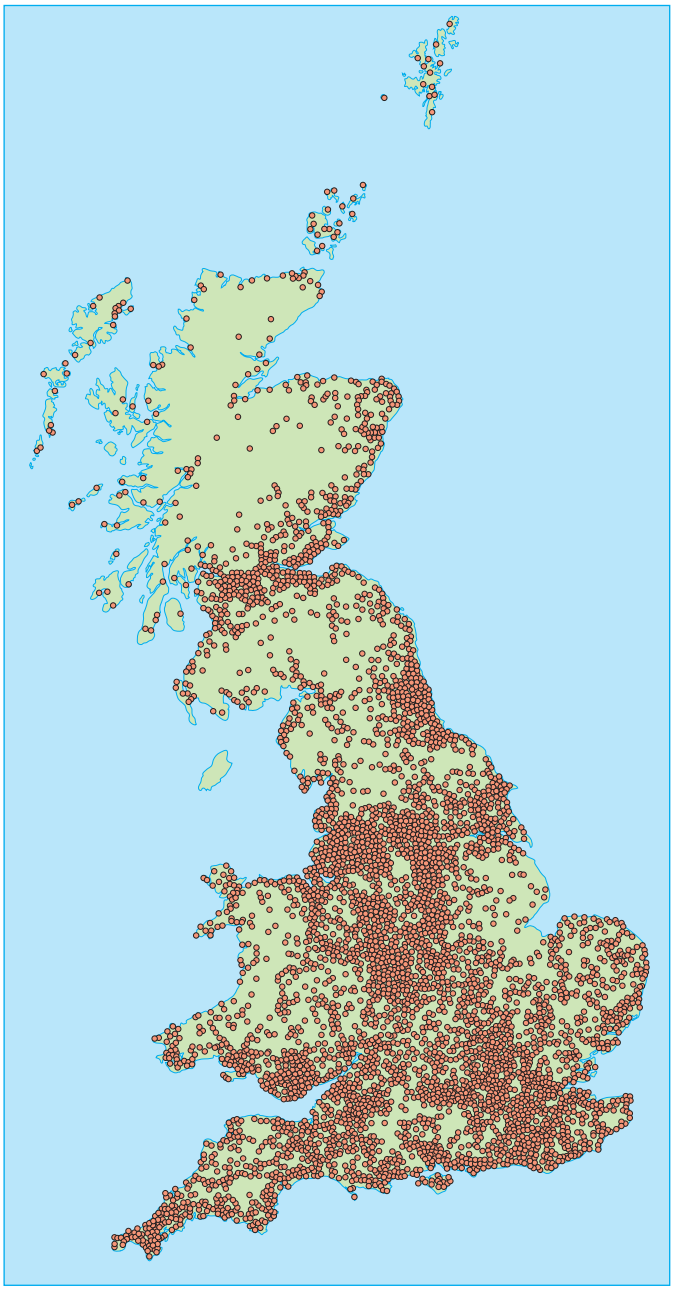
Sick building syndrome

Originally seen and reported as primarily a respiratory occupational disease in the early 1980s, sick building syndrome has more recently seemed more protean and diffuse. Thus, a malaise syndrome similar to situational events was increasingly recognised in later studies; the incidence of complaint was found to be overlaid on quite high prevalence of such symptoms even in buildings not associated with sick building syndrome. Strong linkages to upper limb disorders and stress symptomatologies have also been increasingly reported. It seems likely that these different functional ways of looking at syndromes may be taking different slices out of what is part of the same cake.

Landfills

The figure shows the geographical distribution of UK landfill sites and perhaps also the futility of "nimbyism." (Nimby is an acronymic characterisation—"not in my back yard"—of resistance to the location of any undesired feature in a particular neighbourhood.)

A recent large UK study showed a small association (about 1%) between congenital abnormalities and proximity to a landfill site. It was unclear whether this finding should be seen as reassuring, given that the effect was small and "proximity" encompassed a majority of the UK population. Landfill and other "amenity" sites are particularly likely to be the subject of local symptomatic complaint and anecdotal reporting of cluster events (for example, cancers). Investigationally, the range of reported conditions and (usually) the lack of easily definable toxic exposure may result in unsatisfactory outcomes for all parties involved.



Distribution of UK landfill sites

Organophosphate sheep dips

Organophosphates are well characterised neurotoxins that exert their effects acutely by inhibiting of the widespread neurotransmitter enzyme acetylcholinesterase. This toxic effect has been widely exploited in pesticide applications, as in sheep dips for parasitic infestations. The acute effects have also been widely seen in humans in occupational, domestic, and deliberate overexposures. A typical acute syndrome of stomach cramps, weakness, paralysis, and collapse is directly associated

with measurable cholinesterase suppression in a traditional dose-response relation.

Two syndromes of longer duration have been attributed to long term effects of organophosphate exposure, mainly in sheep dippers who have high and repeated contact with these agents. One, known colloquially as "dipper's flu," is reported to come on some time after exposure, typically up to a day later. As the name indicates, the illness is described as "flu-like" in nature and duration. The other syndrome or syndrome set is reported to be truly chronic, and the syndrome range is typical of that described repeatedly in this chapter. At the anecdotal level, some preponderance of chronic fatigue and cognitive deficits is claimed.

Dipper's flu has been subjected to objective field investigations of exposure and symptoms following dipping. Little difference was observed between symptoms of dippers and unexposed controls when symptoms were grouped (for example, cognitive, visual, flu-like). When symptoms were degrouped and analysed separately, some emerged as more common in dippers but these were not those of flu. Thus, despite quite extensive research, the findings continue to be inconclusive or perverse, and there is no clear dose or dose-surrogate relation. Complaints of neuropathy are more frequent in sheep dippers, especially those handling concentrate, who are also more prone to anxiety and depression. Again, no cause and effect relation has been established, and objective signs of damage have not been in evidence.

Puzzlingly, a plausible mechanism of action has not been found for the longer term or chronic effects attributed to organophosphate exposure. The long term effects are not associated with cholinesterase inhibition, the acute toxic mechanism, in any discernible or direct way. It is possible, speculatively, to postulate some "shadow" effect of cholinesterase inhibition, or some other unknown mechanism of the agent or some contaminant, but the investigations to date have been elusive and discouraging of the existence of such mechanisms.

Conclusion

The foregoing sections describe a substantive sample of occupationally and environmentally ascribed complaints of uncertain origins. Others of no less importance are noted here but have not otherwise been selected for discussion.

To a greater or lesser extent they posit scientific problems associated with differentiating between hazard (innate adverse characteristics) and risk (the likelihood of them happening). Ill understood by society, the difference between hazard and risk is, or seems to be, being marginalised in a society where such issues are now more dominantly subject to perception and the precautionary principle (where ultimately hazard equals risk). The natural course of issues of the type discussed in this chapter is often a cycle of initial concern, resistance, disturbance, investigation, assimilation, and exhaustion. This is shown in the figures, which examine the epidemiological time course of investigations into the nuclear installations issue discussed earlier, and soft tissue sarcoma associated with herbicide application, moving from the sentinel observation towards regression to the mean.

To resolve such issues effectively in the altered perceptual framework of the society in which they flourish probably needs some fundamental reordering of current "expert" and "authoritarian" approaches. The models that have been created to "understand" these issues, while dictated by common sense

Sheep dipping

Claims that long term effects of exposure to organophosphates leading to a variety of chronic syndromes remain unsubstantiated, both epidemiologically and toxicologically

Other environmental ascribed complaints

- Oestrogenic modulators in water and food chains
- Mercury dental amalgams
- Pesticides residues in foodchains

Further reading

- National Radiological Protection Board. *Electromagnetic Fields and the Risk of Cancer.* www.nrpb.org.uk
- Committee on Medical Aspects of Radiation in the Environment (COMARE) reports, available from www.doh.gov.uk/comare/comare.htm
- Hyams KC, Wignall FS, Roswell R. War syndromes and their evaluation from the US Civil War to the Persian Gulf War. *Ann Intern Med* 1996;126:398-405
- Eberlein-Konig B, Przbilla B, Kuhnl P, Golling G, Gebefugi I, Ring I. Multiple chemical sensitivity (MCS) and others: allergological, environmental and psychological investigations in individuals with indoor related complaints. *Int J Hyg Environ Health* 2002;205:213-20
- Campbell D, Cox D, Crum J, Foster K, Riley H. Later effects of the grounding of tanker Braer on health in Shetland. *BMJ* 1994;309:773-4
- David AS. The legend of Camelford: medical consequences of a water pollution accident. *J Psychosom Res* 1995;39:1-9
- Mayon-White RT. How should another Camelford be managed? *BMJ* 1993;307(6901):398-9
- Health and Safety Executive. *Sick building syndrome: a review of the evidence on causes and solutions.* HSE Contract Report No 42/1992. Sudbury: HSE Books, 1992
- Elliott P, Briggs D, Morris S, de Hoogh C, Hurt C, Jensen TK, et al. Risk of adverse birth outcomes in populations living near landfill sites. *BMJ* 2001;323:363-8
- Health effects in relation to landfill sites. IEH report (not published in full). Dept of Health 1999 http://www.doh.gov.uk/lanl.htm
- Institute of Occupational Health. *Symptom reporting following occupational exposure to organophosphate pesticides in sheep dip.* HSE Contract Research Report 371/2001. Sudbury: HSE Books, 2001
- Spurgeon A, Gompertz D, Harrington JM. Modifiers of non-specific symptoms of occupational and environmental syndromes. *Occup Environ Med* 1996;83:361-6

and observation, are themselves authoritarian rather than consensual.

A consensual approach (that is, a process widely accepted, understood, and supported) has been attempted in the United States in dealing, practically and scientifically, with a group known as "downwinders." These were people exposed to downwind deposition of radioiodine released in north west United States in the 1940s. Features of the process are similar to those perceived to have prevented escalation of the Braer disaster but are more rigorously structured. The time, communications, and effort costs are large; the outcomes so far, imperfect.

If attribution and handling these problems in a public health context is difficult then so can be dealing with the "patients"/victims and the "worried well."

Therapeutic handling of sufferers, regardless of the putative source of their illness, usually includes accepting that there is a problem, separating out cause, symptomatology, and illness behaviour, and treating the latter by psychological techniques—an approach similar to that adopted by pain therapists.

The photograph of sheep dipping is used with permission from The University of Queensland Library, Fryer Library. Hume Family Papers. UQFLIO. Photograph album vol 8. Sheep dipping–Yandilla, (1890s). The figure showing the distribution of UK landfill sites is adapted from Eliott P, et al. *BMJ* 2001; 323:363–8. The photograph of Gulf war soldiers is reproduced with permission from Professor Simon Wessely, Academic Department of Psychological Medicine, Guy's, King's and St Thomas's School of Medicine, and Institute of Psychiatry, London.

Appendix I

Features of some important occupational zoonoses

Disease and infectious agent	Main animal reservoir	Workers at risk	Features
Brucellosis Undulant fever *Brucella abortus* *B. melitensis* *B. suis* *B. canis* **Prescribed disease** **RIDDOR reportable**	Cattle Sheep and goats Pigs Dogs	Farmers Butchers Abattoir staff Vets	**Distribution:** Worldwide, especially North and East Africa, Middle East, and Latin America **Mode of acquisition:** Direct contact with infected animals, or ingestion of contaminated milk or dairy products **Incubation period:** Variable, usually 5-60 days **Clinical features**: Acute or insidious onset with intermittent fever, fatigue, arthralgia, and localised suppurative infection of organs. Splenomegaly and lymphadenopathy occurs in about 15% of cases. Neurological symptoms may occur acutely. In the chronic form symptoms include depression, fatigue, and arthritis **Treatment:** Doxycycline with rifampicin or streptomycin Immunisation possible for cattle, but not suitable for humans
Cryptosporidiosis *Cryptosporidium parvum*	Cattle Sheep and goats	Farmers Vets	**Distribution:** Worldwide **Mode of acquisition:** Faeco-oral; ingestion of oocysts excreted in human or animal faeces **Incubation period**: 1-12 days, oocysts (the infectious stage) appear in stool at onset of symptoms, and continue to be excreted in stool for several weeks after symptoms resolve. Oocysts may remain infective for up to 6 months in a moist environment **Clinical features:** Often infection is asymptomatic. Commonest symptom is diarrhoea, often associated with abdominal cramps. Most immunocompetent people will improve within 30 days. Immunocompromised individuals may have severe and protracted illness **Treatment:** Supportive
***Escherichia coli* O157** **Notifiable**	Cattle Sheep and goats Deer Horses	Farmers Healthcare staff	**Distribution:** Worldwide **Mode of acquisition:** Ingestion of contaminated food, direct contact with infected animals, direct person to person spread, and waterborne **Incubation period**: 1-9 days, haemolytic uraemic syndrome may follow after a further 5-10 days **Clinical features:** Asymptomatic, diarrhoeal illness, haemorrhagic colitis, haemolytic uraemic syndrome in up to 10% of infected patients (particularly in children), and thrombotic thrombocytopaenic purpura. Infectious dose probably < 100 organisms, and case fatality 3-17% **Treatment:** Nil specific
Erysipeloid *Erysipelothrix rhusiopathiae*	Fish Wild or domestic animals	Fishermen Butchers Fish handlers Poultry workers Vets	**Distribution:** Worldwide **Mode of acquisition:** Direct contact with infected animal **Clinical features:** Localised cutaneous skin infection/cellulitis with violaceous tinge (fishmonger's finger). Occasionally fever, articular pain, rarely septicaemia and endocarditis. Usually self-limiting **Treatment:** Penicillin, cephalosporins, erythromycin, or tetracycline
Histoplasmosis *Histoplasma capsulatum*	Chicken Bats	Poultry workers	**Distribution:** Americas, Africa, East Asia, Australia—rare in temperate climates **Mode of acquisition:** Inhalation of airborne conidia **Incubation period:** Generally within 3-17 days **Clinical features:** (a) asymptomatic, (b) acute benign respiratory, (c) acute disseminated disease, (d) chronic disseminated disease, (e) chronic pulmonary disease **Treatment:** Itraconazole or ketoconazole for immunocompetent patients with indolent non-meningeal infection. Amphoterocin for those with fulminant or severe infections
Hydatid disease (Tapeworms of genus *Echinococcosis* *E. granulosus* and *E. multilocularis*) **Prescribed disease** **RIDDOR reportable**	Dogs Sheep—intermediate host Sylvatic hosts for *E. multilocularis*	Shepherds Farmers	**Distribution:** Worldwide except Antarctica **Mode of acquisition:** Hand to mouth transfer of eggs after association with infected dogs or through contaminated food, soil, water, or fomites **Incubation period:** Months to years **Clinical features:** Usually asymptomatic until cysts cause noticeable pressure effects; symptomology will depend on size and location of cysts. Eosinophilia common **Treatment:** Surgical resection of cysts combined with albendazole Alveolar hydatid disease (caused by *E. multilocularis*) is usually fatal if not treated

continued

Appendix I continued

Disease and infectious agent	Main animal reservoir	Workers at risk	Features
Listeriosis *Listeria monocytogenes*	Cattle, sheep, and other domestic and wild animals	Farmers Vets	**Distribution:** Worldwide **Mode of acquisition:** Mostly foodborne (soft cheeses, etc.), but also nosocomial, and direct contact with infected animals or aborted animal fetuses **Incubation period:** Variable, but 3-70 days. Infected individuals may shed organism in stool for several months **Clinical features:** General malaise or flu-like symptoms. In pregnant women infection may lead to abortion, intrauterine death, or neonatal sepsis. Immunocompromised individuals may suffer from meningoencepahlitis **Treatment:** Amoxicillin and gentamicin
Lyme disease *Borrelia burgdorferi* **RIDDOR reportable**	Wild rodents Deer	Shepherds Farmers Foresters Outdoor work	**Distribution:** United States, Canada, Europe, former Soviet Union, China, Japan **Mode of acquisition:** Tickborne—*ixodes scapularis, pacificus, ricinis,* and *persulcatus* **Incubation period:** Erythema migrans generally occurs within 7-10 days after tick bite, transmission of *B. burgdorferi* unlikely within 48 hours of tick attachment, therefore prompt removal of tick essential **Clinical features:** Initially erythema migrans (60-80%), associated lymphadenopathy, general malaise, and arthralgia. Aseptic meningitis, cranial nerve lesions, myopericarditis, AV block, cardiomegaly, and arthritis may occur up to 2 years after infection **Treatment:** Penicillin and tetracyclines Vaccine currently available in United States
Newcastle disease *Paramyxovirus*	Domesticated and wild birds	Poultry workers Pet shop staff Vets	**Distribution:** Rare in United Kingdom, occasional outbreaks in import quarantines **Mode of acquisition:** Direct contact with eyes or inhalation **Clinical features:** Mild systemic illness with conjunctivitis **Treatment:** Nil
Nipah virus *Paramyxovirus*	Natural hosts—possibly fruit bats Pigs	Pig farmers Abattoir staff	**Distribution:** South East Asia **Mode of acquisition**: Direct contact with infected blood, body fluids, or tissue **Incubation period**: 4-18 days **Clinical features**: Influenza type symptoms with severe headache, fever, respiratory symptoms, encephalitis. Death occurs in about 50% of those with symptoms **Treatment**: Supportive treatment; ribavarin has been used but effectiveness is uncertain Classified as a Hazard Group 4 agent
Orf *Parapoxvirus* **Prescribed disease**	Sheep and goats	Farm workers Abattoir staff Vets	**Distribution:** Worldwide **Mode of acquisition:** Direct contact with mucous membranes of infected animals **Incubation period:** 3-7 days **Clinical features**: Solitary maculopustular lesion surrounded by erythematous rim. Lesion dries, and crust detaches after 6-8 weeks with no persisting scar. With secondary bacterial infection, cellulitis and regional lymphadenitis occur **Treatment:** Nil
Psittacosis **Avian chlamydiosis** **Ornithosis** *Chlamydia Psittaci* **Prescribed disease** **RIDDOR reportable**	Waterfowl Pheasants Pigeons Psittacine birds	Poultry workers Pet shop staff Vets	**Distribution:** Worldwide **Mode of acquisition:** Inhalation of aerosols contaminated by infected avian faeces or fomites **Incubation period:** 1-4 weeks **Clinical features:** Fever, headache, myalgia, respiratory symptoms. (non-productive cough). Respiratory symptoms are often disproportionately mild when compared with chest *x* ray findings. Complications include encephalitis, myocarditis, and Stephens-Johnson syndrome **Treatment:** Tetracyclines or erythromycin
Ovine enzootic abortion **Prescribed disease** **RIDDOR reportable**	Sheep		Ovine strains can cause severe a septicaemic illness with intrauterine death in pregnant women. Maternal death due to disseminated intravascular coagulation may also occur. Women who are or may be pregnant should avoid exposure to sheep, particularly during lambing
Q Fever *Coxiella burnetii* **Prescribed disease** **RIDDOR reportable**	Sheep and goats Cattle Cats Dogs Wild rodents	Sheep workers Farmers Meat workers Dairy workers Abattoir staff Vets	**Distribution:** Worldwide **Mode of acquisition**: Inhalation of airborne organism, direct contact with infected animals or products **Incubation period:** 2-3 weeks **Clinical features:** Fever, retrobulbar headache, general malaise, atypical pneumonia. Occasionally, acute hepatitis. Chronic symptoms (months or years after original infection) resulting in endocarditis can occur on prosthetic or abnormal valves

continued

Appendix I continued

Disease and infectious agent	Main animal reservoir	Workers at risk	Features
			Treatment: Tetracyclines. Endocarditis will require specialist advice with combination therapy A vaccine is available in some countries for at-risk workers
Ringworm Various species of genera *Trichophyton, Microsporum Epidermophyton*	Dogs Cattle Cats	Vets Farmers	**Distribution:** Worldwide **Mode of acquisition**: Direct skin-to-skin contact **Incubation period:** Variable but usually 3-5 days for infection to become established and 2-3 weeks for symptoms to manifest **Clinical features:** Depends on site and causative agent, but *T. verrucosum* (from cattle) may produce large pustular lesions (kerions). Lesions on trunk or legs consist of prominent red margin with scaly central area **Treatment**: Mild infection responds to topical antifungals. Oral antifungals such as griseofulvin or terbinafine may be necessary when topical therapy fails
Streptococcus suis **Prescribed disease** **RIDDOR reportable**	Pigs	Pig workers Pork processors	**Distribution:** Worldwide **Mode of acquisition:** Direct contact with infected pigs or pork **Clinical features:** Primary skin infection with surrounding erythema and associated septicaemia and meningitis. Sequelae include ataxia and deafness in those with meningitis. Case fatality is extremely high in asplenics. Arthritis, pharyngitis, and diarrhoea may also occur **Treatment:** Penicillin
Toxoplasmosis *Toxoplasma gondii*	Cats	Farm workers Vets	**Distribution:** Worldwide **Mode of acquisition:** Ingestion of undercooked infected meat, contact with contaminated soil, contact with infected animals **Incubation period:** 5-20 days **Clinical features:** Mostly asymptomatic, but some have glandular fever type symptoms. Primary infection in pregnancy may result in fetal infection, abortion, intrauterine death, chorioretinitis, hepatomegaly, hydrocephalus, and mental retardation. Cerebral toxoplasmosis may occur particularly in the immunocompromised. Reactivation of latent infection may also occur **Treatment:** Not routine for uncomplicated acute infection in healthy immunocompetent adults. For toxoplasmic encephalitis a combination of pyrimethamine and sulphadiazine or pyrimethamine with clindamycin or clarithromycin, but expert advice should be sought. Spiramycin may reduce risk of transmission of maternal infection to fetus

Appendix II

Important occupationally acquired infections from human sources

Disease and infectious agent	Features
Measles *Paramyxovirus* **Notifiable disease** **Immunisation available**	**Distribution:** Although the incidence decreased after introduction of vaccination. In the United Kingdom, due to unsubstantiated concerns regarding the combined measles, mumps, and rubella vaccine, immunisation rates have dropped and outbreaks are predicted **Mode of acquisition:** Airborne by droplet spread or direct contact with nose and throat secretions. Measles is one of the most highly communicable diseases **Incubation period**: 7-18 days. **Communicability**: 1 day before the prodromal period to 4 days after the appearance of rash **Clinical features:** Prodromal fever, conjunctivitis, coryza, and Koplik's spots on buccal mucosa. Red maculopapular facial rash starts on day 3-4, and then spreads to trunk and limbs. Complications include pneumonia and encephalitis. Subacute sclerosing panencephalitis is a rare late and fatal complication developing several years after initial infection **The decrease in vaccine uptake in the United Kingdom will mean that non-immune healthcare workers are at high risk of nosocomial infection, but currently there is no consistent screening policy to identify those at risk. In the United States, all non-immune healthcare workers are identified at pre-employment and offered immunisation if non-immune** Human normal immunoglobulin (HNIG) can be offered to those who are non-immune and have compromised immunity
Meningococcal infection *Neisseria menigitidis* **Notifiable disease** **Vaccines available against serogroups A, C, W135 and Y**	**Distribution:** Worldwide there are 13 serogroups; in Europe serogroups B and C predominate. About 10% of the population are asymptomatic carriers **Mode of acquisition:** Person-to-person through respiratory droplets and direct contact with nose and throat secretions. Infectivity is relatively low and transmission requires prolonged close contact **Incubation period:** 2-10 days. **Communicability:** Patients are generally not infectious within 24 hours of antibiotic treatment **Clinical features:** Symptoms of meningitis. The appearance of a petechial rash signifies septicaemia **Healthcare personnel are rarely at risk therefore routine immunisation not indicated. Only intimate contact with infected patients—for example, mouth-to-mouth resuscitation would warrant antibiotic prophylaxis**
Fifth disease **Erythema infectosum** *Parvovirus B19*	**Distribution:** Worldwide, common in childhood **Mode of acquisition:** Person-to person by droplet spread. Rarely by contaminated blood products. It is highly infectious **Incubation period:** 4-20 days. **Communicability:** From 7 days before the appearance of rash until onset of rash. In aplastic crises, infectivity may last for up to a week after the rash appears. In the immunosuppressed with severe anaemia, infectivity may last for months or years **Clinical features:** Initially fever that lasts until rash appears. The rash is maculopapular and generally on the limbs. The cheeks often have a "slapped cheek" appearance. Illness is mild in immunocompetent individuals, although sometimes, persistent joint pain may occur. In those with haemoglobinopathies, transient aplastic crises may occur, and in the immunosuppressed, red cell aplasia and chronic anaemia may occur. Infection in the first 20 weeks of pregnancy can cause hydrops fetalis and fetal loss **Pregnant women <21/40, immunocompromised individuals or those with heamoglobinopathies who have a significant contact with an infected healthcare worker in the 7 days before onset of rash will need further follow-up. In the case of immunocompromise the administration of intravenous immunoglobulin may be considered**
Rubella **Notifiable disease** **Immunisation available**	**Distribution:** Rare in most countries in Western Europe due to vaccination programmes **Mode of acquisition:** Direct person-to-person contact by respiratory droplets **Incubation period:** 2-3 weeks **Communicability:** 1 week before onset of rash to about 4 days later **Clinical features:** Generally a mild fever with sore throat and conjunctivitis precedes a macular rash. Persistent joint infection occasionally occurs, but complete and rapid recovery usual. The main importance clinically is the risk of congenital rubella syndrome
Scabies *Sarcoptes scabiei var. hominis*	**Distribution:** Worldwide **Mode of acquisition:** Transfer of parasites by direct contact with infested skin **Incubation period:** There may be no sign of infection for 2-4 weeks after exposure, although re-exposure may result in rash within a few days **Communicability:** Remains infectious until it is treated **Clinical features:** Rash which is variable (pimples, vesicles, and nodules), burrows may be seen in finger webs, and itching particularly at night. If there is impaired immunity, large numbers of mites may present (Norwegian Scabies) **If healthcare staff are infected they can return to work once treatment is completed**

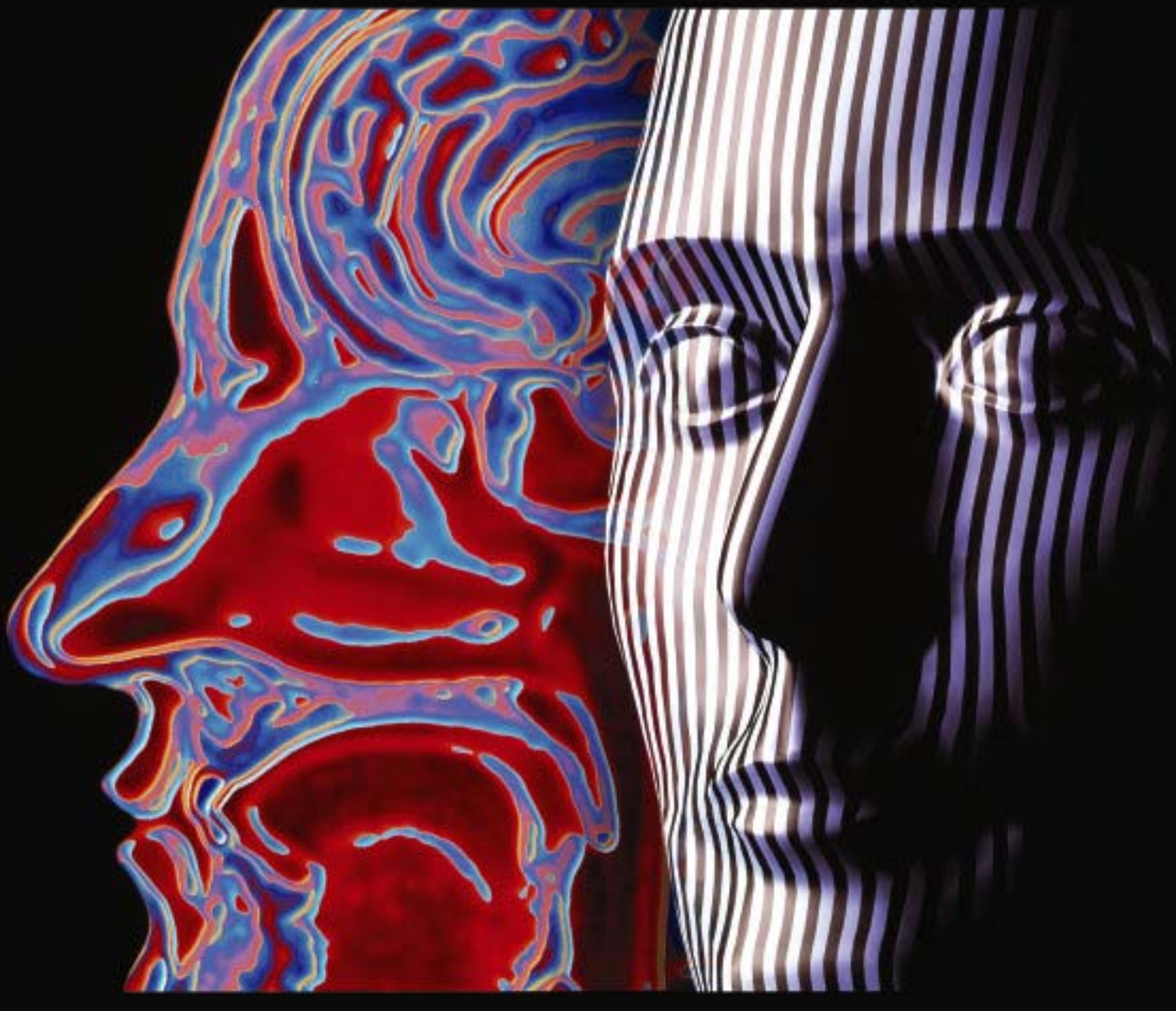
continued

Appendix II continued

Staphylococcus aureus	**Distribution:** Worldwide, but highest rates of resistant strains are in countries with liberal infection control policies such as Japan and Korea. It is carried as a skin commensal at any one time by about 30% of the population. Strains resistant to Penicillinase stable β-lactams are referred to as methicillin resistant *staphylococcus aureus* (MRSA). Recent additional problems include the emergence of resistance to mupirocin, the mainstay of treatment of skin or nasal carriage, and case reports of intermediate-level resistance to vancomycin (VISA) in Japan, France, and United States
Disease and infectious agent	**Features**
	Mode of acquisition: The significance of MRSA is that the organism colonises the skin, nose, and throat of both patients and healthcare staff, spreads readily by direct contact, and hence is an important cause of hospital acquired infections. While patients are usually responsible for spread of infection, the introduction of MRSA into unaffected areas by colonised staff is well documented, and staff hands are an important route of cross-infection **Incubation period:** 4-10 days, but disease may not occur until several months after colonisation **Clinical features:** Infection may cause both trivial and deep-seated infections; particular problems include infected bedsores or surgical wounds **Control of MRSA is therefore essential to patient care, and relies on scrupulously applied infection control programmes and stringent antibiotic policies**
Viral haemorrhagic fevers Lassa, Ebola, Marburg, and Crimean/Congo fevers Notifiable diseases	**Distribution:** Africa, South America, Middle East, and Eastern Europe **Mode of acquisition:** Main concern is that of potential secondary infection in healthcare workers as a result of accidental exposure to infected blood or body fluids **Incubation period:** 3-21 days **Clinical features:** Initial symptoms include general malaise, fever, headache, and muscle and joint pain. Obvious bleeding occurs at a later or terminal stage **In the England and Wales, the Advisory Committee on Dangerous Pathogens provides guidelines on response to a suspected case. Patients at moderate or high risk should be admitted to special isolation facilities, and strict infection control is necessary. For close contacts of high risk cases, daily surveillance for 21 days from the last possible exposure date is necessary**

Index

ABC OF PSYCHOLOGICAL MEDICINE



Edited by Richard Mayou, Michael Sharpe and Alan Carson

BMJ Books

ABC OF PSYCHOLOGICAL MEDICINE

ABC OF PSYCHOLOGICAL MEDICINE

Edited by

RICHARD MAYOU
Professor of Psychiatry, University of Oxford, Warneford Hospital, Oxford

MICHAEL SHARPE
Reader in Psychological Medicine, University of Edinburgh

and

ALAN CARSON
Consultant Neuropsychiatrist, NHS Lothian and Honorary Senior Lecturer, University of Edinburgh

BMJ Books

First published in 2003

by BMJ Books, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0 7279 1556 8

Typeset by Newgen Imaging Systems and BMJ Electronic Production
Printed and bound in Spain by GraphyCems, Navarra

Cover image depicts computer artwork of a face patterned with vertical lines with a magnetic resonance imaging (MRI) scan in the background. The MRI scan allows the internal features of the head to be seen. At the centre is the nasal cavity (red), and above that is the front part of the brain (blue and red). This region of the brain is part of the cerebrum, and is concerned with conscious thought, personality and memory. With permission from Alfred Pasieka/Science Photo Library.

Contents

Contributors

Christopher Bass
Consultant, Department of Psychological Medicine, John Radcliffe Hospital, Oxford

Michael Boyle
General Practitioner, Linlithgow Health Centre, Linlithgow, West Lothian

Tom Brown
Consultant Psychiatrist, St John's Hospital at Howden, Livingston, West Lothian

Alan Carson
Consultant Neuropsychiatrist, NHS Lothian and Honorary Senior Lecturer, University of Edinburgh

Andrew Farmer
Senior Research Fellow, Department of Public Health and Primary Care, University of Oxford

Linda Gask
Reader in Psychiatry, University of Manchester

Russell E Glasgow
Senior Scientist, AMC Cancer Research Center, Denver, Colorado, USA

Elspeth Guthrie
Professor of Psychological Medicine and Medical Psychotherapy, School of Psychiatry and Behavioural Sciences, University of Manchester

Allan House
Professor of Liaison Psychiatry, Academic Unit of Psychiatry and Behavioural Sciences, School of Medicine, University of Leeds

Michael Von Korff
Senior Investigator, Center for Health Studies, Group Health Cooperative of Puget Sound, Seattle, WA, USA

Laurence Leaver
General Practitioner, Jericho Health Centre, Oxford

Una Macleod
Lecturer in General Practice, Department of General Practice, University of Glasgow

Chris J Main
Head of the Department of Behavioural Medicine, Hope Hospital, Salford

Stephanie May
General Practitioner, Stockwell Group Practice, Stockwell Road, London

Richard Mayou
Professor of Psychiatry, University of Oxford, Warneford Hospital, Oxford

Robert Peveler
Professor of Liaison Psychiatry, University of Southampton

Jonathan Price
Clinical Tutor in Psychiatry, Department of Psychiatry, University of Oxford

Gary Rodin
Professor of Psychiatry, University of Toronto, Canada

Michael Sharpe
Reader in Psychological Medicine, University of Edinburgh

Dan Stark
Specialist Regsitrar in Medical Oncology, Academic Unit of Oncology, St James's University Hospital, Leeds

David Thompson
Professor of Medicine, Section of Gastrointestinal Science, Hope Hospital, Salford

Tim Usherwood
Professor in General Practice, University of Sydney, NSW, Australia

Craig A White
Macmillan Consultant in Psychosocial Oncology, Ayrshire and Arran Primary Care NHS Trust

David Wilks
Consultant in Infectious Diseases, Western General Hospital, Edinburgh

Amanda C de C Williams
Senior Lecturer in Clinical Health Psychology, Guy's, King's, and St Thomas's School of Medicine, University of London

Preface

Psychological medicine has a long history. Until the development of pharmacological and other specific treatments, it was a mainstay of a physician's practice. Since then the successes of biomedical theory during the 20th century have led to a loss of interest in the psychological aspects of medicine and core clinical skills have sometimes been neglected. Although many modern doctors are comfortable with the latest advances in molecular medicine, they lack confidence in applying similar intellectual rigour to the psychological problems of their patients. These deficiencies are particularly apparent in the management of patients with chronic disease and of patients whose symptoms seem out of proportion to disease pathology.

Accumulating research evidence now clearly shows that psychological variables make a substantial contribution to the outcome of most common medical conditions. The identification of problems, appropriate formulation and the implementation of appropriate treatment results in not only better outcomes for patients but also in greater satisfaction for the doctors treating them. A rediscovery of the psychological aspects of medicine is underway.

This *ABC of psychological medicine* is a practical and evidence based overview of the psychological aspects of medical practice. It aims to guide practitioners and to provide them with not only relevant information but also an intellectual structure for assessing and managing their patients. The emphasis is on day to day practice and problems rather than psychological theory. The book assumes knowledge of medical assessment, investigation, and treatment.

The opening three chapters describe general principles within which individual assessment and treatment can be formulated. They include the clinical examination and the initiation of treatment but also a critique of the structure within which care is delivered, which can often be as critical as the individual's consultation. The following three chapters describe the core skills of psychological medicine: the assessment and management of anxiety, depression, and functional somatic symptoms. The remaining chapters then describe how these skills are transferred and adapted in specific situations including the care of patients with cancer, trauma, musculoskeletal pain, fatigue, chest pain, abdominal pain, and delirium. This list is not comprehensive but provides a range of examples that should help the reader to adapt the principles to their own practice.

Psychological medicine is an extension of existing clinical knowledge and skills. Indeed many practitioners will recognise it as a formalisation of the medicine they have been practising for many years. We hope that this book will both engage the curiosity and interest of those to whom the subject matter is novel, and encourage and inform those who already understand and apply its principles.

Richard Mayou, Michael Sharpe, Alan J Carson, 2002

Introduction

It is becoming increasingly clear that we can improve medical care by paying more attention to psychological aspects of medical assessment and treatment. The study and practice of such factors is often called psychological medicine. Although the development of specialist consultation-liaison psychiatry (liaison psychiatry in the United Kingdom) and health psychology contribute to psychological medicine, the task is much wider and has major implications for the organisation and practice of care. This book aims to explain some of those implications.

Disorders that are traditionally, and perhaps misleadingly, termed "psychiatric" are highly prevalent in medical populations. At least 25-30% of general medical patients have coexisting depressive, anxiety, somatoform, or alcohol misuse disorders.[1] Several factors account for the co-occurrence of medical and psychiatric disorders. First, a medical disorder can occasionally be a cause of the psychiatric disorder (for example, hypothyroidism as a biological cause of depression). Second, cardiovascular diseases, neurological disorders, cancer, diabetes, and many other medical diseases increase the risk of depression and other psychiatric disorders. Such so called comorbidity is common, but its causal linkage with psychological conditions remains poorly understood. A third factor is coincidence—common conditions such as hypertension and depression may coexist in the same patient because both are prevalent.

Another reason for psychological medicine is the prevalence of symptoms that are unexplained by disease. Although physical symptoms account for more than half of all visits to doctors, at least a third of these symptoms remain medically unexplained.[2,3] This phenomenon is referred to as somatisation—the seeking of health care for somatic symptoms that suggest a medical disorder but represent instead an underlying depressive, anxiety, or somatoform disorder. Most patients with these mental disorders preferentially report somatic rather than emotional symptoms. Further, there are the common but poorly understood symptom syndromes such as fibromyalgia, irritable bowel syndrome, and chronic fatigue syndrome, for which the relative contributions of mind and body are not yet elucidated.[4]

Psychological medicine is important in the management of all these problems; both psychotropic medications and cognitive behavioural treatments have proved effective in the treatment of common physical symptoms and syndromes in numerous studies in general practice.[5,6] Although such treatments have traditionally been considered "psychiatric", they are also beneficial in patients without overt psychiatric disorders. Countries on both sides of the Atlantic have a long way to go in developing psychological medicine, the chasm in America between medical and psychiatric care is particularly deep. The "carve out" or organisational separations of mental health services in the managed care systems in the United States is one example of how ingrained the dualism of mind and body still is and of the reconciliation that must occur.

Psychological medicine does not mean relabelling all such patients as "psychiatric". Many patients prefer to have these problems regarded as "medical" and conceptualised in terms of a neurotransmitter imbalance or a functional bodily disturbance.[7] Concomitant psychological distress is best framed in terms of being a consequence rather than a cause of persistent physical symptoms. Premature efforts to reattribute somatic complaints to psychological mechanisms may be perceived by the patient as rejection. A more aetiologically neutral but psychologically sophisticated approach that initially focuses on symptomatic treatment, reassurance, activation, and restoration of function has proved more effective.[8]

There are better alternatives than simply to relegate such problems to the province of specialist psychiatry. One is to train general practitioners to diagnose and treat common "psychiatric" disorders.[9] Although treatment with psychotropic medication is their most feasible option, general practitioners can also be trained to deliver other psychological treatments. A second option is to use nurses or social workers with specialised training who can work with general practitioners or psychiatrists to manage medication as well as deliver psychotherapies and behavioural interventions. A third model is collaborative care, where the general practitioner's management is augmented but not replaced by visits to a psychiatrist, often on site in the general practitioner's surgery. Stepped care provides an overall principle of management whereby patients only move on to more complex and expensive forms of care where simpler management by the healthcare team is either ineffective or inappropriate. Most studies have been conducted in general medical practices, but patients seen by medical specialists also warrant attention.[3]

Psychological medicine may also be delivered in innovative ways. Promising data exist for behavioural interventions conducted outside the doctor's office, including case management by telephone, cognitive behavioural therapy given through a computer, bibliotherapy—self study by patients—and home visits (for example, for chronic fatigue syndrome).

Medical treatment that integrates a psychological approach has been shown to improve patient outcomes. The benefits of treating common physical symptoms and psychological distress effectively in medical patients include not only improved quality of life and social and work functioning, but also greater satisfaction on the part of patient and doctor and reduced use of healthcare services.[2]

What do we need to do? Better detection of these problems need not be time consuming. For example, screening for depression may require as few as one or two questions. Optimal management of patients with persistent physical symptoms and common mental disorders may require longer or more frequent visits to a doctor, help in educating and following up patients by a nurse case manager, other system changes, and specialist mental health consultations for more complex cases.[10] The competing demands of general practice must be explicitly addressed if we are to enable the general practitioner to practise psychological medicine effectively.[11]

Yet this approach is no different to what is also required for many chronic medical disorders such as diabetes, asthma, and heart disease, for which it has been proved that care in concordance with guidelines requires appreciable reorganisation of medical services.[12]

Neither chronic "medical" nor "psychiatric" disorders can be managed adequately in the current environment of general practice, where the typical patient must be seen in 10–15 minutes or less. The quick visit may work for the patient with a common cold or a single condition, such as well controlled hypertension, but will not suffice for the prevalent and disabling symptoms and disorders comprising psychological medicine. Evidence based treatments exist. Using them in a way that is integrated with general medical care will improve both patients' physical health and their psychological wellbeing.

Kurt Kroenke*
Professor of Medicine, Department of Medicine,
Regenstrief Institute for Health Care,
Indianapolis, IN, USA

* KK has received fees for speaking and research from Pfizer and Eli Lilly.

References

1. Ormel J, Von Korff M, Ustun TB, Pini S, Korten A, Oldehinkel T. Common mental disorders and disability across cultures: results from the WHO collaborative study on psychological problems in general health care. *JAMA* 1994;272:1741-48.
2. Kroenke K. Studying symptoms: sampling and measurement issues. *Ann Intern Med* 2001;134:844-55.
3. Reid S, Wessely S, Crayford T, Hotopf M. Medically unexplained symptoms in frequent attenders of secondary health care: retrospective cohort study. *BMJ* 2001;322:1-4.
4. Wessely S, Nimnuan C, Sharpe M. Functional somatic syndromes: one or many? *Lancet* 1999;354:936-9.
5. O'Malley PG, Jackson JL, Santoro J, Tomkins G, Balden E, Kroenke K. Antidepressant therapy for unexplained symptoms and symptom syndromes. *J Fam Pract* 1999;48:980-90.
6. Kroenke K, Swindle R. Cognitive-behavioral therapy for somatization and symptom syndromes: a critical review of controlled clinical trials. *Psychother Psychosom* 2000;69:205-15.
7. Sharpe M, Carson A. "Unexplained"somatic symptoms, functional syndromes, and somatization: do we need a paradigm shift? *Ann Intern Med* 2001;134:926-30.
8. Von Korff M, Moore JC. Stepped care for back pain: activating approaches for primary care. *Ann Intern Med* 2001;134:911-17.
9. Kroenke K, Taylor-Vaisey A, Dietrich AJ, Oxman TE. Interventions to improve provider diagnosis and treatment of mental disorders in primary care: a critical review of the literature. *Psychosomatics* 2000;41:39-52.
10. Rubenstein LV, Jackson-Triche M, Unutzer J, Miranda J, Minnium K, Pearson ML, et al. Evidence-based care for depression in managed primary care practices. *Health Aff* 1999;18:89-105.
11. Klinkman MS. Competing demands in psychosocial care: a model for the identification and treatment of depressive disorders in primary care. *Gen Hosp Psychiatry* 1997;19:98-111.
12. Wagner EH, Austin BT, Von Korff M. Organizing care for patients with chronic illness. *Milbank Q* 1996;74:511-44.

1 The consultation

Linda Gask, Tim Usherwood

The success of any consultation depends on how well the patient and doctor communicate with each other. There is now firm evidence linking the quality of this communication to clinical outcomes.

The dual focus—Patients are not exclusively physically ill or exclusively emotionally distressed. Often they are both. At the start of a consultation it is usually not possible to distinguish between these states. It is the doctor's task to listen actively to the patient's story, seeking and noticing evidence for both physical illness and emotional distress.

Involving patients—Changes in society and health care in the past decade have resulted in real changes in what people expect from their doctors and in how doctors view patients. In addition, greater emphasis has been placed on the reduction of risk factors, with attempts to persuade people to take preventive action and avoid risks to health. Many patients want more information than they are given. They also want to take some part in deciding about their treatment in the light of its chances of success and any side effects. Some patients, of course, do not wish to participate in decision making; they would prefer their doctor to decide on a single course of action and to advise them accordingly. The skill lies in achieving the correct balance for each patient.

A comprehensive model—The "three function" model for the medical encounter provides a template for the parallel functions of the clinical interview. This is now widely used in medical schools.

Visiting the sick woman, by Quiringh Gerritsz van Brekelenkam (c 1620-68)

Three functions of the medical consultation

1 Build the relationship
- Greet the patient warmly and by name
- Active listening
- Detect and respond to emotional issues

2 Collect data
- Do not interrupt patient
- Consider other factors
- Elicit patient's explanatory model
- Develop shared understanding

3 Agree a management plan
- Provide information
- Make links
- Appropriate use of reassurance
- Negotiate behaviour change
- Negotiate a management plan

Starting the interview

Research has shown the importance of listening to patients' opening statements without interruption. Doctors often ask about the first issue mentioned by their patients, yet this may not be what is concerning them most. Once a doctor has interrupted, patients rarely introduce new issues. If uninterrupted, most patients stop talking within 60 seconds, often well before. The doctor can then ask if a patient has any further concerns, summarise what the patient has just said, or propose an agenda—"I wonder if I could start by asking you some more questions about your headaches, then we need to discuss the worries that your son has been causing you."

Detecting and responding to emotional issues

Even when their problems are psychological or social, patients usually present with physical symptoms. They are also likely to give verbal or non-verbal cues. Verbal cues are words or phrases that hint at psychological or social problems. Non-verbal cues include changes in posture, eye contact, and tone of voice that reflect emotional distress.

It is important to notice and respond to cues at the time they are offered by patients. Failure to do so may inhibit patients from further disclosures and limit the consultation to discussion of physical symptoms. Conversely, physical symptoms must be taken seriously and adequately evaluated. Several of the skills of active listening are valuable in discussing physical, psychological, and social issues with patients. These skills have been clearly shown to be linked to recognition of emotional problems when used by general practitioners.

Responding to patients' "cues"

Verbal cues
- State your observation—"You say that recently you have been feeling fed-up and irritable"
- Repeat the patient's own words—"Not well since your mother died"
- Seek clarification—"What do you mean when you say you always feel tired?"

Non-verbal cues
- Comment on your observation—"I can hear tears in your voice"
- Ask a question—"I wonder if that upsets you more than you like to admit?"

Aspects of interview style that aid assessment of patients' emotional problems

Early in the interview
- Make good eye contact
- Clarify presenting complaint
- Use directive questions for physical complaints
- Begin with open ended questions, moving to closed questions later

Interview style
- Make empathic comments
- Pick up verbal cues
- Pick up non-verbal cues
- Do not read notes while taking patient's history
- Deal with over-talkativeness
- Ask more questions about the history of the emotional problem

Active listening skills

Open ended questions—Questions that cannot be answered in one word require patient to expand
Open-to-closed cones—Move towards closed questions at the end of a section of the consultation
Checking—Repeat back to patient to ensure that you have understood
Facilitation—Encourage patient both verbally ("Go on") and non-verbally (nodding)
Legitimising patient's feelings—"This is clearly worrying you a great deal," followed by, "You have an awful lot to cope with," or, "I think most people would feel the same way"
Surveying the field—Repeated signals that further details are wanted: "Is there anything else?"
Empathic comments—"This is clearly worrying you a great deal"
Offering support—"I am worried about you, and I want to know how I can help you best with this problem"
Negotiating priorities—If there are several problems draw up a list and negotiate which to deal with first
Summarising—Check what was reported and use as a link to next part of interview. This helps to develop a shared understanding of the problems and to control flow of interview if there is too much information

Eliciting a patient's explanatory model

When people consult a doctor, they do so with explanatory ideas about their problems and with anxieties and concerns that reflect these ideas. They are also likely to have hopes and expectations concerning the care that they will receive. It is important not to make assumptions about patients' health beliefs, concerns, and expectations but to elicit these as a basis for providing information and negotiating a management plan.

People's health beliefs and behaviours develop and are sustained within families, and families are deeply affected by the illness of a family member. "Thinking family" can help to avoid difficult and frustrating interactions with family members.

Providing information

Doctors should consider three key questions when providing information to a patient:
• What does the patient already know?
• What does the patient want to know?
• What does the patient need to know?

The first question emphasises the importance of building on the patient's existing explanatory model, adding to what he or she already knows, and correcting inaccuracies. The second and third reflect the need to address two agendas, the patient's and the doctor's. In addition, it is important for the doctor to show ongoing concern and emotional support, making empathic comments, legitimising the patient's concerns, and offering support.

Negotiating a management plan

The ideal management plan is one that reflects current best evidence on treatment, is tailored to the situation and preferences of the patient, and addresses emotional and social issues. Both patient and doctor should be involved in developing the plan, although one or the other may have the greater input depending on the nature of the problem and the inclinations of the patient.

Appropriate use of reassurance

Reassurance is effective only when doctors understand exactly what it is that their patients fear and when they address these fears truthfully and accurately. Often it is not possible to reassure patients about the diagnosis or outcome of disease, but it is always possible to provide support and to show personal concern for them.

Dealing with difficult emotions: denial, anger, and fear

Denial—When patients deny the seriousness of their illness you should never be tempted to force them into facing it. The decision on how to address denial must be based on how adaptive the denial is, what kind of support is available to the patient, and how well prepared the patient is to deal with the fears that underlie the denial.

Think family

When interviewing an individual
- Ask how family members view the problem
- Ask about impact of the problem on family function
- Discuss implications of management plan for the family

When a family member comes in with patient
- Acknowledge relative's presence
- Check that patient is comfortable with relative's presence
- Clarify reasons for relative coming
- Ask for relative's observations and opinions of the problem
- Solicit relative's help in treatment if appropriate
- If patient is an adolescent accompanied by an adult always spend part of consultation without the adult present
- Never take sides

Negotiating a management plan

Ascertain expectations
- What does patient know?
- What does patient want?—Investigation? Management? Outcomes?

Advise on options
- Elicit patient's preferences

Develop a plan
- Involve patient
- Tailor preferred option to patient's needs and situation
- "Think family"

Check understanding
- Ensure that patient is clear about plan
- Consider a written summary

Advise on contingency management
- What should patient do if things do not go according to plan?

Agree arrangements for follow up and review

Reassurance is an essential skill of bedside medicine. (Hippocrates (469-399 BC), the "father of bedside medicine")

Anger—If patients or relatives become angry, try to avoid being defensive. Acknowledge the feelings that are expressed and ask about the reasons for these. Take concerns seriously and indicate that you will take appropriate action.

Fear—Many patients are frightened that they may have some serious disease. It is crucial to ensure that you have addressed what a patient is really worried about as well as checking that the patient has correctly understood what you are concerned about.

Motivation

Efforts to help people reduce alcohol consumption, stop smoking, and manage chronic illness have highlighted the importance of good interviewing skills in motivating patients to change their behaviour. This is not to say that patients no longer have the responsibility for such change, but doctors should recognise that they bear some responsibility for ensuring that patients get the best possible help in arriving at the decision to change.

Making the link between emotions and physical symptoms

Particular strategies may be needed to help people who present with physical symptoms of psychological distress but who have not made the link between these and their emotional and life problems. However, it is essential that you do not go faster than the patient and try to force the patient to accept your explanation.

Feeling understood—Ensuring that the patient feels understood is essential. It is crucial to get the patient on your side and show that you are taking his or her problems seriously. Start from the patient's viewpoint and find out what the patient thinks may be causing the symptoms, while at the same time picking up any verbal and non-verbal cues of emotional distress.

Broadening the agenda can begin when all the information has been gathered. The aim is to broaden the agenda from one where the problem is seen essentially as physical to one where both physical and psychological problems can be acknowledged. Acknowledging the reality of the patient's pain or other symptoms is essential and must be done sensitively. Summarise by reminding the patient of all the symptoms, both physical and emotional, that you have elicited and link them to life events if this is possible.

Negotiating explanations can involve various techniques. Only one or two will be appropriate for each patient, and different techniques may be useful at different times. Simple explanation is the commonest, but it is insufficient to say "Anxiety causes headaches." A three stage explanation is required in which anxiety is linked to muscle tension, which then causes pain. A similar approach can be used to explain how depression causes lowering of the pain threshold, which results in pain being felt more severely than it otherwise would be.

Once the patient and doctor have agreed that psychological distress is an important factor in the patient's illness, they can start to examine management options to address this. Even if the patient has significant physical disease, it is important to detect and manage psychological comorbidity.

Visiting the sick woman is held at the Hermitage and is reproduced with permission of Bridgeman Art Library.

Helping patients to change their behaviour

Explore motivation for change
- Build rapport and be neutral
- Help draw up list of problems and priorities
- Is problem behaviour on patient's agenda?
- If not, raise it sensitively
- Does patient consider the behaviour to be a problem?
- Do others?

Clarify patient's view of the problem
- Help draw up a balance sheet of pros and cons
- Empathise with difficulty of changing
- Reinforce statements that express a desire to change
- Resist saying why you think patient ought to change
- Summarise frequently
- Discuss statements that are contradictory

Promote resolution

If no change is wanted negotiate if, when, and how to review
- Enable informed decision making
- Give basic information about safety or risks of behaviour
- Provide results of any examination or test
- Highlight potential medical, legal, or social consequences
- Explain likely outcome of potential choices or interventions
- Get feedback from patient
- Give patient responsibility for decision

Key stages in linking somatic symptoms of emotional distress

- Helping patient to feel understood
- Broadening agenda to cover physical, psychological, and social issues
- Negotiating explanations for how physical symptoms, psychological distress, and social problems may be linked via physiological mechanisms

Evidence based summary

- The style with which a doctor listens to a patient will influence what the patient says
- Effective communication between doctor and patient leads to improved outcome for many common diseases
- Patients' compliance will be improved if the management plan has been negotiated jointly

Lang F, Floyd MR, Beine KL. Clues to patients' explanations and concerns about their illnesses—a call for active listening. *Arch Fam Med* 2000;9:222-7

Stewart MA. Effective physician-patient communication and health outcomes: a review. *Can Med Assoc J* 1995;152:1423-33

Roter D, Hall JA, Merisca R, Nordstron B, Cretin D, Svarstad B. Effectiveness of interventions to improve patient compliance: a meta-analysis. *Med Care* 1998;36:1138-61

Further reading

- Cole SA, Bird J. *The medical interview: the three function approach.* St Louis, MO: Harcourt Health Sciences, 2000
- Gask L, Morriss R, Goldberg D. *Reattribution: managing somatic presentation of emotional distress.* 2nd ed. Manchester: University of Manchester, 2000. (Teaching videotape available from Nick.Jordan@man.ac.uk)
- Usherwood T. *Understanding the consultation.* Milton Keynes: Open University Press, 1999

2 Beginning treatment

Jonathan Price, Laurence Leaver

Traditionally, the management of newly presenting patients has two stages—assessment and then treatment. However, this two stage approach has limitations. When underlying disease pathology is diagnosed there may be delays in starting effective treatment. If no disease is found reassurance is often ineffective. In both cases many patients are left feeling uncertain and dissatisfied. Lack of immediate information and agreed plans may mean that patients and their families become anxious and draw inappropriate conclusions, and an opportunity to engage them fully in their management is missed.

If simple diagnosis is supplemented with fuller explanation, patient satisfaction and outcomes are improved. This can be achieved by integrating assessment and treatment. The aim of an integrated consultation is that the patient leaves with a clear understanding of the likely diagnosis, feeling that concerns have been addressed, and knowledge of the treatment and prognosis (that is, the assessment becomes part of the treatment). This approach can be adopted in primary and secondary care and can be applied to patients with or without an obvious disease explanation for their symptoms. The integrated approach may require more time, but this is offset by a likely reduction in patients' subsequent attendance and use of resources.

This article describes principles and practical procedures for effective communication and simple interventions. They can be applied to various clinical situations—such as straightforward single consultation, augmenting brief medical care, and promoting an effective start to continuing treatment and care.

Mismatch of expectations and experiences

What patients want	What some patients get
To know the cause	No diagnosis
Explanation and information	Poor explanation that does not address their needs and concerns
Advice and treatment	Inadequate advice
Reassurance	Lack of reassurance
To be taken seriously by a sympathetic and competent doctor	Feeling that doctor is uninterested or believes symptoms are unimportant

Disease centred versus patient centred consultations

Disease centred—Doctor concentrates on standard medical agenda of diagnosis through systematic inquiries about patient's symptoms and medical history

Patient centred—Doctor works to patient's agenda, including listening and allowing patient to explain all the reasons for attending, feelings, and expectations. Decision making may be shared, and plans are explicit and agreed. Patient centred consultations need take no longer than traditional disease centred consultations

Taking time to listen to and address patients' ideas, concerns, and expectations can improve outcomes (*Charcot at the Saltpêtrière* by Luis Jimenez y Aranda, 1889, in the Provincial Museum of Art, Seville)

General principles

Integrating physical and psychological care

Somatic symptoms are subjective and have two components, a somatic element (a bodily sensation due to physiology or pathology) and a psychological element (related to thoughts and beliefs about the symptoms). Traditional management focuses only on the somatic component, with the aim of detecting and treating underlying pathology. Addressing the psychological component in the consultation as well, with simple psychological interventions, is likely to reduce distress and disability and reduce the need for subsequent specialist treatment.

Providing continuity

Seeing the same doctor on each visit increases patient satisfaction. Continuity may also improve medical outcomes, including distress, compliance, preventive care, and resource use. Problems resulting from lack of continuity can be minimised by effective communication between doctors.

Involving the patient

The psychological factors of beliefs and attitudes about illness and treatment are major determinants of outcome. Hence, strategies that increase understanding, sense of control, and participation in treatment can have large benefits. One example is written management plans agreed between doctor and patient. This approach is the basis of the Department of Health's "Expert Patient Programme," which aims to help patients to "act as experts in managing their own condition, with appropriate support from health and social care services."

Communication between doctors

- Reduce need for communication between doctors by providing continuity of care whenever possible
- Brief, structured letters are more likely to be read than lengthy, unstructured letters
- Letters from primary to secondary care should provide relevant background information and a clear reason for referral
- Letters from secondary to primary care should provide only essential information, address the needs of referrer, and outline a proposed management plan and what has been discussed with patient
- Avoid using letters for medical records purposes rather than communication
- The telephone can be a prompt and effective means of communication and is particularly useful in complex cases

Thinking "family"
Relatives' illness beliefs and attitudes are also crucial to outcome and are therefore worth addressing. Key people may be invited to join a consultation (with the patient's permission) and their concerns identified, acknowledged, and addressed. Actively involving relatives, who will spend more time with the patient than will the doctor, allows them to function as co-therapists.

Effective communication

Gaining and demonstrating understanding
Simple techniques can be used to improve communication. The first two stages of the three function approach (see previous article) are appropriate. The first stage is building a relationship in which a patient gives his or her history and feels understood. The second stage is for the doctor to share his or her understanding of the illness with the patient. In cases that are more complicated it may be most effective to add an additional brief session with a practice or clinic nurse.

Providing information for patients
Patients require information about the likely cause of their illness, details of any test results and their meaning, and a discussion of possible treatments. Even when this information has been given in a consultation, however, many patients do not understand or remember what they are told. Hence, the provision of simple written information can be a time efficient way of improving patient outcomes.

One way of providing written information is to copy correspondence such as referral and assessment letters to the patient concerned. For those not used to doing this, it may seem a challenge, but any changes needed to make the letters understandable (and acceptable) to patients are arguably desirable in any case. Letters should be clearly structured, medical jargon minimised, pejorative terms omitted, and common words that may be misinterpreted (such as "chronic") explained.

Well written patient information materials (leaflets and books) are available, as are guidelines for their development. The National Electronic Library of Health (www.nelh.nhs.uk) is a new internet resource that aims to provide high quality information for healthcare consumers and is linked to NHS Direct Online (www.nhsdirect.nhs.uk/main.jhtml). There are also many books to recommend—such as *Chronic Fatigue Syndrome (CFS/ME): The Facts* (see Further reading list). Information is most helpful if it addresses not only the nature of the problem, its prognosis, and treatment options, but also self care and ways of coping.

The assessment as treatment

Reassurance
Worry about health (health anxiety) is a common cause of distress and disability in those with and without serious disease. Reassurance is therefore a key component of starting treatment.

The first step is to elicit and acknowledge patients' expectations, concerns, and illness beliefs. This is followed by history taking, examination, and if necessary investigation. Premature reassurance (such as "I'm sure its nothing much") may be construed as the doctor not taking the problem seriously. Finally, the explanation should address all of a patient's concerns and is best based on the patient's understanding of how his or her body functions, which may differ from the doctor's.

A modest increase in consultation time, provision of written information, and perhaps the use of trained nursing staff to

Gaining understanding of patients' concerns

- Read referral letter or notes, or both, before seeing patient
- Encourage patients to discuss their presenting concerns without interruption or premature closure
- Explore patients' presenting complaints, concerns, and understanding (beliefs)
- Inquire about disability
- Inquire about self care activities
- Show support and empathy
- Use silence appropriately
- Use non-verbal communication such as eye contact, nods, and leaning forward

Showing your understanding of patients' concerns

- Relay key messages—such as, "The symptoms are real," "We will look after you," and "You're not alone"
- Take patients seriously and make sure they know it
- Don't dismiss presenting complaints, whether or not relevant pathology is found
- Explain your understanding of the problem—what it is, what it isn't, treatment, and the future. A diagram may help
- Consider offering a positive explanation in the absence of relevant physical pathology
- Reassure
- Avoid mixed messages
- Encourage and answer questions
- Share decisions
- Communicate the management plan effectively, both verbally and in writing
- Provide self care information, including advice on lifestyle change
- Explain how to get routine or emergency follow up, and what to look out for that would change the management plan

Providing information

- Invite and answer questions
- Use lay terms, and build on patient's understanding of illness wherever possible
- Avoid medical jargon and terms with multiple meanings, such as "chronic"
- Involve relatives
- Provide written material when available
- Provide a written management plan when appropriate

The complexity of reassurance

General reassurance
- To know it will be OK
- To know I will be looked after
- To know there are others like me

Reassurance about cause
- To know what it is
- To know what it is not
- To know it's not serious
  "There are several possible causes, not just cancer"
  "It's not cancer"
  "It will get better"

Reassurance about cure
- To know it can be treated
- To know it will be treated
- To know how it will be treated
- To know the complaint will go away

facilitate information giving, can all enhance doctor-patient communication and, therefore, reassurance. Although extra time and effort may be needed, it may well reduce subsequent demand on resources.

Being positive

Doctors themselves are potentially powerful therapeutic agents. There is evidence that being deliberately positive in a consultation may increase this effect. In one randomised trial, general practice patients received either a positive consultation (firm diagnosis and good prognosis) or a non-positive consultation (no firm diagnosis and uncertain prognosis). Two weeks later, the positive consultation, which was simple and brief, had improved symptoms, with a number needed to treat of four (95% confidence interval 3 to 9).

Using tests as treatment

Tests should ideally be informative and reassuring for both doctors and patients. However, there is increasing evidence that tests may not reassure some patients and may even increase their anxiety. This is most likely with patients who are already anxious about their health. When weighing the pros and cons of ordering a test, doctors should take account of the potential psychological impact on their patient (both positive and negative).

Providing explanations after negative investigation

Even when tests are reported as normal, some patients are not reassured. Such patients may benefit from an explanation of what is wrong with them, not just what is not wrong. A cognitive behavioural model can be used to explain how interactions between physiology, thoughts, and emotion can cause symptoms without pathology. Simple headache provides an analogy: the pain is real, and often distressing and disabling, but is usually associated with "stress." Diagnoses such as "tension headache" and "irritable bowel syndrome" can be helpful in reducing patients' anxiety about sinister causes for their symptoms.

Planning for the future

Maintaining and increasing activities

Sometimes patients unnecessarily avoid or reduce their activities for fear it will make their illness worse. This coping strategy magnifies disability. Planning a graded return towards normal activities is one of the most effective ways of helping such patients. A plan should specify clearly what activity, for how long, when, with whom, and how often. It is best if the plan is written down and reviewed regularly. A collaborative approach increases the chances of success.

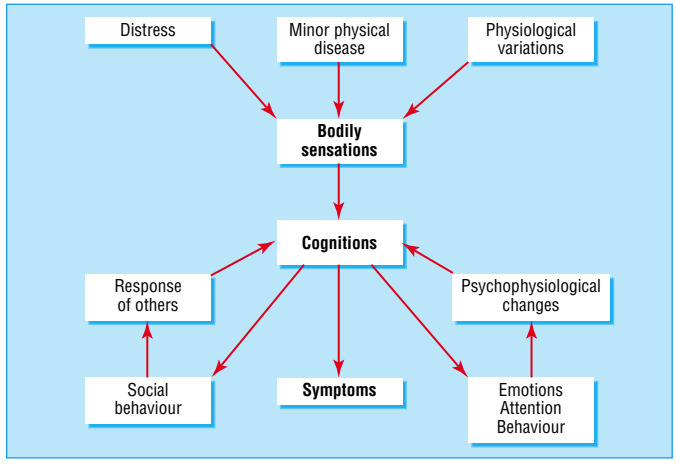
Follow up

Positively following up patients who have presented for the first time can be an effective use of time. It allows review and modification of the management plan and may be particularly effective if the same doctor is seen.

The painting of Charcot is reproduced with permission of the Wellcome Library.

Characteristics of brief psychological intervention

- Brief, single session intervention
- Suitable for more complex problems, such as in secondary care
- Delivered with or soon after clinic attendance
- Integrated with usual care
- Uses cognitive understanding of health anxiety
- Minimises negative aspects of patient experience
- Reinforces positive aspects of patient experience
- Provides explicit explanation and reassurance



A simple cognitive model of physical symptoms. A cognitive model is one in which the patient's thoughts and beliefs are seen as central to the aetiology, perception, and presentation of the problem

Evidence based summary

- The quality of communication, both in history taking and in discussing a management plan, influences patient outcome
- Patients should be encouraged to take an active role in maintaining or improving their own health, and doctors should ensure they are given the necessary information and opportunities for self management
- Reassurance involves eliciting and acknowledging patients' expectations, concerns, and illness beliefs

Coulter A, Entwistle V, Gilbert D. Sharing decisions with patients: is the information good enough? *BMJ* 1999;318:318-22

Di Blasi Z, Harkness E, Ernst E, Georgiou A, Kleijnen J. Influence of context effects on health outcomes: a systematic review. *Lancet* 2001;357:575-762

Stewart M. Effective physician-patient communication and health outcomes: a review. *Can Med Assoc J* 1995;152:1423-33

Thomas KB. General practice consultations: is there any point in being positive? *BMJ* 1997;294:1200-2

Further reading

- Balint M. *The doctor, his patient, and the illness.* Tunbridge Wells: Pitman Medical, 1957
- Department of Health. *The NHS plan—A plan for investment. A plan for reform.* London: DoH, 2000
- Campling F, Sharpe M. *Chronic fatigue syndrome (CFS/ME): the facts.* Oxford: Oxford University Press, 2000
- Department of Health. *The expert patient: a new approach to chronic disease management for the 21st century.* 2001 (www.ohn.gov.uk/ohn/people/ep_report.pdf

3 Organising care for chronic illness

Michael Von Korff, Russell E Glasgow, Michael Sharpe

A major and increasing task for health services is the management of chronic illness. Although the details of chronic illness management will depend on the illness in question, many of the principles are common to all chronic conditions.

Principles of effective management

Whatever health services may offer, most of the day to day responsibilities for the care of chronic illness fall on patients and their families. Planners and organisers of medical care must therefore recognise that health care will be most effective if it is delivered in collaboration with patients and their families. To enable patients to play an active role in their care, health services must not only provide good medical treatment but also improve patients' knowledge and self management skills. This can be done by supplementing medical care with educational and cognitive behavioural interventions. Chronic disease treatment programmes have tended to underestimate the need for this aspect of care, and, consequently, many treatment programmes have been psychologically naive and, as a result, less effective than they could have been.

Services also need to be not merely reactive to patients' requests but proactive with planned follow up. Finally, to be most efficient, interventions are best organised in a stepped fashion—that is, the more complex and expensive interventions are given only when simpler and cheaper ones have been shown to be inadequate or inappropriate.

Treating chronic conditions must involve the family

Common elements of effective chronic illness management

- Collaboration between service providers and patients
- A personalised written care plan
- Tailored education in self management
- Planned follow up
- Monitoring of outcome and adherence to treatment
- Targeted use of specialist consultation of referral
- Protocols for stepped care

Collaboration with patients and families

To win the collaboration of patients and their families, those providing care need to elicit, negotiate and agree on a definition of the problem they are working on with each patient. They must then agree on the targets and goals for management and develop an individualised collaborative self management plan. This plan should be based on established cognitive behavioural principles and on the evidence relating to the management of the chronic condition.

In order to implement collaborative care, patients and their families require access to the necessary information and services to enable them to play a full and informed role. The need for collaborative care in which patients play an active role has been highlighted in Britain with the development of the concept of the "expert patient."

Principles of collaboration

- Understanding of patients' beliefs, wishes, and circumstances
- Understanding of family beliefs and needs
- Identification of a single person to be main link with each patient
- Collaborative definition of problems and goals
- Negotiated agreed plans regularly reviewed
- Active follow up with patients
- Regular team review

Encouraging self care

Active self care is critical to the optimal management of chronic illness. Interventions to optimise self care are based on cognitive behavioural principles.

They start with an assessment of patients' attitudes and beliefs about their illness and their chosen coping behaviours. This assessment then guides the provision of information, the resolution of misunderstandings and misinterpretations, and collaborative goal setting. These are agreed between patient and members of the healthcare team.

The outcome of this initial assessment takes the form of a personal action plan, a written agreement between those delivering care and the patient. The patient keeps a copy of the plan, and the healthcare team keeps another. The plan can be written on brief, standardised forms. The plan is not static but is

The UK "expert patient" programme*

- Encouragement of self care protocols, nationally and locally
- Development of electronic and written self care material
- Training programmes, national and local
- Integrating self care into local health planning
- Nurse led telephone service (NHS Direct)

* From; Department of Health. *The expert patient: a new approach to chronic disease management for the 21st century* (www.ohn.gov.uk/ohn/people/expert)

developed over time: the initial goals and the care plan designed to achieve them are refined in view of the patient's progress and the identification of factors that are either helpful or unhelpful in achieving the desired outcome.

Active follow up
The personal action plan guides the patient's follow up contacts. Active planned follow up ensures that the plan is carried out and that modifications to it are made as needed. These steps are repeated in an iterative, ongoing, and flexible way rather than all at once in a single visit. Because the care of chronic illness is a long term process, the work of supporting self care does not need to be done all at once but can be spread over many contacts.

Individualised stepped care
Stepped care provides a framework for using limited resources to greatest effect. Professional care is stepped in intensity—that is, it starts with limited professional input and systematic monitoring and is then augmented for patients who do not achieve an acceptable outcome. Initial and subsequent treatments are selected according to evidence based guidelines in light of a patient's progress.

The principle of increasing intensity of professional input for those who do not respond to initial management is familiar in primary care. However, organised stepped care requires the systematic monitoring of progress and higher levels of coordination between specialist care, care management, and primary care than generally exist. The primary care team, a specialist consultant (when needed), and a care manager (when needed) work together to provide the level of professional support needed to achieve a favourable outcome. Stepped care is individualised according to each patient's preferences and progress.

Plan for collaborative self care

1 Assessment
- Assess patient's self management beliefs, attitudes, and knowledge
- Identify personal barriers and supports
- Collaborate in setting goals
- Develop individually tailored strategies and problem solving

2 Goal setting and personal action plan
- List goals in behavioural terms
- Identify barriers to implementation
- Make plans that address barriers to progress
- Provide a follow up plan
- Share the plan with all members of the healthcare team

3 Active follow up to monitor progress and support patient

Levels of stepped care

1 Systematic routine assessment and preventive maintenance
2 Self care with low intensity support
3 Care management in primary care
4 Intensive care management with specialist advice
5 Specialist care

Assumptions of stepped care

- Different individuals require different levels of care
- The optimal level of care is determined by monitoring outcomes
- Moving from lower to higher levels of care based on patient outcomes can increases effectiveness and lower costs

Skills required by those delivering care

The team providing care must not only be familiar with a patient's condition but must also possess the psychological skills to help the patient achieve self care. They also need access to specialists in psychological and psychiatric management to provide supervision and consultation in selected cases. The necessary psychological skills include
• Anxiety management
• Recognition and treatment of depression
• Cognitive behavioural analysis
• Cognitive behavioural principles of step by step change
• Ability to monitor patient's progress.

Changes in the organisation of care

Achieving collaboration between healthcare providers and chronically ill patients requires organisational changes in six related areas.

Organisation of care—Clinical leadership should encourage efforts to improve quality, including development of incentives for improved care and reorganisation of acute care to encourage self care.

Clinical information systems—A disease (or disorder) registry should be set up that identifies the population to be served and includes information on the performance of guideline based care, including self care tasks. The registry should permit identification of patients with specific needs, reminder systems, and tailored treatment planning.

Example of changes in organisation of care for patients with diabetes

Organisation of care
- Primary care clinic initiates year long effort to reorganise diabetes care
- Team is set up and meets regularly to make changes, monitor progress, and address obstacles

Clinical information systems
- Team develops a register of all patients with diabetes in the clinic, with records of HbA_{1c} values, eye and foot examinations, and goals and key elements of patients' personal action plans

Delivery system design
- Clinic nurses assigned responsibility for diabetes case management
- Doctors agree to provide planned visits for all diabetic patients at least once a year, including preventive services (such as eye and foot examinations, ordering HbA_{1c} tests, screening for depression)
- Clinic support staff maintain the register and print out a status report before each visit

Decision support
- Team agrees on standard evidence based guidelines and adapts them to clinic and liaison with the specialist diabetic clinic
- Team agrees a standard form for planned visits

Community resources
- Nurse case managers plan training in diabetes self management. The nurses are trained to co-lead the course at regular intervals

Self care support
- Nurse case managers decide that every diabetic patient will have a personal action plan developed within a year
- Each nurse sees one patient a week until this goal is accomplished
- Nurses telephone patients who have not been seen for six months and those who need extra support to achieve their goals

Delivery system design—Practice team roles should be changed in the organisation of visits and in follow up care. Useful innovations include group visits, planned visits, and telephone delivered care.

Decision support—Evidence based practice guidelines and protocols should be made effective by integrating information and reminders into visits. There should be collaborative support from relevant medical specialties.

Community resources—Links should be established with community resources, especially for vulnerable populations such as elderly, low income, and deprived populations.

Self care support—Tailored educational resources, skills training, and psychosocial support are effective. Successful self care programmes rely on collaboration; patient centred interventions for managing illness are especially beneficial.

Is this approach feasible for the large numbers of patients seen in busy primary and secondary care settings? There is growing experience with integrating support for self care to the delivery of routine medical care. Specific techniques such as cognitive behavioural interventions and the use of nurses and other staff as care managers have been found to be both feasible and effective. However, the full implementation of this approach in primary care requires substantial organisational changes. These enable medical and other expertise to be used more effectively and efficiently. They also enable doctors to obtain greater satisfaction in being responsible for higher quality care.

Evidence that it works

Collaborative self care has been used to guide efforts to improve the quality of chronic illness care in many different healthcare settings and for many different chronic conditions including diabetes, heart failure, geriatric care, depression, and asthma. This approach gives patients the confidence and skills for self care and for getting what they need from the healthcare system (that is, becoming active, informed patients). Such effective support of patients is more likely to occur when the providers of care themselves have the information, training, resources, and time to deliver effective interventions (that is, are a well prepared, proactive practice team).

There is now considerable evidence and practical experience that supports fundamental changes in the way we organise and deliver health care to better support patients who are living with a chronic condition. Consequently, we need to include psychological and behavioural expertise as essential supplements to basic medical treatment.

Patient centred care is more than a respectful attitude or a style of clinical interviewing. It means that healthcare systems are organised to maximise the effectiveness of patients to manage their chronic illness themselves.

Psychological medicine will make its full contribution only when an awareness of the importance of psychological and behavioural factors is fully integrated into general medical care.

Work on this article was supported by grants from the Robert Wood Johnson Foundation National Program for Improving Chronic Illness Care, NIMH grants MH51338 and MH41739, and NIH grant P01 DE08773.

The picture is reproduced with permission of CC Studio/SPL.

Making evidence based care time and cost effective

Problems
- Time for patient care
- Time for assessing evidence
- Unrealistic patient expectations and demands
- Lack of patient understanding of behavioural basis of self care
- Lack of involvement of patients in clinical decisions
- Lack of professional skills
- Access to disparate community and medical services

Solutions
- Treatment protocols
- Involvement of healthcare team
- Use of self help procedures
- Formalising links with local health, social, and voluntary agencies
- Liaison with specialist medical, psychiatric, and psychological services
- Continuing professional development

Evidence based summary

- Collaborative and adaptive approaches to self care that are structured and integrated into medical services improve outcomes for many chronic diseases
- Systematic setting of therapeutic goals and monitoring of clinical treatment and outcomes are integral to this approach
- Such an approach to health care will often require changes to the way in which teams and primary and secondary care services interact

Department of Health. *The expert patient: a new approach to chronic disease management for the 21st century* (www.ohn.gov.uk/ohn/people/ep_report.pdf)

Gibson PG, Coughlan J, Wilson AJ, Abramson M, Bauman A, Hensley MJ, et al. Self-management education and regular practitioner review for adults with asthma. *Cochrane Database Syst Rev* 2000;(2): CD001117

Further reading

- Department of Health. *The expert patient: a new approach to chronic disease management for the 21st century*. (www.ohn.gov.uk/ohn/people/ep_report.pdf)
- Lorig KR, Sobel DS, Stewart AL, Brown BW Jr, Bandura A, Ritter P, et al. Evidence suggesting that a chronic disease self-management program can improve health status while reducing hospitalisation. *Med Care* 1999;37:5-14
- Von Korff M, Gruman J, Schaefer J, Curry S, Wagner EH. Collaborative management of chronic illness. *Ann Intern Med* 1997; 127:1097-102
- Wagner EH, Glasgow RE, Davis C, Bonomi AE, Provost L, McCulloch D, et al. Quality improvement in chronic illness care: a collaborative approach. *Jt Comm J Qual Improv* 2001;27:63-80
- Wolpert HA, Anderson BJ. Management of diabetes: are doctors framing the benefits from the wrong perspective? *BMJ* 2001;323: 994-6

4 Depression in medical patients

Robert Peveler, Alan Carson, Gary Rodin

Depressive illness is usually treatable. It is common and results in marked disability, diminished survival, and increased healthcare costs. As a result, it is essential that all doctors have a basic understanding of its diagnosis and management. In patients with physical illness depression may

- Be a coincidental association
- Be a complication of physical illness
- Cause or exacerbate somatic symptoms (such as fatigue, malaise, or pain).

Clinical features and classification

The term depression describes a spectrum of mood disturbance ranging from mild to severe and from transient to persistent. Depressive symptoms are continuously distributed in any population but are judged to be of clinical significance when they interfere with normal activities and persist for at least two weeks, in which case a diagnosis of a depressive illness or disorder may be made. The diagnosis depends on the presence of two cardinal symptoms of persistent and pervasive low mood and loss of interest or pleasure in usual activities.

Adjustment disorders are milder or more short lived episodes of depression and are thought to result from stressful experiences.

Major depressive disorder refers to a syndrome that requires the presence of five or more symptoms of depression in the same two week period.

Dysthymia covers persistent symptoms of depression that may not be severe enough to meet the criteria for major depression, in which depressed mood is present for two or more years. Such chronic forms of depression are associated with an increased risk of subsequent major depression, considerable social disability, and unhealthy lifestyle choices such as poor diet or cigarette smoking.

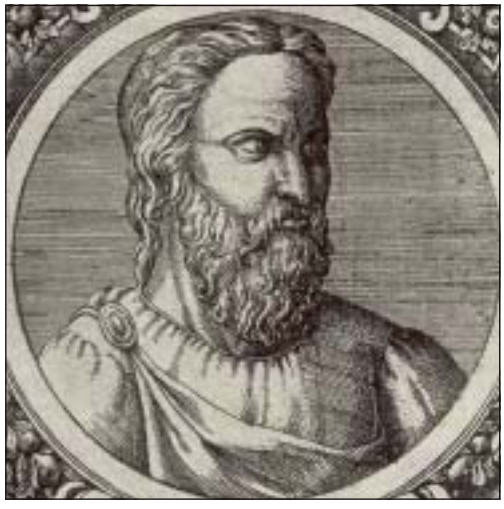
Manic depressive (bipolar) disorder relates to the occurrence of episodes of both major depression and mania.

Epidemiology

The World Health Organization estimates that depression will become the second most important cause of disability worldwide (after ischaemic heart disease) by 2020. Major depressive disorder affects 1 in 20 people during their lifetime. Both major depression and dysthymia seem to be more common in women.

Depressive illness is strongly associated with physical disease. Up to a third of physically ill patients attending hospital have depressive symptoms. Depression is even more common in patients with

- Life threatening or chronic physical illness
- Unpleasant and demanding treatment
- Low social support and other adverse social circumstances
- Personal or family history of depression or other psychological vulnerability
- Alcoholism and substance misuse
- Drug treatments that cause depression as a side effect, such as antihypertensives, corticosteroids, and chemotherapy agents.



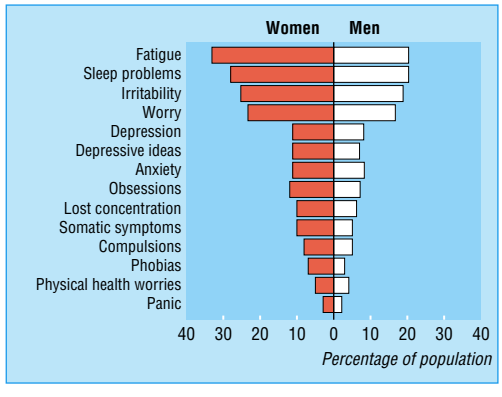
Aretaeus of Cappadocia (circa 81-138 AD) is credited with the first clinical description of depression

Criteria for major depression*

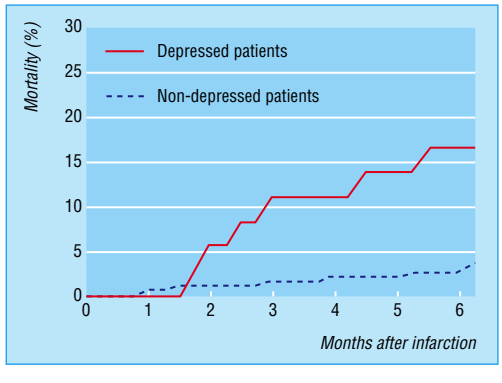
Five or more of the following symptoms during the same two week period representing a change from normal

- Depressed mood†
- Substantial weight loss or weight gain
- Insomnia or hypersomnia
- Feelings of worthlessness or inappropriate guilt
- Recurrent thoughts of death or suicide or suicide attempt
- Decreased interest or pleasure†
- Psychomotor retardation or agitation
- Fatigue or loss of energy
- Diminished ability to think or concentrate

*From *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition
†One of these symptoms must be present



"Neurotic" symptoms, including depression, are continuously distributed in the UK population

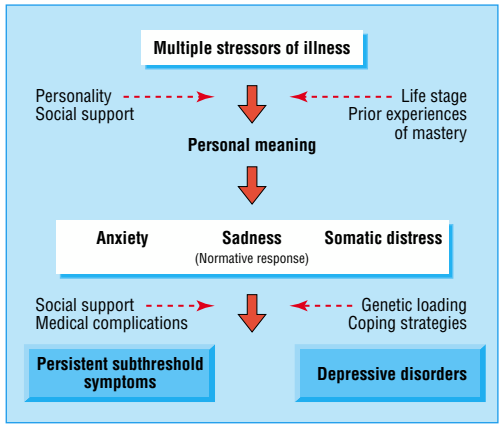


The association between depression and mortality after myocardial infarction

Risk factors

Anxiety, sadness, and somatic discomfort are part of the normal psychological response to life stress, including medical illness. Clinical depression is a final common pathway resulting from the interaction of biological, psychological, and social factors. The likelihood of this outcome depends on such factors as genetic and family predisposition, the clinical course of a concurrent medical illness, the nature of the treatment, functional disability, the effectiveness of individual coping strategies, and the availability of social and other support.

In the attempts to understand the relation between physical illness and depression there has been much debate about the direction of causality. In particular, there has been speculation that certain illnesses—such as stroke, Parkinson's disease, multiple sclerosis, and pancreatic cancer—may cause depression via direct biological mechanisms. Stroke has perhaps received the most attention, but studies have failed to convincingly show direct aetiological mechanisms.



Pathways to depression

Recognition and diagnosis

In spite of its enormous clinical and public health importance, depressive illness is often underdiagnosed and undertreated, particularly when it coexists with physical illness. This often causes great distress for patients who have mistakenly assumed that symptoms such as weakness or fatigue are due to an underlying medical condition.

All medical practitioners must be able to diagnose and manage depressive illness effectively. This depends on

- Alertness to clues in interviews
- Patients' manner
- The use of screening questions in those at risk—in particular, two questions about low mood and lack of pleasure in life can detect up to 95% of patients with major depression.

Self report screening instruments, such as the Beck depression inventory (BDI) and the hospital anxiety and depression scale (HADS) cannot replace systematic clinical assessment, but they are useful in drawing attention to depression and other emotional disturbances in clinical settings where mood is not routinely assessed. Doctors must be aware that persistent low mood and lack of interest and pleasure in life cannot be accounted for by severe physical illness alone. The usual response to illness and treatment is impressive resilience.

If there is doubt about the diagnosis, a doctor may resort to an empirical trial of treatment to establish whether there is benefit. The wider availability of safer drugs and psychological treatments makes this option more attractive than in the past.

Reasons why depression is missed

- Difficulty distinguishing psychological symptoms of depression, such as sadness and loss of interest, from a "realistic" response to stressful physical illness
- Confusion over whether physical symptoms of depression are due to an underlying medical condition
- Negative attitudes to diagnosis of depression
- Unsuitability of clinical setting for discussion of personal and emotional matters
- Patients' unwillingness to report symptoms of depression

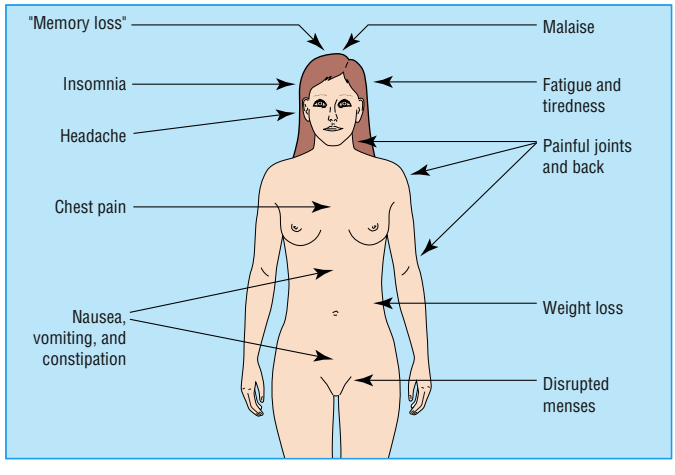
Screening questions for depression

- How have you been feeling recently?
- Have you been low in spirits?
- Have you been able to enjoy the things you usually enjoy?
- Have you had your usual level of energy, or have you been feeling tired?
- How has your sleep been?
- Have you been able to concentrate on newspaper articles or your favourite television or radio programmes?

Management

The main aims of treatment are to improve mood and quality of life, reduce the risk of medical complications, improve compliance with and outcome of physical treatment, and facilitate the "appropriate" use of healthcare resources. The development of a treatment plan depends on systematic assessment that should, whenever possible, not only involve the patients but also their partners or other key family members.

Milder or briefer adjustment disorders can be managed by primary care staff without recourse to specialist referral. Education, advice, and reassurance are of value. It is important that primary care staff are familiar with the properties and use of the commoner antidepressant drugs, and the value of brief psychological treatments such as cognitive behaviour therapy, interpersonal therapy, and problem solving.



Physical symptoms that may be due to depression

Patients with more enduring or severe symptoms will usually require specific forms of treatment, usually drug treatment. Staff should also be able to assess suicidal thinking and risk. For patients with suicidal ideation or those whose depression has not responded to initial management, specialist referral is the next step in management.

Drug treatment

Antidepressants have been shown to be effective in treating major depressive disorder irrespective of whether the mood disturbance is "understandable." There have been far fewer trials of antidepressants in patients who are also physically unwell, but the available evidence is in keeping with the treatment of depression generally.

One of the commonest questions is which antidepressant should be used. For non-specialists, the range of available drugs, and the claims made about them can be bewildering. There are four main classes of antidepressant

- Tricyclics
- Selective serotonin reuptake inhibitors
- Monoamine oxidase inhibitors
- Others (noradrenaline reuptake inhibitors).

Choice of agent

Data from the Cochrane Collaboration and other systematic reviews show that the differences in overall tolerability between different preparations is minimal. In general, patients are slightly less likely to drop out of trials because of unacceptable side effects when taking a selective serotonin reuptake inhibitor but are slightly less likely to drop out because of treatment inefficacy when taking a tricyclic. Rather than continuously experimenting with a range of different drugs, clinicians should stick to prescribing one drug from each class in order to become familiar with their dosing regimens, actions, interactions, and side effects. Clinicians should also be aware that in certain situations one class of drug may be more advisable than others.

Adequacy of treatment

The debate about different preparations has obscured a potentially more important issue—that of drug dose and compliance. Most prescriptions for antidepressants are for inadequate doses and for inadequate time periods. This problem is compounded by only a minority of patients complying with the prescribed treatment. A recent household survey by the Royal College of Psychiatrists showed that many people believed that antidepressants were addictive and could permanently damage the brain.

Explanation

To treat patients successfully with antidepressants, doctors must be able to show their patient that they have understood the patient's problems, considered the issues, and are advising the best available treatment (see previous chapters). Before starting treatment, patients should be given an explanation of side effects and be reassured that side effects tend to be worse during the first two weeks of treatment and then diminish. They need to be warned that they are unlikely to feel benefits from treatment in the first four weeks. They should be given follow up appointments during this period in order to encourage compliance.

Duration of treatment

After initial treatment has led to remission of symptoms, subsequent treatment can be divided into two phases. Firstly, four to six months of continuous treatment at full dose are

Clinical assessment of suicidal intent

Low level risk

Clinical picture	*Action*
• Suicidal ideation but no suicide attempts • Supportive environment • Physically healthy • No history of psychiatric illness	Consider referral to mental health professional for routine appointment (not always necessary)

Moderate level risk

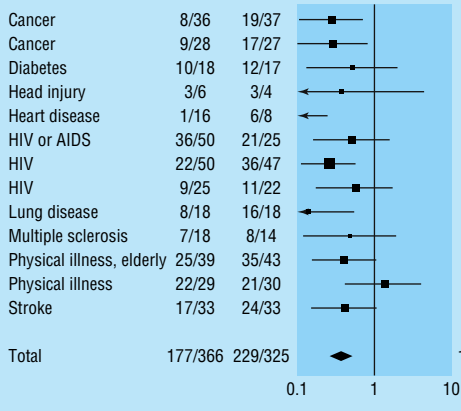
Clinical picture	*Action*
• Low lethality suicide attempt (patient's perception of lethality) • Frequent thoughts of suicide • Previous suicide attempts • Persistent depressive symptoms • Serious medical illness • Inadequate social support • History of psychiatric illness	Refer to mental health professional, to be seen as soon as possible

High level risk

Clinical picture	*Action*
• Definite plan for suicide (When? Where? How?) • Major depressive disorder, severe • High lethality suicide attempt or multiple attempts • Advanced medical disease • Social isolation • History of psychiatric illness	Refer to mental health professional for immediate assessment

Comparison: Antidepressants *v* Placebo
Outcome: Lack of improvement at end of study

Condition	No of patients: Drug	No of patients: Control	Odds ratio (95% CI)	Weighting (%)	Odds ratio (95% CI)
Cancer	8/36	19/37		11.2	0.29 (0.11 to 0.75)
Cancer	9/28	17/27		9.0	0.30 (0.10 to 0.85)
Diabetes	10/18	12/17		5.4	0.54 (0.14 to 2.07)
Head injury	3/6	3/4		1.7	0.39 (0.03 to 4.54)
Heart disease	1/16	6/8		3.0	0.04 (0.01 to 0.26)
HIV or AIDS	36/50	21/25		8.0	0.52 (0.17 to 1.60)
HIV	22/50	36/47		15.2	0.26 (0.12 to 0.59)
HIV	9/25	11/22		7.6	0.57 (0.18 to 1.80)
Lung disease	8/18	16/18		5.3	0.14 (0.04 to 0.56)
Multiple sclerosis	7/18	8/14		5.2	0.49 (0.12 to 1.95)
Physical illness, elderly	25/39	35/43		10.5	0.42 (0.16 to 1.11)
Physical illness	22/29	21/30		7.7	1.34 (0.43 to 4.18)
Stroke	17/33	24/33		10.2	0.41 (0.15 to 1.10)
Total	177/366	229/325		100.0	0.37 (0.27 to 0.51)



Meta-analysis of randomised controlled trials of drug treatment of depression in the physically ill

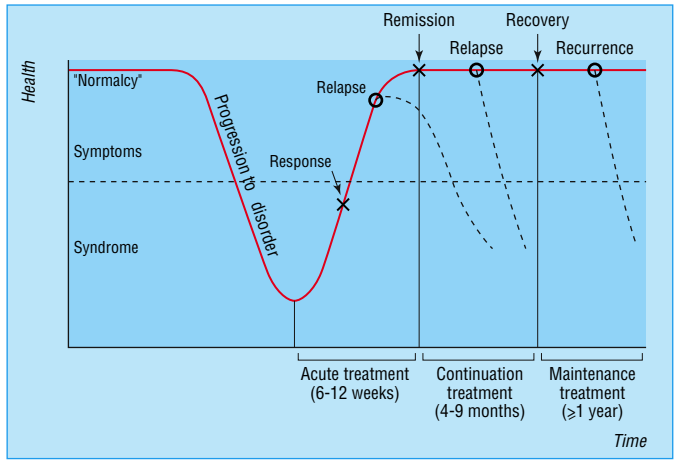


Chart of clinical course indicating remission, recovery, and relapse

necessary to consolidate remission and prevent early relapse. Secondly, consideration must be given to preventive maintenance treatment, to reduce the risks of recurrence of depression. This is usually indicated if the patient has had two or more episodes of depression within the past five years. Psychological treatment may also help to prevent recurrence and can be used in combination with drug treatment.

Psychological treatment

Psychological treatment can range from discussion and simple problem solving to more specialised cognitive or dynamic behavioural psychotherapies. In many cases, brief treatment by non-specialists in primary and secondary care can be effective. Such interventions may include education and reassurance about the common reactions to the threats and losses associated with illness and empathic listening to patients' views, uncertainties, and beliefs about the illness. Education and advice about the medical condition and associated depression may prevent needless worry, reduce feelings of helplessness, and diminish irrational fears. Therapeutic approaches that support or promote active coping strategies are an important aspect of treatment in physically ill patients.

Cognitive behavioural principles may be used by non-specialists to correct distorted thinking and to encourage behaviours that contribute to patients' sense of mastery and wellbeing. Training in briefer forms of treatment using cognitive behavioural principles for primary care staff may be a worthwhile investment.

Cognitive behaviour therapy, interpersonal therapy, and problem solving have all been shown to be effective for treating depression, although there has been only limited evaluation of their effectiveness in physically ill populations. Although time consuming by comparison with drug treatment, psychological treatment may reduce relapse rates and may be cost effective in the long run. Some patients may require preliminary treatment with drugs to enable them to make best use of psychological treatment.

Service organisation

Depression is so common in physically ill patients that it is not feasible for all cases to be managed by mental health specialists. There are advantages to collaborative management with primary care staff working closely with mental health specialists. Community based mental health services may be less accessible to general hospitals and often lack specialist knowledge about assessment and treatment when an important physical illness is also present. Liaison psychiatry services are often well placed to provide support, training, and psychiatric expertise to general hospital patients in a timely fashion.

Problem solving in psychological treatment

- Define and list the problems
- Choose a problem for action
- List alternative courses of action
- Evaluate courses of action and choose the best
- Try the action
- Evaluate the results
- Repeat until major problems have been solved

Evidence based summary

- Depressive illness is an important cause of morbidity and disability in physically ill patients
- All patients with depression should be examined for suicidal ideation
- Depression is treatable in physically ill patients

Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. The functioning and well-being of depressed patients. Results from the medical outcomes study. *JAMA* 1989;262:914-9

Carson AJ, Best S, Warlow C, Sharpe M. How common is suicidal ideation among neurology outpatients? *BMJ* 2000;320:1311-2

Gill D, Hatcher S. Antidepressants for depression in medical illness. *Cochrane Database Syst Rev* 2000;(4):CD001312

Further reading

- Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al. Lifetime and 12-month prevalence of DSM-IIIR psychiatric disorders in the United States: results of the national comorbidity survey. *Arch Gen Psychiatr* 1994;51,8-19
- Rodin G, Craven J, Littlefield C. *Depression in the medically ill: an integrated approach.* New York NY: Brunner/Mazel, 1991
- Royal College of Physicians, Royal College of Psychiatrists. *The psychological care of medical patients: recognition of need and service provision.* London: RCP, RCPsych, 1995

The diagram of the distribution of neurotic symptoms in the UK population is adapted from Jenkins et al *Psychol Med* 1997;27:765-74. The graph of association between depression and mortality after myocardial infarction. is adapted from Frasure Smith et al *JAMA* 1993;270: 1819-25. The diagram showing pathways to depression is adapted from Rodin et al *Depression in the medically ill* 1991. The meta-analysis of trials comparing antidepressants is adapted from Gill and Hatcher *Cochrane Database Syst Rev* 2000;(4):CD001312.

5 Anxiety in medical patients

Allan House, Dan Stark

Doctors often consider anxiety to be a normal response to physical illness. Yet, anxiety afflicts only a minority of patients and tends not to be prolonged. Any severe or persistent anxious response to physical illness merits further assessment.

What is anxiety?

Anxiety is a universal and generally adaptive response to a threat, but in certain circumstances it can become maladaptive. Characteristics that distinguish abnormal from adaptive anxiety include

- Anxiety out of proportion to the level of threat
- Persistence or deterioration without intervention (>3 weeks)
- Symptoms that are unacceptable regardless of the level of threat, including
  - Recurrent panic attacks
  - Severe physical symptoms
  - Abnormal beliefs such as thoughts of sudden death
- Disruption of usual or desirable functioning.

One way to judge whether anxiety is abnormal is to assess whether it is having a negative effect on the patient's functioning.

Abnormal anxiety can present with various typical symptoms and signs, which include

- Autonomic overactivity
- Behaviours such as restlessness and reassurance seeking
- Changes in thinking, including intrusive catastrophic thoughts, worry, and poor concentration
- Physical symptoms such as muscle tension or fatigue.

Classification of abnormal anxiety

Abnormal anxiety can be classified according to its clinical features. In standardised diagnostic systems there are four main patterns of abnormal anxiety.

Anxious adjustment disorder—Anxiety is closely linked in time to the onset of a stressor.

Generalised anxiety disorder—Anxiety is more pervasive and persistent, occurring in many different settings.

Panic disorder—Anxiety comes in waves or attacks and is often associated with panicky thoughts (catastrophic thoughts) of impending disaster and can lead to repeated emergency medical presentations.

Phobic anxiety—Anxiety is provoked by exposure to a specific feared object or situation. Medically related phobic stimuli include blood, hospitals, needles, doctors and (especially) dentists, and painful or unpleasant procedures.

Additionally, anxiety often presents in association with depression. Mixed anxiety and depressive disorders are much more common than anxiety disorders alone. Treatment for the depression may resolve the anxiety. Anxiety can also be the presenting feature of other psychiatric illnesses common in physically ill people, such as delirium or drug and alcohol misuse.

William Cullen (1710-90) coined the term neurosis (though the term as he used it bears little resemblance to modern concepts of anxiety disorders)

Somatic and psychological symptoms of anxiety disorders

In all anxiety disorders

- Palpitations, pounding heart, accelerated heart rate
- Trembling or shaking
- Difficulty in breathing
- Chest pain or discomfort
- Feeling dizzy, unsteady, faint, light headed
- Fear of losing control, going crazy, passing out
- Sweating
- Dry mouth
- Feeling of choking
- Nausea or abdominal discomfort
- Feeling that objects are unreal or that self is distant
- Fear of dying
- Numbness or tingling sensations
- Hot flushes or cold chills

In more severe or generalised anxiety disorders

- Muscle tension or aches and pains
- Feeling keyed up, on edge, or mentally tense
- Exaggerated response to minor surprises or being startled
- Persistent irritability
- Restlessness, inability to relax
- Sensation of difficulty swallowing, lump in the throat
- Difficulty concentrating or "mind going blank" from anxiety or worry
- Difficulty in getting to sleep because of worry

Distinguishing features of anxiety disorders

Anxious adjustment disorder

Prevalence in general population—Not known

Cardinal features

- Onset of symptoms within 1 month of an identifiable stressor
- No specific situation or response

Generalised anxiety disorder

Prevalence in general population—31 cases/1000 adults

Cardinal features

- Period of 6 months with prominent tension, worry, and feelings of apprehension about everyday problems
- Present in most situations and no specific response

Panic disorder

Prevalence in general population—8 cases/1000 adults

Cardinal features

- Discrete episode of intense fear or discomfort with crescendo pattern; starts abruptly and reaches a maximum in a few minutes
- Occurs in many situations, with a hurried exit the typical response

Phobia

Prevalence in general population—11 cases/1000 adults

Cardinal features

- No specific symptom pattern
- Occurs in specific situations, with an avoidance response

Detecting anxiety and panic

Who is at risk?—Certain groups are more vulnerable to anxiety disorders: younger people, women, those with social problems, and those with previous psychiatric problems. However, such associations are less consistent in the setting of life threatening illness, perhaps because susceptibility to anxiety becomes less important as the stressor becomes more severe. Pathological anxiety is commoner among patients with a chronic medical condition than in those without.

Excluding physical causes—There are many presentations with physical complaints whose aetiology may be due to anxiety. Equally, several physical illnesses can cause anxiety or similar symptoms. When such disorders cannot be reliably distinguished from anxiety by clinical examination they need to be excluded through appropriate investigation. A firm diagnosis of anxiety should therefore be made only when a positive diagnosis can be supported by the presence of a typical syndrome and after appropriate investigation.

Use of screening questionnaires—Screening, with self completed questionnaires, has been widely used to improve detection of psychiatric morbidity, including anxiety. Such questionnaires are acceptable to patients and can be amenable to computerised automation in the clinic. The hospital anxiety and depression scale, the general health questionnaire, and many quality of life instruments include anxiety items. No one questionnaire has been consistently shown to be preferable to another.

Iatrogenic anxiety—Anxiety symptoms can be caused by poor communication (see chapters 1 and 2) and by prescribed drugs. Well known causes include corticosteroids, β adrenoceptor agonists, and metoclopramide, but doctors should remember that many less commonly used drugs can cause psychiatric syndromes.

Treatment of anxiety and panic

General management

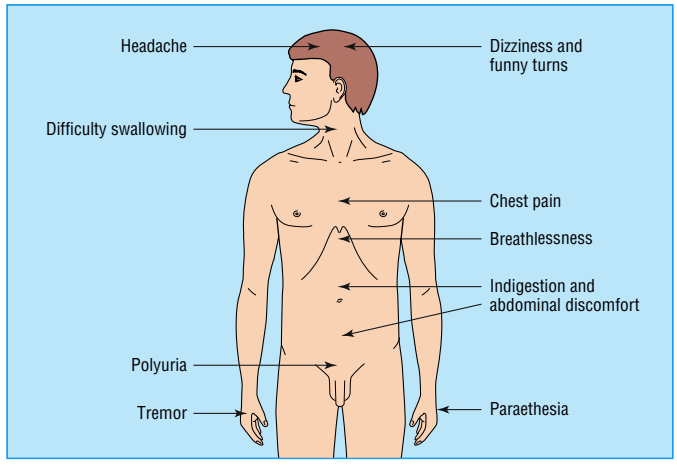
Treating anxiety is part of the management of most medical conditions. It can lead to direct improvement of symptoms or improve patient compliance. It is important to intervene if a positive diagnosis of anxiety is made. Without treatment, anxiety is associated with increased disability, increased use of health service resources, and impaired quality of life.

Involving a mental health professional is not always possible for anxious patients, particularly those in general hospital settings. The range of available services is often limited, and not all patients are prepared to accept referral. Since many patients have to be managed without recourse to psychiatric services, treating anxiety should be considered a core skill for all doctors.

Giving information is often the first step in helping anxious patients, so much so that it has been said that knowledge is reassurance. While information must be tailored to the wishes of the individual, many patients want more information than they are given. Such a simple step as showing people where they are to be cared for can reduce anxiety.

Effective communication is central to information giving, with evidence that anxiety is associated with poor communication. Training doctors to use open questions, discuss psychological issues, and summarise—and to avoid reassurance, "advice mode," and leading questions—has been shown to lead to greater disclosure and enduring change in patients with psychological problems.

Reassurance is one of the most widely practised clinical skills. Doctors often need to tell patients that their symptoms are not due to occult disease. Simple reassurance, however, may be



Common physical problems that may be caused by anxiety

Medical conditions mimicking or directly resulting in anxiety

- Poor pain control—Such as ischaemic heart disease, malignant infiltration
- Hypoxia—May be episodic in both asthma and pulmonary embolus
- Hypocapnia—May be due to occult bronchial hyperreactivity
- Central nervous system disorders (structural or epileptic)
- Anaemia
- Hypoglycaemia
- Hyperkalaemia
- Alcohol or drug withdrawal
- Vertigo
- Thyrotoxicosis
- Hypercapnia
- Hyponatraemia

Self reported questionnaires used to assess anxiety

Hospital anxiety and depression scale

Advantages
- Excludes somatic symptoms of disease
- Brevity (14 items in all, 7 concerning anxiety)
- Widespread use in cancer and other physical illnesses
- More effective than many other instruments
- Used as a screen and a measure of progress

Disadvantages
- Recent concern that, used alone, it is poor at detecting depression

State-trait anxiety inventory

Advantages
- Specific to anxiety
- Used as a screen and a measure of progress

Disadvantages
- Used alone does not detect depression
- Longer (20-40 items) than many other self reported questionnaires

General health questionnaire

Advantages
- Brevity (12 or 28 items)
- Excludes somatic symptoms of disease
- Used as a screen and a measure of progress

Disadvantages
- May not be accurate in detecting chronic problems

Common drug causes of anxiety

- Anticonvulsants—Carbamazepine, ethosuximide
- Antimicrobials—Cephalosporins, ofloxacin, aciclovir, isoniazid
- Bronchodilators—Theophyllines, β_2 agonists
- Digitalis—At toxic levels
- Oestrogen
- Insulin—When hypoglycaemic
- Non-steroidal anti-inflammatory drugs—Indomethacin
- Antidepressants—Specific serotonin reuptake inhibitors
- Antihistamines
- Calcium channel blockers—Felodipine
- Dopamine
- Inotropes—Adrenaline, noradrenaline
- Levodopa
- Corticosteroids
- Thyroxine

Many drugs can cause palpitation or tremor, but these should be easily distinguished from anxiety by clinical examination

ineffective for anxious patients; their anxiety may be reduced initially by the consultation, but it rapidly returns. Several theoretical models of this problem have been suggested, based on the patterns of thinking ("cognitions") of people who are difficult to reassure.

Preparation for unpleasant procedures can remove the additional burden of facing the unknown. It may also allow planning of short term tactics for dealing with anxiety provoking circumstances. Anxious patients are highly vigilant and overaware of threatening stimuli. They often use "quick fix" techniques based on avoidance of threat to reduce anxiety; such strategies are generally maladaptive and result in increasing disability. In some medical situations, however, such avoidance may not be a bad thing if the threat is temporary. A similar effect is seen with use of benzodiazepine to provide temporary relief from anxiety symptoms that will not recur because the stressor is not persistent.

Behavioural treatments are among the most effective treatments for anxiety disorders. Many patients restrict their activities in response to anxiety, which often has the effect of increasing both the level of anxiety and the degree of disability in the longer term. The principle of treatment is that controlled exposure to the anxiety producing stimulus will eventually lead to diminution in symptoms. Although specific behavioural treatments will normally be conducted by specialists, other clinicians should be aware of the basic principles. It is important to encourage and help patients to maintain their normal activities as much as possible, even if this causes temporary increases in anxiety.

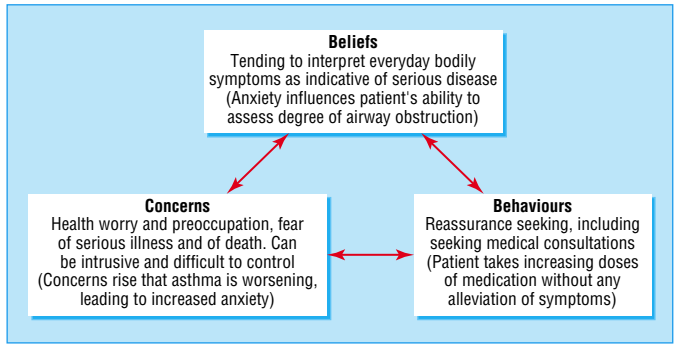
Drug treatments—Several drugs can be used to treat anxiety, each with its own advantages and disadvantages. Long term benzodiazepine dependence and misuse are considered by many to be a problem in medical practice. Although the evidence for this is conflicting, the use of benzodiazepines may be reserved for the short term treatment of anxiety and for emergencies.

Drug withdrawal—Dependence on other substances, particularly analgesics and alcohol, occurs fairly frequently in the context of anxiety. This often results from self medication for anxiety. In this situation withdrawal from the existing "treatment" will be an important part of the anxiety management programme.

Role of specialist psychological treatment

Clinical studies indicate that psychological interventions for anxiety can be effective both in general psychiatric settings and for physically ill patients. The most popular, and those with the best evidence to support them, are based on the principles of behaviour, cognitive behaviour, or interpersonal therapy.

In behaviour and cognitive behaviour therapies the main aim is to help patients identify and challenge unhelpful ways of thinking about and coping with physical symptoms and their meaning, about themselves, and about how they should live their lives. In interpersonal therapies the main focus is on relationships with family members and friends—how such relationships are affected by illness and how they influence patients' current emotional state. Patients need to know that such therapies may be both brief and practical. Fewer than six sessions may be enough, concentrating on symptoms or the immediate problems associated with them and learning new ways of dealing with problems. In only a minority of cases is more extended therapy needed, usually when anxiety is longer standing and only partially due to associated physical disease.



Characteristic features of health anxiety (using the example of asthma)

Drug interventions in anxious medical patients

β Blockers
- Benefits unproved in randomised controlled trials
- Help to control palpitation and tremor, but not anxiety itself
- Often used for performance anxiety, such as in interviews or examinations

Tricyclic antidepressants (such as imipramine)
- Likely to be beneficial (number needed to treat = 3)
- Anxiolytic effect is slow in onset (weeks)
- Not dependency inducing
- Useful in panic disorder or in anxiety with depression
- Anticholinergic effects can be ameliorated by a low starting dose

Selective serotonin reuptake inhibitors (such as sertraline)
- Benefit unproved but suggested
- Less anticholinergic effects than tricyclic antidepressants
- Start at low dose in anxious patients

Short acting benzodiazepines (such as alprazolam)
- Effectiveness and relative lack of toxicity well established
- All benzodiazepines can induce dependency
- Rapid onset of effect, but problems may recur on withdrawal
- Less likely to accumulate in liver

Antipsychotics (such as haloperidol)
- Benefits unproved in randomised controlled trials
- Useful adjunct to benzodiazepines
- Less respiratory depression that benzodiazepines
- Not dependency inducing
- Risk of acute dystonia, akathisia, and parkinsonism
- Avoid long term use because of risk of tardive dyskinesia

Buspirone
- Limited evidence for effectiveness from randomised controlled trials, few clinicians are convinced
- Causes some nausea and dizziness

Further reading

- Noyes R, Hoehn-Saric R. Anxiety in the medically ill: disorders due to medical conditions and substances. In: Noyes R, Hoehn-Saric R, eds. *The anxiety disorders*. Cambridge: Cambridge University Press, 1998:285-334
- Colon EA, Popkin MK. Anxiety and panic. In: Rundell JR, Wise MG, eds. *Textbook of consultation-liaison psychiatry*. Washington DC: American Psychiatric Press, 1996:403-25
- Westra HA, Stewart SH. Cognitive behavioural therapy and pharmacotherapy: complementary or contradictory approaches to the treatment of anxiety? *Clin Psychol Rev* 1998;18:307-40
- Cochrane Depression Anxiety and Neurosis Group. See list of reviews at www.update-software.com/abstracts/g240index.htm

6 Functional somatic symptoms and syndromes

Richard Mayou, Andrew Farmer

Concern about symptoms is a major reason for patients to seek medical help. Many of the somatic symptoms that they present with—such as pain, weakness, and fatigue—remain unexplained by identifiable disease even after extensive medical assessment. Several general terms have been used to describe this problem—somatisation, somatoform, abnormal illness behaviour, medically unexplained symptoms, and functional symptoms. We will use the term functional symptoms, which does not assume psychogenesis but only a disturbance in bodily functioning.

Some common functional symptoms and syndromes

- Muscle and joint pain (fibromyalgia)
- Low back pain
- Tension headache
- Atypical facial pain
- Chronic fatigue
- Non-cardiac chest pain
- Palpitation
- Non-ulcer dyspepsia
- Irritable bowel
- Dizziness
- Insomnia

Classification of functional syndromes

Most functional symptoms are transient, but a sizeable minority become persistent. Persistent symptoms are often multiple and disabling and may be described as functional syndromes. Although different medical and psychiatric classifications of functional syndromes exist, these are simply alternative ways of describing the same conditions.

Medical syndromes (such as fibromyalgia and chronic fatigue, chronic pain, and irritable bowel syndromes) highlight patterns of somatic symptoms, often in relation to particular bodily systems. Although they are useful in everyday medical practice, recent studies show there is substantial overlap between them.

Psychiatric syndromes (such as anxiety, depression, and somatoform disorders) highlight psychological processes and the number of somatic symptoms irrespective of the bodily system to which they refer. Depression and anxiety often present with somatic symptoms that may resolve with effective treatment of these disorders. In other cases the appropriate psychiatric diagnostic category is a somatoform disorder.

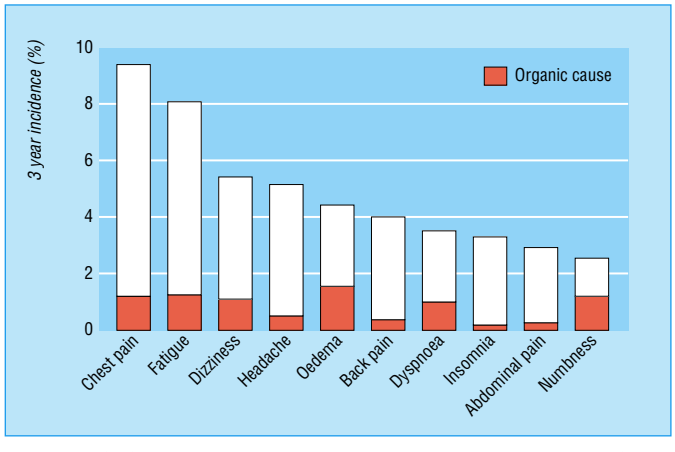
The existence of parallel classificatory systems is confusing. Both have merits, and both are imperfect. For many functional symptoms, a simple description of the symptom qualified with the descriptors single or multiple and acute or chronic may suffice. When diagnosis of a functional syndrome seems appropriate a combination of medical and psychiatric descriptors conveys the most information, such as "irritable bowel syndrome with anxiety disorder."

A major obstacle to effective management is patients feeling disbelieved by their doctor. Patients who present with symptoms that are not associated with disease may be thought by some to be "putting it on." The deliberate manufacture of symptoms or signs, however, is probably rare in ordinary practice.

René Descartes, who formulated the philosophical principle of separation of brain and mind. This has led to continuing dualism—separation of body and mind—in Western medicine and difficulty in accepting the interaction of physical and psychological factors in aetiology

Epidemiology

Community based studies report annual prevalences of 6-36% for individual troublesome symptoms. In primary care only a small proportion of patients presenting with such symptoms ever receive a specific disease diagnosis. The World Health Organization found functional symptoms to be common and disabling in primary care patients in all countries and cultures studied. Up to half of these patients remain disabled by their symptoms a year after presentation, the outcome being worse for those referred to secondary and tertiary care. The clinical and public health importance of functional symptoms has been greatly underestimated.



Three year incidence of 10 common presenting symptoms and proportion of symptoms with a suspected organic cause in US primary care

Causal factors

The cause of functional symptoms and syndromes is not fully understood, and it is therefore best to remain neutral regarding aetiological theories. In practice, functional symptoms are often attributed to single cause, which may be pathological (such as "a virus") or psychological (such as "stress"). This simplistic and dualistic approach is unhelpful both in explaining the cause to a patient and in planning treatment. The available evidence suggests that biological, psychological, interpersonal, and healthcare factors are all potentially important.

The dualistic, single factor view has tended to emphasise psychological over biological factors, as exemplified by the commonly used term "somatisation." However, recent evidence suggests that biological factors (especially reversible functional disturbance of the nervous system) are relevant to many functional syndromes, as they are to depression and anxiety disorders. A pragmatic doctor therefore asks not whether symptoms are "physical" or "mental" but whether they are fixed or are reversible by appropriate intervention.

The role of interpersonal factors in general, and of doctors and the health system in particular, in exacerbating functional symptoms has received less attention than it deserves. Raising fears of disease, performing unnecessary investigations and treatments, and encouraging disability are probably common adverse effects of medical consultations. However, denying the reality of patients' symptoms may damage the doctor-patient relationship and drive patients from evidence based care into the arms of the unhelpful, unscientific, and unscrupulous.

Aetiological factors can also be usefully divided into the stage of illness at which they have their effect. That is, they may be predisposing, precipitating, or perpetuating. Predisposing and precipitating factors are useful in producing a fuller understanding of why a patient has the symptom, while perpetuating factors are the most important for treatment.

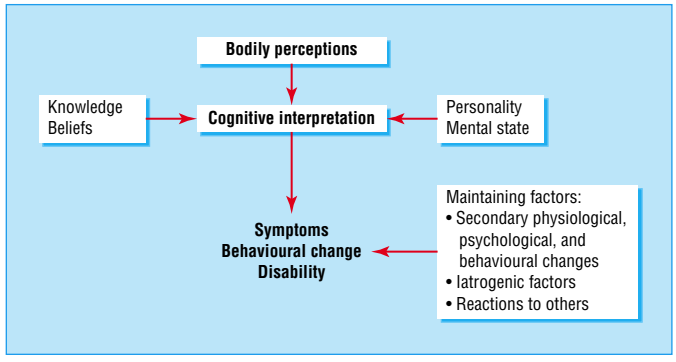
Precipitating factors—Symptoms may arise from an increased awareness of physiological changes associated with stress, depression, anxiety and sometimes disease and injury. They become important to the patients when they are severe and when they are associated with fears of, or belief in, disease.

Predisposing factors increase the chance that such symptoms will become important. Some people are probably biologically and psychologically predisposed to develop symptoms. Fear of disease may result from previous experience—for example, a middle aged man with a family history of heart disease is likely to become concerned about chest pain.

Perpetuating factors are those that make it more likely that symptoms and associated disability persists. Patients' understandable attempts to alleviate their symptoms may paradoxically exacerbate them. For example, excessive rest to reduce pain or fatigue may contribute to disability in the longer term. Doctors may also contribute to this by failing to address patients' concern or unwittingly increasing fear of disease (such as by excessive investigation). The provision of disability benefits can also be a financial disincentive for some patients to return to jobs they dislike, and the process of litigation may maintain a focus on disability rather than recovery.

Detection and diagnosis

Almost any symptom can occur in the absence of disease, but some, such as fatigue and subjective bloating, are more likely to be functional than others. Surprisingly, the more somatic symptoms a person has, the less likely it is that these symptoms reflect the presence of disease and the more likely there is associated depression and anxiety.



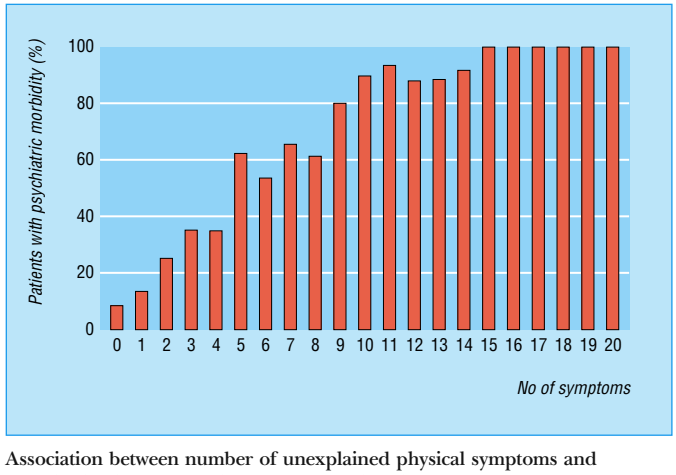
Interactive aetiology of "unexplained symptoms"

Iatrogenic factors in development of medically unexplained symptoms

- Appearance of uncertainty and inability to provide an explanation
- Expressed concern about disease explanations
- Failure to convince patient that the complaint is accepted as genuine
- Reassurance without a positive explanation being given
- Ambiguous and contradictory advice
- Excessive investigation and treatment

Individualised aetiological formulation for patient with chronic pain

Causes	Predisposing factors	Precipitating factors	Perpetuating factors
Biological	Genetic?	Injury at work	Effect of immobility Physiological mechanisms
Psychological	Lack of care as child	Trauma	Fear of worsening pain Avoid activity
Interpersonal	Family history of illness Dissatisfaction with work	Response of employer	Oversolicitous care Litigation process
Medical system	—	Misleading explanation of pain	Focus solely on somatic problems



Association between number of unexplained physical symptoms and psychiatric disorder (anxiety and depression) in an international study of primary care attenders

Patients with functional symptoms can be detected by maintaining an awareness of the problem when seeing new patients and by the use of somatic symptom questionnaires (large numbers of symptoms are more likely to be functional).

Management

Although it is essential to consider disease as the cause of the patient's symptoms an approach exclusively devoted to this can lead to difficulties if none is found. Making explicit from the start the possibility that the symptoms may turn out to be functional keeps the option of a wider discussion open. Even if more specialist treatment is needed, then the problem has, from the outset, been framed in a way that enables psychological treatment to be presented as part of continuing medical care rather than as an unacceptable and dismissive alternative. In this way it is possible to avoid an anxious disabled patient being treated by a bewildered frustrated doctor.

Investigation

An appropriate physical examination and necessary medically indicated investigation are clearly essential. Thereafter, before further investigation is done, the potential adverse psychological effect on the patient should be balanced against the likelihood and value of new information that may be obtained.

Reassurance and explanation

Most patients are reassured by being told that the symptoms they have are common and rarely associated with disease and that their doctor is familiar with them. This is especially so if accompanied by the promise of further review should the symptoms persist.

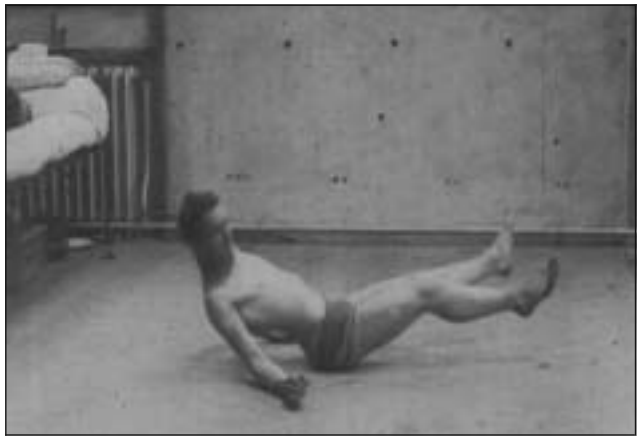
Reassurance needs to be used carefully, however. It is essential to elicit patients' specific concerns about their symptoms and to target reassurance appropriately. The simple repetition of bland reassurance that fails to address patients' fears is ineffective. If patients have severe anxiety about disease (hypochondriasis) repeated reassurance is not only ineffective but may even perpetuate the problem.

A positive explanation for symptoms is usually more helpful that a simple statement that there is no disease. Most patients will accept explanations that include psychological and social factors as well as physiological ones as long as the reality of symptom is accepted. The explanation can usefully show the link between these factors—for example, how anxiety can lead to physiological changes in the autonomic nervous system that cause somatic symptoms, which, if regarded as further evidence of disease, lead to more anxiety.

Further non-specialist treatment

A minority of patients need more than simple reassurance and explanation. Treatment should address patients' illness fears and beliefs, reduce anxiety and depression, and encourage a gradual return to normal activities.

There is good evidence that antidepressants often help, even when there are no clear symptoms of depression. Practical advice is needed, especially on coping effectively with symptoms and gradually returning to normal activity and work. Other useful interventions include help in dealing with major personal, family, or social difficulties and involving a close relative in management. Other members of the primary care or hospital team may be able to offer help with treatment, follow up, and practical help.



Functional somatic symptoms were common after combat in the first world war, such as this soldier's "hysterical pseudohypertrophic muscular spasms." The course and outcome of such symptoms can now be seen to have been substantially determined by varied medical and military approaches to prevention and treatment

Principles of assessment

- Identify patients' concerns and beliefs
- Review history of functional symptoms
- Explicitly consider both disease and functional diagnoses
- Appropriate medical assessment with explanation of findings
- Ask questions about patients' reaction to and coping with symptoms
- Use screening questions for psychiatric and social problems
- Consider interviewing relatives

Principles of treatment

- Explain that the symptoms are real and familiar to doctor
- Provide a positive explanation, including how behavioural, psychological, and emotional factors may exacerbate physiologically based somatic symptoms
- Offer opportunity for discussion of patient's and family's worries
- Give practical advice on coping with symptoms and encourage return to normal activity and work
- Identify and treat depression and anxiety disorders
- Discuss and agree a treatment plan
- Follow up and review

Non-specialist specific treatments

- Provide information and advice
- Agree a simple behavioural plan with patient and family
- Give advice about anxiety management
- Encourage use of diaries
- Advise about graded increase in activities
- Prescribe antidepressant drug
- Explain use of appropriate self help programmes

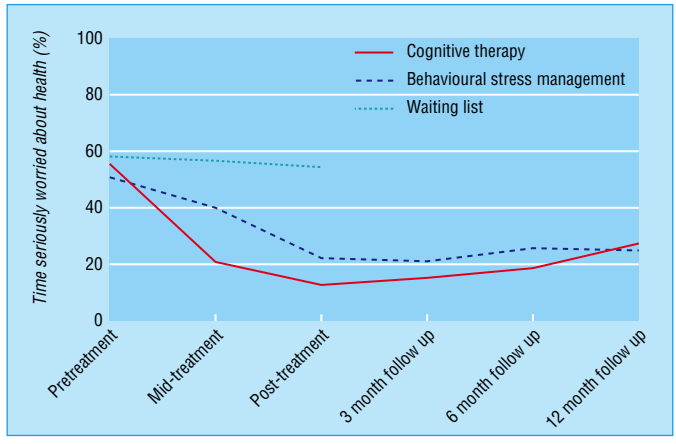
Specialist treatments

- Full and comprehensive assessment and explanation based on specialist assessment
- Cognitive behaviour therapy
- Supervised programme of graded increase in activity
- Antidepressants when these were previously not accepted or ineffective
- Illness specific interventions (such as rehabilitation programme for chronic pain)

Referral for specialist treatment

There is always a temptation to refer difficult patients to another doctor. However, this can result in greater long term difficulties if not carefully planned. When there is a good reason for further medical or psychiatric referral, then a clear explanation to the patient of the reason and an appropriately worded referral letter are essential.

Psychiatric treatments that may be required include more complex antidepressant drug regimens and specialist psychological interventions. Cognitive behaviour therapy has been shown to be effective in randomised controlled trials for a variety of functional syndromes (such as non-cardiac chest pain, irritable bowel, chronic pain, and chronic fatigue) and for patients with hypochondriasis.



Randomised controlled trial of cognitive and behavioural treatments for hypochondriasis

Functional symptoms accompanying disease

Functional symptoms are also common in those who also have major disease. For example, after a heart attack or cardiac surgery, minor muscular chest aches and pains may be misinterpreted as evidence of angina, leading to unnecessary worry and disability. Explanation and advice, perhaps in the context of a cardiac rehabilitation programme, may make a substantial contribution to patients' quality of life.

Conclusion

An understanding of the interaction of biological, psychological, interpersonal, and medical factors in the predisposition, precipitation, and perpetuation of functional somatic symptoms allows convincing explanations to provided for patients and effective treatment to be planned.

Important components of general management include effective initial reassurance, a positive explanation, and practical advice. It is also important to identify early those who are not responding and who require additional specific interventions.

The difficulty that health systems have in effectively dealing with symptoms that are not attributable to disease reflects both intellectual and structural shortcomings in current care. The most salient of these is the continuing influence of mind-body dualism on our education and provision of care. In the longer term, scientific developments will break down this distinction. For the time being, it places primary care in a pivotal role in ensuring appropriate care for these patients.

The graph of incidence of common presenting symptoms in US primary care is adapted from Kroenke and Mangelsdorff, *Am J Med* 1989:86: 262-6. The graph of association between number of unexplained physical symptoms and psychiatric disorder is adapted from Kisely et al, *Psychol Med* 1997;27:1011-9. The picture of a shellshocked soldier is reproduced with permission of British Pathe. The graph of effects of cognitive behaviour treatment for hypochondriasis is adapted from Clark DM et al, *Br J Psychiatry* 1998;173:218-25.

Evidence based summary points

- Functional somatic symptoms are common in primary care in all countries and cultures
- Cognitive behaviour therapies are of general applicability
- Antidepressants are of value whether or not patient is depressed

Gureje O, Simon GE, Ustun TB, Goldberg DP. Somatization in cross-cultural perspective: a World Health Organization study in primary care. *Am J Psychiatry* 1997;154:989-95

Kroenke K, Swindle R. Cognitive-behavioral therapy for somatization and symptom syndromes: a critical review of controlled clinical trials. *Psychother Psychosom* 2000;69:205-15

O'Malley PG, Jackson JL, Santoro J, Tomkins G, Balden E, Kroenke K. Antidepressant therapy for unexplained symptoms and symptom syndromes. *J Fam Pract* 1999;48:980-90

Further reading

- Kroenke K, Mangelsdorff D. Common symptoms in ambulatory care: incidence, evaluation, therapy and outcome. *Am J Med* 1989:86: 262-6
- Mayou R, Bass C, Sharpe M. *Treatment of functional somatic symptoms.* Oxford: Oxford University Press, 1995
- Sharpe M, Carson AJ. "Unexplained" somatic syndromes, somatisation: do we need a paradigm shift? *Ann Intern Med* 2001;134:296
- Wessely S, Nimnuan C, Sharpe M. Functional somatic syndromes: one or many? *Lancet* 1999;354:936-9

7 Chronic multiple functional somatic symptoms

Christopher Bass, Stephanie May

The previous article in this series described the assessment and management of patients with functional somatic symptoms. Most such patients make no more than normal demands on doctors and can be helped with the approach outlined. However, a minority have more complex needs and require additional management strategies. These patients typically have a longstanding pattern of presenting with various functional symptoms, have had multiple referrals for investigation of these, and are regarded by their doctors as difficult to help.

Terminology

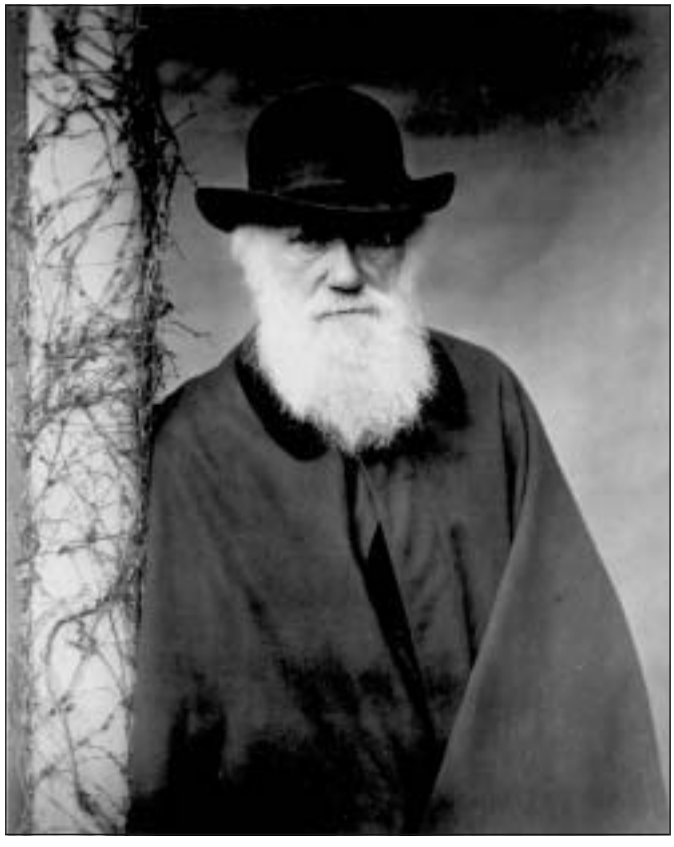
Because such patients may evoke despair, anger, and frustration in doctors, they may be referred to as "heartsink patients," "difficult patients," "fat folder patients," and "chronic complainers." The use of these terms is inadvisable. If patients read such descriptions in their medical notes they are likely to be offended and lose faith in their doctor and may make a complaint. In psychiatric diagnostic classifications these patients are often referred to as having somatisation disorder. We prefer the term "chronic multiple functional symptoms" (CMFS).

Epidemiology and detection

The prevalence of CMFS depends upon the number of different symptoms required for diagnosis and on the setting. Whilst each primary care doctor will have an average of 10-15 of such patients, they are more common in specialist medical settings where they may account for as many as 10% of referrals.

Most patients with CMFS are women. They often have recurrent depressive disorder and a longstanding difficulty with personal relationships and may misuse substances. There is an association with an emotionally deprived childhood and childhood physical and sexual abuse. Some patients will clearly have general disturbances of personality.

The risk of iatrogenic harm from over-investigation and over-prescribing for somatic complaints makes it important that patients with CMFS are positively identified and their management planned, usually in primary care. Potential CMFS patients may be identified simply by the thickness of their paper notes, from records of attendance and hospital referral, and by observation of medical, nursing, or clerical staff.



Charles Darwin (1809-82) suffered from chronic anxiety and varied physical symptoms that began shortly after his voyage in the *Beagle* to South America (1831-6). Despite many suggested medical explanations, these symptoms, which disabled him for the rest of his life and largely confined him to his home, remain medically unexplained

Management in primary care

Assessment

It is helpful if one doctor is identified as a patient's principal carer. Once a patient is identified as possibly having CMFS a systematic assessment is desirable. The case notes should be reviewed and the patient seen for one or more extended consultations.

Case notes—Patients with CMFS often have extensive case notes. Unless these are reviewed, much potentially useful information may remain hidden. It is also helpful to compile a summary of these records and to evaluate critically the accuracy of any previously listed complaints and diagnoses. The summary should include key investigations performed to date and any information about patients' personal and family circumstances.

Long appointment—During one or more long appointment a patient's current problems and history should be fully explored.

"Fat files" are a simple indicator of a high level of contact with medical services, which may indicate multiple chronic functional somatic complaints

Patients should be encouraged to talk not only about their symptoms but also about their concerns, emotional state, and social situation and the association of these with their symptoms. At the end of the assessment, patient and doctor should agree a current problem list, which can then be recorded in the notes.

Management

The initial long interviews serve not only to derive a problem list but also to foster a positive relationship between doctor and patient. Thereafter, the doctor should arrange to see the patient at regular, though not necessarily frequent, fixed intervals. These consultations should not be contingent on the patient developing new symptoms. Consultation outside these times should be discouraged.

Planned review

All symptoms reported by patients during these consultations must be acknowledged as valid. A detailed review of symptoms enhances the doctor-patient relationship and minimises the likelihood of missing new disease.

Reassurance that "nothing is wrong" may be unhelpful, possibly because a patient's aim may be to develop an understanding relationship with the doctor rather than relief of symptoms. Focused physical examination can be helpful, but there is a risk of patients receiving multiple diagnostic tests and referrals to specialists, and these should be minimised. Patients also often accumulate unnecessary prescribed drugs, and if so these should be reduced gradually over time.

If a satisfactory rapport can be established with a patient, new information about his or her emotional state, relationship difficulties, or childhood abuse may be revealed. In such cases the doctor may need to offer the patient a further long appointment to reassess the need for specialist psychological care.

Support for doctors

General practitioners managing patients with CMFS should arrange ongoing support for themselves, perhaps from a partner or another member of the primary care team with whom they can discuss their patients. A doctor and, for example, a practice nurse can jointly manage some of these patients if there is an agreed management plan and clear communication.

Referral to psychiatric services

Not all doctors will consider that they have the necessary skills or time to manage these patients effectively. Review by an appropriate specialist can then be helpful. Unfortunately, the decline in the number of "general physicians" and specialist mental health services' increasing focus on psychotic illness mean there are few appropriate specialists to refer to.

If referral is sought two questions must be considered: "Are there any local and appropriate psychiatric services?" and "How can I prepare the patient for this referral?" If available, liaison psychiatry services are often the most appropriate and experienced in this area of practice. To prepare the patient, a discussion emphasising the distressing nature of chronic illness and the expertise of the services in this area, together with a promise of continuing support from the primary care team, can help to make the referral seem less rejecting. If possible, the psychiatrist should visit the practice or medical department and conduct a joint consultation.

Assessment of chronic multiple functional somatic symptoms

- Elicit a history of the current complaints, paying special attention to recent life events
- Find out what the patient has been told by other doctors (as well as friends, relatives, and alternative practitioners). Does this accord with the medical findings?
- Elicit an illness history that addresses previous experience of physical symptoms and contact with medical services (such as illness as a child, illness of parents and its impact on childhood development, operations, time off school and sickness absence)
- Explore psychological and interpersonal factors in patient's development (such as quality of parental care, early abusive experiences, psychiatric history)
- Interview a partner or reliable informant (this may take place, consent permitting, in the patient's presence)
- After the interview attempt a provisional formulation

Useful interviewing skills for doctors managing patients with multiple physical complaints

- Adopt a flexible interviewing style—"I wonder if you've thought of it like this?"
- Try to remind the patient that physical and emotional symptoms often coexist—"I'm struck by the fact that, in addition to the fatigue, you've also been feeling very low and cannot sleep"
- Try "reframing" the physical complaints to indicate important temporal relationship between emergence of patient's somatic and emotional symptoms and relevant life events
- Respond appropriately to "emotional" cues such as anger
- Explore patient's illness beliefs and worst fears—"What is your worst fear about this pain?"

Management strategy for patients with chronic multiple functional somatic symptoms

- Try to be proactive rather than reactive—Arrange to see patients at regular, fixed intervals, rather than allowing them to dictate timing and frequency of visits
- During appointments, aim to broaden the agenda with patients—This involves establishing a problem list and allowing patients to discuss relevant psychosocial problems
- Stop or reduce unnecessary drugs
- Try to minimise patients' contacts with other specialists or practitioners—This will reduce iatrogenic harm and make containment easier if only one or two practitioners are involved
- Try to co-opt a relative as a therapeutic ally to implement your management goals
- Reduce your expectation of cure and instead aim for containment and damage limitation
- Encourage patients (and yourself) to think in terms of coping and not curing

Explanations to the patient

Present patient's problems as a summary with an invitation to comment:

"So let me see if I've understood you properly: you have had a lot of pain in your abdomen, with bloating and distension for the past four years. You have been attending the (GP) surgery most weeks because you've been very worried about cancer (and about your husband leaving you). You also told me that these pains often occur when you are anxious and panicky, and at these times other physical complaints such as trembling and nausea occur.

"I'm struck by the fact that all these complaints began soon after you had a very frightening experience in hospital, when your appendix was removed and you felt that 'No one was listening to my complaints or pain.'

"Have I got that right, or is there anything I've left out?"

Summary of a 15 year "segment" of the life of a patient with chronic multiple functional somatic symptoms

Date (age)	Symptoms (life events)	Referral	Investigations	Outcome
1970 (18)	Abdominal pain	GP to surgical outpatients	Appendicectomy	Normal appendix
1973 (21)	Pregnant (boyfriend in prison)	GP to obstetrics and gynaecology outpatients	Termination of pregnancy	—
1975-7 (23-25)	Bloating, abdominal pain, blackouts (stressful divorce)	GP to gastroenterology and neurology outpatients	All tests normal	Diagnosis of irritable bowel syndrome and unexplained syncope. Treated with fibre
1979 (27)	Pelvic pain (wants to be sterilised)	GP to obstetrics and gynaecology outpatients	Sterilised, ovaries preserved	Pelvic pain persists for 2 years after surgery
1981 (29)	Fatigue (problems at work)	GP to infectious disease clinic	Nothing abnormal detected	Diagnosis of myalgic encephalomyelitis made by patient. Joins self help group
1983 (31)	Aching, painful muscles	GP to rheumatology clinic	Mild cervical spondylosis. No treatment	Treated with amitriptyline 50 mg on referral to pain clinic. Some improvement
1985 (34)	Chest pain and breathlessness (son truanting from school)	Accident and emergency to chest clinic	Nothing abnormal detected, probable hyperventilation	Refer to psychiatric services

Specialist assessment

Before interviewing a patient, it is useful to request both the general practice and hospital notes and summarise the medical history. A typed summary of the "illness history" can be kept as a permanent record in the notes. This summary can guide future management and is especially useful when a patient is admitted subsequently as an emergency or when the receiving doctor has no prior knowledge of the patient.

Several important interviewing skills should be used during the assessment. These skills can be learnt using structured role playing and video feedback. They form the basis of a technique called reattribution, which has been developed to help the management of patients with functional somatic symptoms.

Specialist management

If a patient can understand and agree an initial shared formulation of the problems, an important first stage is reached. From this a plan of management can be negotiated. It is best to adopt a collaborative approach rather than a didactic or paternalistic manner. If it is difficult to arrive at an understanding of why the patient developed these symptoms at this particular time, then an alternative approach may have to be adopted. In essence this involves the doctor attempting to address those factors that are maintaining the symptoms.

Assessment and management go hand in hand. One of the main aims of management is to modify patients' often unrealistic expectations of the medical profession and to remind them of the limits to medicine. In many cases hopes may have been falsely raised, and patients expect either a cure or at least a considerable improvement in symptoms. Although this is desirable, it may not be attainable. Instead, the doctor should attempt to broaden the agenda, with an emphasis on helping patients to address personal concerns and life problems as well as somatic complaints. It is also necessary to encourage them to concentrate on coping rather than seeking a cure.

This process requires patience, and a capacity to tolerate frustration and setbacks. It may require several discussions in which the same issues are reviewed. In the long term, however, it can be rewarding for both patient and doctor.

Common problems in management

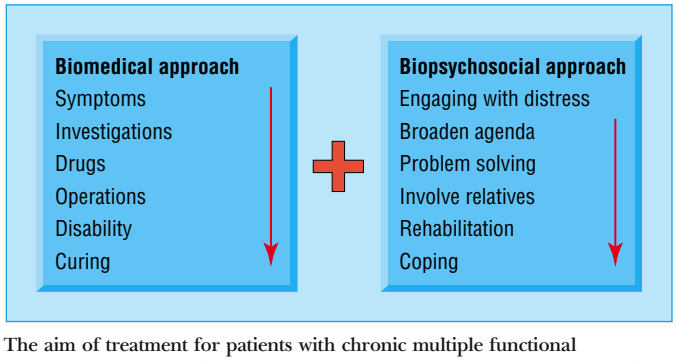
Management may be complicated by various factors. Firstly, preoccupation and anxious concern about symptoms may lead patients to make unhelpful demands of their doctor, which prove difficult to resist.

What is the cause of functional somatic symptoms?

- A variety of biological, psychological, and social factors have been shown to be associated with functional symptoms; the contribution of these factors will vary between patients
  Recent developments in neuroscience show altered functioning of the nervous system associated with functional symptoms, making the labelling of these as "entirely psychological" increasing inappropriate
- With our current knowledge, it is best to maintain "aetiological neutrality" about the cause of functional symptoms
- The main task of treatment is to identify those factors that may be maintaining a patient's symptoms and disability

Maintaining factors that should be focus of treatment in patients with multiple somatic symptoms

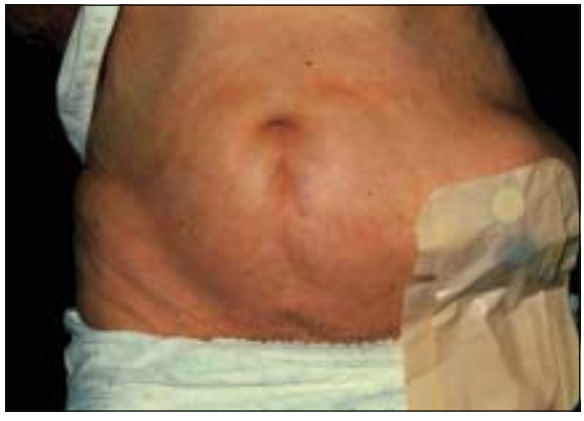
- Depression, anxiety, or panic disorder
- Chronic marital or family discord
- Physical inactivity
- Occupational stress
- Abnormal illness beliefs
- Iatrogenic factors
- Pending medicolegal and insurance claims



The aim of treatment for patients with chronic multiple functional symptoms is to add a biopsychosocial perspective to the existing biomedical approach

Secondly, there may be evidence of longstanding interpersonal difficulties, as indicated by remarks such as "Nobody cares" or "It's disgusting what doctors can do to you." Such comments may suggest that the patient's relationship with the doctor may reflect poor quality parental care or emotional deprivation in childhood. They are important for two reasons: firstly, the doctor may take these remarks personally, become demoralised or angry, and retaliate, which will destroy the doctor-patient relationship; and, secondly, the attitudes revealed may require more detailed psychological exploration.

Finally, iatrogenic factors may intervene that are beyond the treating doctor's control. Because these patients have often visited several specialists, conventional and alternative, they may have been given inappropriate information and advice, inappropriate treatment, or, in some cases, frank misdiagnosis.



Failing to recognise and institute appropriate management for patients with multiple functional somatic symptoms may lead to iatrogenic harm from excessive and inappropriate medical and surgical intervention

Factitious disorders and malingering

Factitious disorders

Factitious disorders are characterised by feigned physical or psychological symptoms and signs presented with the aim of receiving medical care. They are therefore different from functional symptoms. The judgment that a symptom is produced intentionally requires direct evidence and exclusion of other causes. Most patients with factitious disorders are women with stable social networks, and more than half of these work in medically related occupations. Once factitious disorder is diagnosed, it is important to confront the patient but remain supportive. When factitious disorder is established in a person working in health care it is advisable to organise a multidisciplinary meeting involving the patient's general practitioner, a physician and surgeon, a psychiatrist, and a medicolegal representative.

If, and only if, the deliberate feigning of symptoms and signs can be established (such as by observation of self mutilation) should patients be confronted. It is helpful if both a psychiatrist and the referring doctor (who should have met to discuss the aims, content, and possible outcomes of the meeting beforehand) can carry out the confrontation jointly. This "supportive confrontation" is done by gently but firmly telling the patient that you are aware of the role of their behaviour in the illness whilst at the same time offering psychological care to help with this. After confrontation, patients usually stop the behaviour or leave the clinic. Only sometimes do they engage in the psychiatric care offered.

Malingering

A distinction should be made between factitious disorders and malingering. Malingerers deliberately feign symptoms to achieve a goal (such as to avoid imprisonment or gain money). Malingering is behaviour and not a diagnosis. The extent to which a doctor feels it necessary to confront this issue will depend on the individual circumstances.

Conclusion

Patients with multiple longstanding functional symptoms are relatively uncommon, but their interaction with the health system is memorable in that it often leaves both them and their doctors frustrated. Their effective management requires that special attention be paid to their interpersonal difficulties (including those arising in their relationship with the doctor), the limiting of unhelpful demands, and the avoidance of iatrogenic harm. As with any chronic illness, confident management and getting to know a patient as a person can change what is often a frustrating task into a rewarding one.

Münchausen's syndrome

- Münchausen's syndrome is an uncommon subtype of factitious illness in which the patient, who is often a man with sociopathic traits and an itinerant lifestyle, has a long career of attending multiple hospitals with factitious symptoms and signs
- Management is as for factitious disorder, but engagement with psychiatric treatment is rare

Evidence based summary

- Prevalence of chronic multiple functional somatic symptoms depends on how many functional symptoms are required—the fewer symptoms the higher the prevalence
- Patients with chronic multiple functional somatic symptoms (somatisation disorder) can be effectively managed in primary care, with resulting cost savings

Kroenke K, Spitzer RL, deGruy FV, Hahn SR, Linzer M, Williams JB, et al. Multisomatoform disorder. An alternative to undifferentiated somatoform disorder for the somatizing patient in primary care. *Arch Gen Psychiatry* 1997;54:352-8

Smith GR, Monson RA, Ray DC. Psychiatric consultation in somatization disorder—a randomized controlled study. *N Engl J Med* 1986;314:1407-13

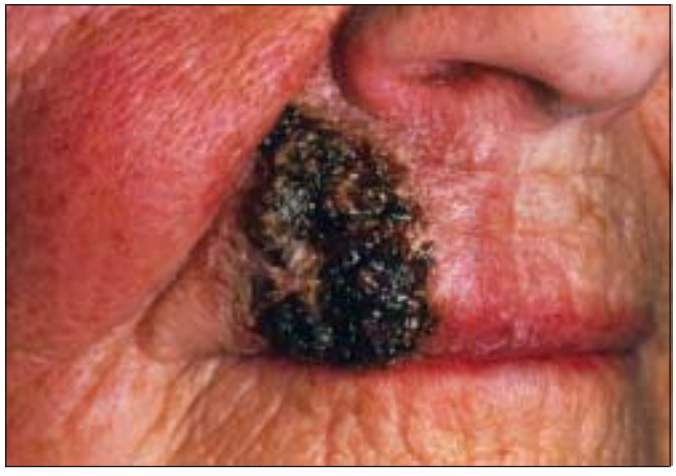
Suggested reading

- Bass C. Management of somatisation disorder. *Prescribers J* 1996;36:198-205
- Dixon DM, Sweeney KG, Pereira Gray DJ. The physician healer ancient magic or modern science? *Br J Gen Pract* 1999;49:309-12
- Fink P, Rosendal M, Toft T. Assessment and treatment of functional disorders in general practice: the extended reattribution and management model—an advanced educational program for nonpsychiatric doctors. *Psychosom* 2002;43:93-131
- Smith GR. Management of patients with multiple symptoms. In: Mayou R, Bass C, Sharpe M, eds. *Treatment of functional somatic symptoms*. Oxford: Oxford University Press, 1995:175-87
- Tate P. *The doctor's communication handbook*. Oxford: Radcliffe Medical Press, 1994

8 Cancer

Craig A White, Una Macleod

Cancer is the most feared of diseases. Unsurprisingly, it causes considerable psychological distress in patients, families, carers, and often those health professionals who care for them. Only a minority of cancer patients develop psychiatric illness, but other psychologically and socially determined problems are common. These include unpleasant symptoms such as pain, nausea, and fatigue; problems with finances, employment, housing, and childcare; family worries; and existential and spiritual doubts. Well planned care that fully involves patients and their families can minimise these problems.



Squamous cell cascinoma on lip after radiotherapy. As well as the fear of cancer itself, an additional source of distress can be the potentially disfiguring nature of the disease and its treatment

Psychological consequences

Though often dismissed as "understandable," distress is a treatable cause of reduced quality of life and poorer clinical outcome. Some patients delay seeking help because they fear or deny their symptoms of distress. Presentation can be obvious, as depressed or anxious mood can manifest as increased severity of somatic complaints such as breathlessness, pain, or fatigue. Adjustment disorder is the commonest psychiatric diagnosis, and neuropsychiatric complications may occur. The risk of suicide is increased in the early stages of coping with cancer.

Depression

Depression is a response to perceived loss. A diagnosis of cancer and awareness of associated losses may precipitate feelings similar to bereavement. The loss may be of parts of the body (such as a breast or hair), the role in family or society, or impending loss of life. Severe and persistent depressive disorder is up to four times more common in cancer patients than in the general population, occurring in 10-20% during the disease.

Anxiety, fear, and panic

Anxiety is the response to a perceived threat. It manifests as apprehension, uncontrollable worry, restlessness, panic attacks, and avoidance of people and of reminders of cancer, together with signs of autonomic arousal. Patients may overestimate the risks associated with treatment and the likelihood of a poor outcome. Anxiety may also exacerbate or heighten perceptions of physical symptoms (such as breathlessness in lung cancer), and post-traumatic stress symptoms (with intrusive thoughts and avoidance of reminders of cancer) occasionally follow diagnosis or treatment that has been particularly frightening.

Certain cancers and treatments are associated with specific fears. Thus, patients with head and neck cancers may worry about being able to breathe and swallow. Patients may develop phobias and conditioned vomiting in relation to unpleasant treatments such as chemotherapy.

Neuropsychiatric syndromes

Delirium and dementia may arise from brain metastases, which usually originate from lung cancer but also from tumours of the breast and alimentary tract and melanomas. Brain metastases occasionally produce psychological symptoms before metastatic disease is discovered. Certain cancers (notably cancers of the lung, ovary, breast, or stomach and Hodgkin's lymphoma) sometimes produce neuropsychiatric problems in the absence of metastases (paraneoplastic syndromes). The aetiology is thought to be an autoimmune response to the tumour.

"Distress is an unpleasant emotional experience of a psychological, social, or spiritual nature that may interfere with a patient's ability to cope with cancer and its treatment. Distress extends along a continuum, ranging from common normal feelings of vulnerability, sadness, and fear to problems that can become disabling, such as depression, anxiety, panic, social isolation, and spiritual crisis"

US National Comprehensive Cancer Network

Challenges faced by people with cancer

- Maintaining activity and independence
- Coping with treatment side effects
- Accepting cancer and maintaining a positive outlook
- Seeking and understanding medical information
- Regulating the feelings associated with cancer experiences
- Seeking support
- Managing stress

Who becomes distressed?

The severity of emotional distress is more closely related to a patient's pre-existing vulnerability than to the characteristics of the cancer. Distress is also more likely to occur at specific points in a patient's experience of cancer:

Diagnosis—Investigation and diagnosis are particularly stressful and can cause shock, anger, and disbelief as well as emotional distress. These resolve without intervention in most patients, but especially high levels of distress at this time are predictive of later emotional problems. It can help if doctors explain that patients' feelings are expected and normal ("I would expect you to have times when you feel tearful and cannot get it out of your head").

During treatment—Treatment itself can be a potent cause of distress. It may involve hospital attendance but also unpleasant surgery, radiotherapy, or chemotherapy. Side effects include hair loss and disfigurement. Patients worry about whether treatment is working and are likely to become distressed at times of apparent treatment failure.

End of treatment—At the end of apparently successful treatment some patients can experience "rebound" distress associated with the fear that the cancer might recur or spread. The ending of a prolonged relationship with the cancer service staff can lead to a sense of loss and vulnerability. It is only at this time that some patients become fully aware of the impact of their cancer experience.

After treatment—Like those with other life threatening illnesses, patients who survive cancer may reorder their life priorities and experience psychological benefits including a greater appreciation of some aspects of their life. Others need help to overcome continuing worries, including preoccupation with loss and illness, a tendency to avoid reminders of cancer, and difficulties coping with intimacy, return to work, and fears of recurrence. Fear of recurrence can manifest as a form of health anxiety with misinterpretation of physiological sensations (such as believing that pain associated with a muscle strain represents a recurrence of cancer) and the anxious seeking of reassurance.

Recurrence—Patients who believe they have been cured (that is, those most likely to be surprised by recurrence) are at greater risk of severe distress if recurrence occurs. Most patients report recurrence of cancer as more distressing than receiving the initial diagnosis.

Terminal disease—About 40% of people who develop cancer will die as a result. The terminal phase commonly brings fear of uncontrolled pain, of the process of dying, of what happens after death, and of the fate of loved ones. Depression is common in the terminal phase, especially in those with poorly controlled physical symptoms.

Risk factors for psychiatric disorder

Patient
- History of psychiatric disorder
- Social isolation
- Dissatisfaction with medical care
- Poor coping (such as not seeking information or talking to friends)

Cancer
- Limitation of activities
- Disfiguring
- Poor prognosis

Treatment
- Disfiguring
- Isolating (such as bone marrow transplant)
- Side effects

Issues to be considered in planning care

- Patient's and family's understanding of the illness and its treatment
- Patient's and family's understanding of help available
- Explanation of how symptomatic relief will be provided
- How the patient can be fully involved in care
- Who will be managing the treatment plan
- Routine and emergency contact arrangements
- Practical help in everyday activities
- Support at home—role of hospital and residential care
- Involving and supporting family and friends

Depression is common in the terminal phase of cancer, especially in patients with poorly controlled physical symptoms (*Resignation* by Carl Wilhelm Wilhelmson (1866-1928))

Management

People with cancer benefit from care in which psychological and medical care are coordinated. Apart from the obvious

Psychological care for cancer patients

In primary care
- Need for agreed local protocols
- Multidisciplinary skills and resources
- Individually agreed collaborative care for each patient
- Regular liaison with specialist units and local agencies
- Local training for all involved

In specialist units
- Training in psychological aspects of care for all staff
- Regular review of all individual treatment plans
- Protocols for routine management of "at risk" patients (such as relapse after chemotherapy)
- Involvement of specialist nurses and other staff with psychological expertise
- Access to psychiatrists and clinical psychologists with special interest in managing cancer problems for consultation and supervision
- Use of self help methods and voluntary agencies

benefits to quality of life, there is some evidence that encouraging an active approach to living with cancer can improve survival.

Most of the psychological care of cancer patients will be delivered in primary care. As for all chronic illness, a multidisciplinary approach and management protocols that include psychological as well as medical assessment and intervention are required. These protocols need not be specific for cancer as the issues are common to many medical conditions. The important point is that the staff involved have the skills to address psychological as well as medical problems. The danger is that psychological care can be neglected by the medical focus on cancer treatment. A case manager, whether nurse or doctor, who can coordinate the often diverse agencies involved in a cancer patient's care can ensure that treatment is delivered efficiently.

Assessment

Depressive and anxiety disorders are often unrecognised. There is therefore a need for active screening by simply asking patients about symptoms of anxiety and depression. A self rated questionnaire such as the hospital anxiety and depression scale (HADS) may be helpful. Doctors should be aware that patients may be distressed because of factors unrelated to cancer.

Treatment

Information—Doctors often underestimate the amount and frankness of information that most patients need and want. It is best given in a staged fashion with checks on patients' understanding and desire to hear more at each stage. Repetition and written information may be helpful. Summaries of agreed management plans have been found to improve patients' satisfaction and their adherence to medical treatment.

Social support—Most patients will receive this from family and friends. They may, however, not want to "burden others" and consequently may need encouragement to use this support by talking about their illness. Additional support can be provided by specific cancer related services such as the primary care team and specialist nurses.

Addressing worries—Staff often find it most difficult to help patients who talk about worries that reflect the reality of cancer (such as, "I am going to die"). It is important to do so because this may help planning and may reveal misconceptions, such as the inevitability of uncontrolled pain, that can then be addressed by giving accurate information about methods of pain control.

Managing anxiety—Accurate information (such as which physical symptoms are due to anxiety and which are due to cancer) and practical help are important. Anxious patients can be helped by relaxation strategies, including breathing exercises. Severe persistent anxiety may merit the short term prescription of anxiolytic drugs such as diazepam.

Managing depression—Depressive disorders should be managed in the same ways as they are in patients without cancer. Discussion, empathy, reassurance, and practical help are essential. Antidepressants have been shown to be effective in patients with cancer in randomised trials, although surprisingly few trials have been conducted. If in doubt about what drug to choose or about possible interactions with cancer treatment, it is important to check with a pharmacist. Specialist psychological intervention, such as formal cognitive-behavioural therapy, may also be required to treat persistent depression or anxiety.

Specialist referral

Structured psychological interventions (such as psycho-education and cognitive-behavioural based therapies)

Questions for assessing patients' anxiety and depression

- How are you feeling in yourself? Have you felt low or worried?
- Have you ever been troubled by feeling anxious, nervous, or depressed?
- What are your main concerns or worries at the moment?
- What have you been doing to cope with these? Has this been helpful?
- What effects do you feel cancer and its treatment will have on your life?
- Is there anything that would help you cope with this?
- Who do you feel you have helping you at the moment?
- Is there anyone else outside of the family?
- Have you any questions? Is there anything else you would like to know?

Principles of treatment

- Sympathetic interest and concern
- A clearly identified principal therapist who can coordinate all care
- Effective symptomatic relief
- Elicit and understand patient's beliefs and needs
- Collaborative planning of continuing care
- Information and advice—oral and written
- Involve patient in treatment decisions
- Involve family and friends
- Early recognition and treatment of psychological complications
- Clear arrangements to deal with urgent problems

Useful sources of information

- National Comprehensive Cancer Network. Distress management guidelines (www.nccn.org/physician_gls/index.html)
- National Cancer Institute. Cancer.gov (www.cancer.gov/cancer_information/)
- Cancer BACUP (www.cancerbacup.org.uk)
- Cancer Help UK (www.cancerhelp.org.uk/)
- Macmillan Cancer Relief (www.macmillan.org.uk/)
- Cancer Research UK (www.cancer.org.uk)
- International Psycho-Oncology Society (www.ipos-aspboa.org/iposnews.htm)

Specialist treatments

- Antidepressant drugs
- Effective drug treatment of pain, nausea, and other symptoms
- Problem solving discussion
- Cognitive-behavioural treatment of psychological complications
- Joint and family interviews to encourage discussion and planning
- Group support and treatment
- Cognitive-behavioural methods to help cope with chemotherapy and other unpleasant treatments

Referral decisions

- What specialist expertise in psycho-oncology is available at my local cancer centre or unit?
- What has helped when this patient has had problems before?
- Are there local cancer support groups that could help?
- Does this patient have problems that might benefit from specialist psychological or psychiatric intervention?
- Does this patient want to be referred to specialist services?
- Does this patient prefer individual or group based psychological intervention?

have been shown to reduce anxiety and depression in cancer patients and to improve adherence to medical treatment.

Patients with severe or persistent distress may need referral to an experienced clinical psychologist or psychiatrist. An increasing number of mental health professionals are attached to cancer centres and units, and other staff such as appropriately trained specialist nurses play an increasingly important role.

Increasing numbers of non-NHS agencies also offer psychological care for patients with cancer. When referring patients to such services it is important to check their quality and to ensure that their contribution is coordinated within an overall care plan.

The picture of skin cancer is reproduced with permission of Dr P Marazzi and Science Photo Library. *Resignation* is held at the Nationalmuseum, Stockholm, and is reproduced with permission of Bridgeman Art Library.

Evidence based summary

- Antidepressants are effective in treating depressed mood in cancer patients
- Cognitive-behavioural treatments are effective in relieving distress, especially anxiety, and in reducing disability
- Psychological interventions can be effective in relieving specific cancer related symptoms such as breathlessness

McDaniel JS, Musselman DL, Porter MR, Reed DA, Nemeroff CB. Depression in patients with cancer. Diagnosis, biology, and treatment. *Arch Gen Psychiatry* 1995;52:89-99

Sheard T, Maguire P. The effect of psychological interventions on anxiety and depression in cancer patients: results of two meta-analyses. *Br J Cancer* 1999;80:1770-80

Bredin M, Corner J, Krishnasamy M, Plant H, Bailey C, A'Hern R. Multicentre randomised controlled trial of nursing intervention for breathlessness in patients with lung cancer. *BMJ* 1999;318:901-4

Further reading

- Barraclough J. *Cancer and emotion : a practical guide to psycho-oncology*. 3rd ed. Chichester: John Wiley, 1998
- Burton M, Watson M. *Counselling patients with cancer*. Chichester: John Wiley, 1998
- Faulkener A, Maguire P. *Talking to cancer patients and their relatives*. Oxford: Oxford Medical Publications, 1994
- Holland JC. *Psycho-oncology*. Oxford: Oxford University Press, 1998
- Lewis S, Holland JC. *The human side of cancer: living with hope, coping with uncertainty*. London: Harper Collins, 2000
- Scott JT, Entwistle V, Sowden AJ, Watt I. Recordings or summaries of consultations for people with cancer. Cochrane Database of Systematic Reviews. 2001

9 Trauma

Richard Mayou, Andrew Farmer

Minor physical trauma is a part of everyday life, and for most people these injuries are of only transient importance, but some have psychiatric and social complications. Most people experience major trauma at some time in their lives.

Psychological, behavioural, and social factors are all relevant to the subjective intensity of physical symptoms and their consequences for work, leisure, and family life. As a result, disability may become greater than might be expected from the severity of the physical injuries.

Psychological and interpersonal factors also contribute to the cause of trauma, and clinicians should be alert to these and their implications for treatment. Tactful questioning, careful examination, and detailed record keeping are essential, especially for non-accidental injury by a patient or others:

- Ask for a detailed description of the cause of the incident
- Ask about previous trauma
- Ask about substance misuse—alcohol and drugs
- Look for patterns of injuries that may be non-accidental, deliberate self harm, or inflicted by others
- Check records
- If suspicious speak to other informants
- Discuss findings and suspicions with a colleague.

Detail of *Very Slippy Weather* by James Gillray (1757-1815)

Lifetime prevalence of specific traumatic events (n=2181)

Type of trauma	Prevalence
Assault	38%
Serious car or motor vehicle crash	28%
Other serious accident or injury	14%
Fire, flood, earthquake, or other natural disaster	17%
Other shocking experience	43%
Diagnosed with a life threatening illness	5%
Learning about traumas to others	62%
Sudden, unexpected death of close friend or relative	60%
Any trauma	90%

Dealing with the acute event

At a major incident it is important that members of the emergency services, especially ambulance staff and police, should seem calm and in control. This helps to relieve distress and prevent victims from suffering further injury. Explanation and encouragement can reduce fear at the prospect of being taken to hospital by ambulance. The needs of uninjured relatives and others involved should also be considered. Clearly recorded details of the incident, injury, and the extent of any loss of consciousness may be useful in later assessment as well as in the preparation of subsequent medicolegal reports.

Many people attend hospital emergency departments for minor cuts, bruises, or pain, or for "a check up" after being involved in an incident, whereas others attend their general practitioner. Immediate distress is common. Clear explanation, advice, and discussion at the outset can prevent later problems in returning to normal activities and enable early recognition of psychological and social consequences. A sympathetic approach is needed that includes suitable analgesia, reassurance about the likely resolution of symptoms, and encouragement to return to normal activity. Some patients may already be considering compensation, and records should be kept with this in mind.

Immediate effects of frightening trauma

- Causes a varied picture of anxiety, numbness, dissociation (feeling distanced from events, having fragmentary memories), and sometimes apparently inappropriate calmness
- Those who believe they are the innocent victims of others' misbehaviour are often angry, and this may be exacerbated by subsequent frustrations
- The term "acute stress disorder" is now used for a combination of distress, intrusive memories (flashbacks, nightmares), avoidance, and numbing in the months after the trauma. It occurs in 20-50% of those who have suffered major trauma
- The severity of emotional symptoms is much more closely related to how frightening the trauma was than to the severity of the injury; even uninjured victims may suffer considerable distress
- Severe distress is usually temporary but indicates a risk of long term post-traumatic symptoms

Immediate management

- Physical treatment, including adequate analgesia
- Sympathetic discussion of acute distress
- Explanation and appropriate reassurance about treatment and prognosis
- Appropriate encouragement for graded return to work and other activities
- Indicate what help will be available for continuing psychological symptoms and social problems
- Information and support to relatives

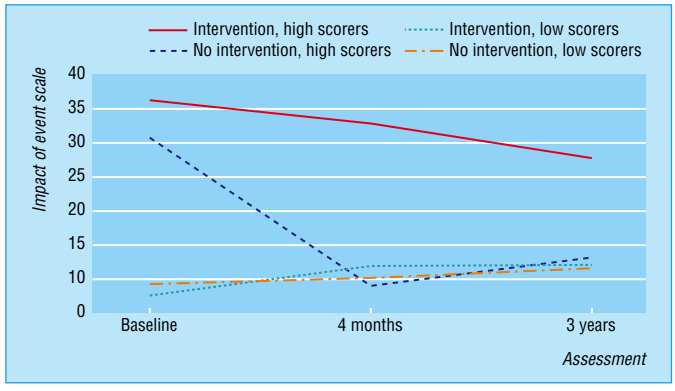
Advice about return to work and other activities

Patients with painful injuries that should improve within days or weeks are often uncertain how to behave and how soon to return to work. The assessment is an opportunity to give advice about this. Patients need information on the cause of their symptoms, their likely impact on daily life, and a positive plan for return to normal activity; this includes discussing the type of work normally done, the employer's attitude to time away from work, and opportunities for a graded increase in activity. Good, rapid communication between hospital and primary care is essential.

Immediate psychological interventions

Many employers and medical and voluntary groups recommend routine "debriefing" after frightening trauma. However, the evidence shows this is not only ineffective but may be harmful.

It is better, therefore, to concentrate on the immediate relief of distress through support and sympathetic reassurance and on practical help, while encouraging further early consultation if problems persist. This is especially so in groups who may be regularly exposed to frightening and distressing circumstances, such as members of the armed forces, police, and ambulance staff. Severe immediate distress and perception of the trauma as having been very frightening indicate an increased risk of chronic post-traumatic symptoms, and early review is recommended to identify those who need extra help. Victims of crime can be helped by referral to the charity Victim Support.



Effect of immediate debriefing on victims of road traffic injury. Those with high initial scores on the impact of events scale (intrusive thoughts and avoidance) had worse outcome than untreated controls at 4 months and 3 years

Later consequences and care

Treatment should include clear, agreed plans for mobilisation and return to optimal activity. Physiotherapists are often involved in rehabilitation and need to be aware of the psychological as well as the physical factors that are perpetuating disability. If necessary, a multidisciplinary approach should be established.

Chronic pain and disability

A small number of those who have suffered trauma continue to complain of physical symptoms and disabilities that are difficult to explain. Investigations are negative or ambiguous, and the relationship between doctors and patients may become fraught. Doctors may feel their patient is disabled for psychological reasons, whereas patients may feel that doctors do not believe that their symptoms are real and that they are unsympathetic and are not offering appropriate treatment.

Arguments about whether symptoms are physical or psychological are rarely helpful. Instead, it is essential to agree a coordinated behavioural and rehabilitative approach with patient and family that aims to achieve the maximum improvement. Unfortunately, there is a shortage of appropriate multidisciplinary specialist services for such people. This leaves primary care teams in the key role in monitoring progress and implementing a biopsychosocial approach to rehabilitation.

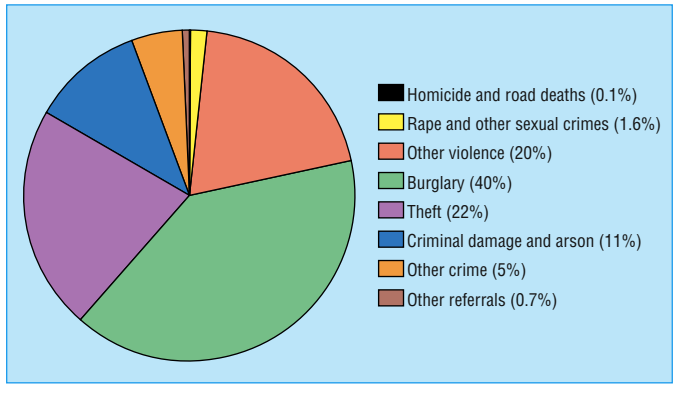
Psychological symptoms and syndromes

Depression, post-traumatic stress disorder, and phobic anxiety are common after frightening trauma and can be severe, whether or not there is evidence of previous psychological and social vulnerability. These psychological complications are not closely related to the severity of any physical injury. The general principles of assessment are those for similar psychological problems occurring in the absence of trauma.

Depression—A failure to recognise depression is distressingly common, perhaps because care focuses on physical injuries. Inquiries about depressive symptoms should therefore be routine.

Post-traumatic stress disorder is also common and disabling. It is characterised by intrusive memories of the trauma, avoidance of reminders of it, and chronic arousal and distress. It may be complicated by alcohol misuse. It usually has an early onset in the first few weeks (acute stress disorder). Many people improve rapidly but, if symptoms are still present two or three months after the injury, they are likely to persist for much longer. A few cases have a delayed onset. Psychological treatment is effective.

Phobic anxiety may be associated with post-traumatic stress disorder but can occur separately. A particularly common form



Reasons for people being offered help by Victim Support 1997-8

Unexplained and disproportionate disability and pain

- Lack of explanation or overcautious advice often leads to misunderstandings and secondary disability
- Delays in assessment and treatment exacerbate problems and make treatment more difficult
- Lack of coordination (between general practice, physiotherapy, hospital, etc) frequently exacerbates problems
- Low mood, misunderstandings, and inactivity worsen pain and disability
- Agree on consistent, collaborative plans with patient and family
- Early access to specialist rehabilitation and pain clinics providing high quality cognitive and behavioural psychological treatments

Psychologically determined consequences of trauma

- Acute anxiety, numbing, arousal (acute stress disorder)
- Anxiety disorder
- Major depressive disorder
- Post-traumatic symptoms and disorder
- Avoidance and phobic anxiety
- Pain and apparently disproportionate disability
- Unexplained physical symptoms
- Impact on family (such as family arguments, depression in family members)

Cognitive behavioural approach to treating post-traumatic stress disorder

- *Talking it through*—Encourage victim to discuss and relive feelings about the incident
- *Tackling avoidance*—Discuss graded increase in activities, such as return to travel after a road crash
- *Coping with anxiety*—Anxiety management techniques (relaxation, distraction)
- *Dealing with anger*—Encourage discussion of incident and of feelings
- *Overcoming sleep problems*—Emphasise importance of regular sleep habits and avoidance of excessive alcohol and caffeine
- *Treat associated depression*—Antidepressant drugs, limited role for hypnotics immediately after trauma

is anxiety about travel, both as a driver and as a passenger, after a road traffic crash. This anxiety may lead to distress and limitation of activities and lifestyle. Early advice about the use of anxiety management techniques and the need for a graded return to normal travel is helpful, but more specialist behavioural treatment may be required and is usually effective.

Detection of psychological problems
During a clinical assessment, a few brief screening questions can be useful as a guide to identify depression, anxiety, post-traumatic stress disorder and drinking problems. It is often helpful to speak to someone close to the victim who can offer an independent view.

Personal injury and compensation
Victims who believe that others are to blame for their trauma increasingly consult specialist lawyers, who are alert to psychiatric complications such as post-traumatic stress disorder and phobic avoidance. Acrimonious discussion about a small number of controversial cases of alleged exaggeration and simulation has obscured a more productive discussion of psychiatric disorder.

Head injury
Most head injuries are mild. These were once believed to be without consequences, but recent evidence has suggested that almost half of patients experiencing mild head injuries (Glasgow coma scale 13-15) remain appreciably disabled a year later. The effects of more severe head injuries on personality and cognitive performance may be greater than is apparent in a clinical interview and commonly affect "executive" functions such as social judgment and decision making.

Such deficits are often not detected by standard bedside screening tools such as the mini-mental state examination. Patients with head injury should therefore not be pushed to return to demanding activities too quickly, and there should be a low threshold for seeking a specialist opinion or undertaking psychometric assessment.

Consequences for others

Family members may also suffer distress, especially if they have been involved in the traumatic incident. Seeing the relatives of the traumatised person is usually helpful in the management of persistent problems.

Those involved in treating trauma will encounter particularly distressing incidents with severely injured victims and distraught relatives. These often occur when those involved in treatment are working under considerable pressure. Clear procedures for training and support of staff are essential. For those working in large emergency services the provision of regular specialist support is advisable.

Types of trauma

The pattern of consequences varies with the type of trauma experienced. All services that see trauma emergencies need management plans for psychological as well as medical care. This includes planning for major events in which there are many victims and for the much commoner road traffic and other incidents in which there are often several victims, some of whom may be severely injured and who may well be related or know one another. Emergency departments and primary care need procedures for helping the patients and for supporting the staff that are involved.

Treating avoidance and phobic anxiety

- *Diary keeping*—Encourage detailed diary of activity and associated problems as a basis for planning and monitoring progress
- *Anxiety (stress) management*—Relaxation, distraction, and cognitive procedure for use in stressful situations
- *Graded practice*—Discuss a hierarchy of increasing activities; emphasise importance of not being overambitious and need to be consistent in following step by step plan

Compensation

- Simulation of disability and exaggeration are uncommon in routine clinical contacts
- Many victims want recognition of their suffering as much as financial compensation
- Innocent victims of trauma are generally slower to return to work than those victims who accept that they were to blame
- Financial and social consequences of trauma and blighting of ambitions may be considerable and are often unrecognised
- Compensation procedures and reports may hinder development and agreement about treatment and active rehabilitation
- Compensation may allow interim payments and funding of specialist care to treat complications and prevent chronic disability

Head injury

- Assessment should involve questions about possible unconsciousness and post-traumatic amnesia
- Cognitive consequences of minor head injury are often not recognised
- Minor impairments may be obscured in clinical situations but be disabling in work and everyday activities
- Recovery may be prolonged
- Complaints of confusion and poor memory can be due to depression
- Specialist assessment may be needed

Relatives' needs

Immediately after severe or frightening trauma
- Make comfortable
- Inform relatives of trauma in a sympathetic manner
- Practical assistance
- Clear information

Later
- Information about injuries, treatment, and prognosis
- Discuss effects on everyday life
- Discuss needs for practical help and availability
- Ask about possible psychiatric problems and indicate help available

Types of trauma

- *Occupational*—Return to work often slower than in other types of injury. Liaison with employer essential. Compensation issues may impede return to work
- *Sporting*—May be associated with physical unfitness or with inappropriate activity for age
- *Domestic*—Assess role of alcohol, consider possible family and other problems, assess risk of further incidents
- *Assault (including sexual)*—Assess role of alcohol, keep detailed records, suggest availability of help for major, and especially for sexual, assault
- *Road traffic crash*—Psychological complications may occur even if no significant physical injury. Whiplash injuries should be treated by well planned mobilisation and encouragement, together with alertness to possible psychological complications

Disasters

All medical services and other institutions should have a disaster plan that is readily available and regularly reviewed. It should include a specification for immediate psychological care and information, together with proactive follow up so that psychological problems are identified early. Those involved in coping with disasters also require support and encouragement, and a minority may require specialist psychological help. The disaster plan should also set out procedures for giving information to relatives and offering them practical help.

Conclusion

The psychological aspects of trauma may be important, even when injury seems trivial. Clear, sympathetic care, which takes account of patients' needs, can do much to promote optimal recovery. Specialist advice should be sought for persistent problems within the first few months of an injury. Long delays in providing adequate assessment and treatment lead to unnecessary suffering and disability and may make such problems much more difficult to treat.

The print *Very Slippy Weather* is reproduced with permission of Leeds Museum and Art Galleries and Bridgeman Art Library. The table of lifetime prevalence of traumatic events is adapted from Breslau et al. *Arch Gen Psychiatry* 1998;55:626-32. The graph of effect of immediate debriefing on the psychiatric wellbeing of victims of road traffic injury is adapted from Mayou et al *Br J Psychiatry* 2000;176:590-4. The figure showing reasons for people being offered help by Victim Support is adapted from *Information in the Criminal Justice System in England and Wales. Digest 4*, London: Home Office, 1999.

Evidence based summary

- Cognitive behaviour therapy is effective in treating post-traumatic stress disorder
- Early critical incident debriefing after trauma is potentially harmful

Sherman JJ. Effects of psychotherapeutic treatments for PTSD: a meta-analysis of controlled clinical trials. *J Trauma Stress* 1998;11:413-36

Wessely S, Rose S, Bisson J. Brief psychological interventions ("debriefing") for trauma-related symptoms and the prevention of post traumatic stress disorder *Cochrane Database Syst Rev* 2999;(2):CD00050

Suggested reading

- Mayou RA, Bryant B. Outcome in consecutive emergency department attenders following a road traffic accident. *Br J Psychiatry* 2001;179:528-34
- McDonald AS, Davey GCL. Psychiatric disorders and accidental injury. *Clin Psychol Rev* 1996;16:105
- NIH Consensus Development Panel on Rehabilitation of Persons with Traumatic Brain Injury. Rehabilitation of persons with traumatic brain injury. *JAMA* 1999;282:974-83

10 Fatigue

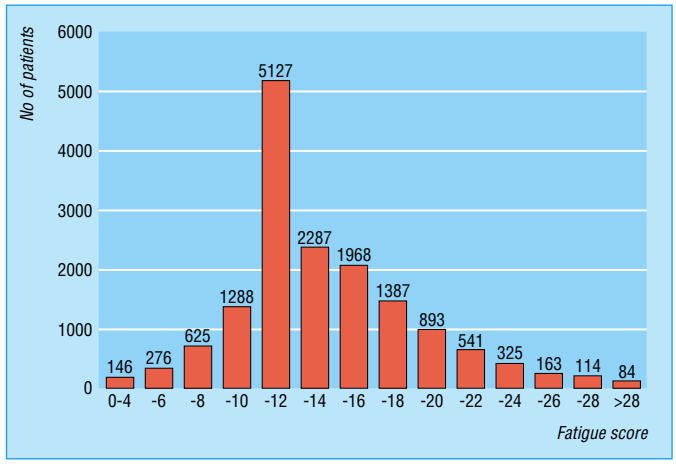
Michael Sharpe, David Wilks

Fatigue can refer to a subjective symptom of malaise and aversion to activity or to objectively impaired performance. It has both physical and mental aspects. The symptom of fatigue is a poorly defined feeling, and careful inquiry is needed to clarify complaints of "fatigue," "tiredness," or "exhaustion" and to distinguish lack of energy from loss of motivation or sleepiness, which may be pointers to specific diagnoses (see below).

Prevalence—Like blood pressure, subjective fatigue is normally distributed in the population. The prevalence of continuously significant fatigue depends on the threshold chosen for severity (usually defined in terms of associated disability) and persistence. Surveys report that 5-20% of the general population suffer from such persistent and troublesome fatigue. Fatigue is twice as common in women as in men but is not strongly associated with age or occupation. It is one of the commonest presenting symptoms in primary care, being the main complaint of 5-10% of patients and an important subsidiary symptom in a further 5-10%.

Fatigue as a symptom—Patients generally regard fatigue as important (because it is disabling), whereas doctors do not (because it is diagnostically non-specific). This discrepancy is a potent source of potential difficulty in the doctor-patient relationship. Fatigue may present in association with established medical and psychiatric conditions or be idiopathic. Irrespective of cause, it has a major impact on day to day functioning and quality of life. Without treatment, the prognosis of patients with idiopathic fatigue is surprisingly poor; half those seen in general practice with fatigue are still fatigued six months later.

Weary 1887 by Edward Radford (1831-1920)



Distribution of the complaint of fatigue in the population

Causes of fatigue

The physiological and psychological mechanisms underlying subjective fatigue are poorly understood. Fatigue may rather be usefully regarded as a final common pathway for a variety of causal factors. These can be split into predisposing, precipitating, and perpetuating factors.

Predisposing factors include being female and a history of either fatigue or depression.

Precipitating factors include acute physical stresses such as infection with Epstein-Barr virus, psychological stresses such as bereavement, and social stresses such as work problems.

Perpetuating factors include physical inactivity, emotional disorders, ongoing psychological or social stresses, and abnormalities of sleep. These factors should be sought as part of the clinical assessment.

Other physiological factors such as immunological abnormalities and slightly low cortisol concentration are of research interest but not clinical value.

Diagnoses associated with fatigue

Among patients who present with severe chronic fatigue as their main complaint, only a small proportion will be suffering from a recognised medical disease. In no more than 10% of patients presenting with fatigue in primary care is a disease cause found. The rate is even lower in patients seen in secondary care.

Fatigue is a major symptom of many psychiatric disorders, but for a substantial proportion of patients with fatigue the

Medical conditions that may present with apparently unexplained fatigue

- *General*—Anaemia, chronic infection, autoimmune disease, cancer
- *Endocrine disease*—Diabetes, hypothyroidism, hypoadrenalism
- *Sleep disorders*—Obstructive sleep apnoea and other sleep disorders
- *Neuromuscular*—Myositis, multiple sclerosis
- *Gastrointestinal*—Liver disease
- *Cardiovascular*—Chronic heart disease
- *Respiratory*—Chronic lung disease

symptom remains unexplained or idiopathic. In general, the more severe the fatigue and the larger the number of associated somatic (and unexplained) complaints, then the greater the disability and the greater the likelihood of a diagnosis of depression.

Chronic fatigue syndromes

Chronic fatigue syndrome is a useful descriptive term for prominent physical and mental fatigue with muscular pain and other symptoms. It overlaps with another descriptive term, fibromyalgia, that has often been used when muscle pain is predominant but in which fatigue is almost universal. There is also substantial overlap of the diagnoses with other symptom based syndromes, the so called functional somatic syndromes.

The term myalgic encephalomyelitis (or encephalopathy) has been used in Britain and elsewhere to describe a poorly understood illness in which a prominent symptom is chronic fatigue exacerbated by activity. This is a controversial diagnosis that some regard as simply another name for chronic fatigue syndrome and that others regard as a distinct condition. This article will focus on chronic fatigue syndrome.

Prevalence and outcome—Chronic fatigue syndrome can be diagnosed in up to 2% of primary care patients. Untreated, the prognosis is poor, with only about 10% of patients recovering in a two to four years. A preoccupation with medical causes seems to be a negative prognostic factor.

Psychiatric diagnoses commonly associated with fatigue

- Depression
- Anxiety and panic
- Eating disorders
- Substance misuse disorders
- Somatisation disorder

Diagnostic criteria for chronic fatigue syndrome

Inclusion criteria

- Clinically evaluated, medically unexplained fatigue of at least 6 months' duration that is
  Of new onset (not life long)
  Not result of ongoing exertion
  Not substantially alleviated by rest
  Associated with a substantial reduction in previous level of activities
- Occurrence of 4 or more of the following symptoms
  Subjective memory impairment, sore throat, tender lymph nodes, muscle pain, joint pain, headache, unrefreshing sleep, post-exertional malaise lasting more than 24 hours

Exclusion criteria

- Active, unresolved, or suspected medical disease or psychotic, melancholic, or bipolar depression (but not uncomplicated major depression), psychotic disorders, dementia, anorexia or bulimia nervosa, alcohol or other substance misuse, severe obesity

Assessment and formulation

History—The nature of the fatigue is an important clue to diagnosis, and it is therefore important to clarify patients' complaints. Fatigue described as loss of interest and enjoyment (anhedonia) points to depression. Prominent sleepiness suggests a sleep disorder. The history should also cover

- Systematic inquiry for diseases and medications often associated with fatigue
- Symptoms of depression anxiety and sleep disorder
- Patients' own understanding of their illness and how they cope with it
- Current social stresses.

Examination—Both a physical and mental state examination must be performed in every case, to seek medical and psychiatric diagnoses associated with fatigue.

Routine investigations—If there are no specific indications for special investigations, a standard set of screening tests is adequate.

Special investigations—Immunological and virological tests are generally unhelpful as routine investigations. Sleep studies can be useful in excluding other diagnoses, especially obstructive sleep apnoea and narcolepsy.

Psychological assessment—It is important to inquire fully about patients' understanding of their illness (questions may include "What do you think is wrong with you?" and "What do you think the cause is?"). Patients may be worried that the fatigue is a symptom of a severe, as yet undiagnosed, disease or that activity will cause a long term worsening of their condition.

Formulation—A formulation that distinguishes predisposing, precipitating, and multiple perpetuating factors is valuable in providing an explanation to patients and for targeting intervention.

Screening tests for fatigue

- Full blood count
- Erythrocyte sedimentation rate or C reactive protein
- Liver function tests
- Urea, electrolytes, and calcium
- Thyroid stimulating hormone and thyroid function tests
- Creatine kinase
- Urine and blood tests for glucose
- Urine test for protein

Factors to consider in a formulation of chronic fatigue

	Predisposing cause	Precipitating cause	Perpetuating cause
Biological	Biological vulnerability	Acute disease	Pathophysiology Excessive inactivity Sleep disorder Side effects of drug treatment Untreated disease
Psychological	Vulnerable personality	Stress	Depression Unhelpful beliefs about cause Fearful avoidance of activity
Social	Lack of support	Life events Social or work stress	Reinforcement of unhelpful beliefs Social or work stress

General management

Persistent fatigue requires active management, preferably before it has become chronic. When a specific disease cause of fatigue

can be identified this should be treated. If no disease diagnosis can be made, or if medical treatment of disease fails to relieve the fatigue, a broader biopsychosocial management strategy is required. A discussion with the patient about fatigue and its treatment can be supplemented with written material (see below).

Patients should be told that they are suffering from a common and treatable condition that the doctor takes seriously and for which behavioural treatment can be helpful. While patients may be concerned about possible disease and the need for medical investigation and treatment, it can be explained that no disease has been found, and hence there is no disease based treatment, but that with help there is a great deal that the patients can do themselves.

Identifying unhelpful beliefs—Potentially unhelpful beliefs should be discussed. If a patient has a simple aetiological model (such as "It is all due to a virus") an alternative approach based on a biopsychosocial formulation can be outlined. This has the advantage of highlighting potential perpetuating factors, as these may be regarded as obstacles to recovery. Doctor and patient can then work together to overcome these. It is rarely productive to argue over the best name for the illness; instead, the emphasis should be on agreeing a positive and open minded approach to rehabilitation.

Managing activity and avoidance—Gradual increases in activity can be advised unless there is a clear contraindication. It is critical, however, to distinguish between carefully graded increases carried out in collaboration with patients and "forced" exercise. It is also important to explain that erratic variation between overactivity on "good" days and subsequent collapse does not help long term recovery and that "stabilising" activity is a prerequisite to graded increases.

Depression and anxiety—If there is evidence of depression a trial of an antidepressant drug is worth while. Patients with fatigue are often sensitive to the side effects of antidepressants. However, if they are given adequate information about what to expect when treatment begins, with small doses, most patients can tolerate them. Randomised trials have shown psychological therapies such as cognitive behaviour therapy to be equally effective for mild to moderate depression.

Managing occupational and social stresses—Patients who remain in work may be overstressed by it. Those who have left work may be inactive and demoralised and may not wish to return to the same job. These situations require a problem solving approach to consider how to manage work demands, achieve a return to work, or to plan an alternative career.

Drug Treatments for Fatigue—A variety of pharmacological drugs including stimulants and steroids have been advocated for the treatment of fatigue. There is a limited evidence base for any of these pharmacological treatments, most of which may lead to substantial adverse effects. The role for these drugs is therefore limited and they should only be prescribed with great caution.

Referral for specialist management

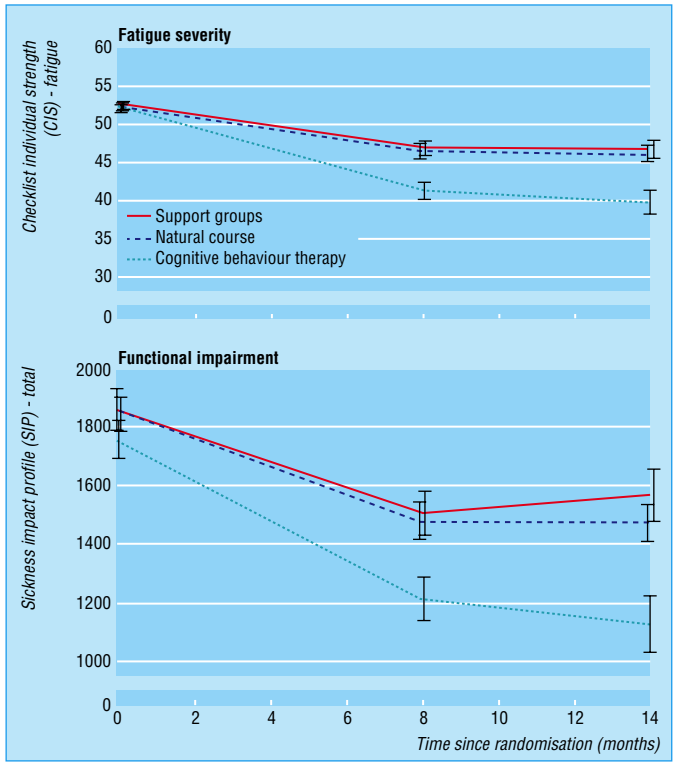
Most patients with fatigue are managed in primary care, but certain groups may require referral to specialist care:

- Children with chronic fatigue
- Patients in whom the general practitioner suspects occult disease
- Patients with severe psychiatric illness
- Patients requiring specialist management of sleep disorders
- Patients unresponsive to management in primary care.

Management of chronic fatigue

1 Assessment
  Empathise
  History
  Examination
  Limited investigation
  Biopsychosocial formulations
2 Treat treatable medical and psychiatric conditions
3 Help patient to overcome perpetuating factors
  Educate
  Reduce distress
  Gradual increase in activity
  Solve social and occupational problems
4 Follow up

Patients should be encouraged to gradually increase their activity ("Mrs Bradbury's establishment for the recovery of ladies nervously affected")



Efficacy of cognitive behaviour therapy for treating chronic fatigue syndrome

What is cognitive behaviour therapy?

- Brief pragmatic psychological therapy
- Targets beliefs and behaviours that might perpetuate symptoms
- An established treatment for depression and anxiety
- Has been adapted for somatic complaints of pain and fatigue
- Requires a skilled therapist

Referral may be to a physician or psychiatrist as is deemed most appropriate. Psychologists may be able to offer cognitive behaviour therapy. Where available, joint medical and psychiatric clinics are ideally suited to the assessment of chronic fatigue and related problems. It is essential there is close liaison between primary and specialist care to ensure a clear, consistent, and encouraging approach by all concerned.

Rehabilitation

Rehabilitation based on behavioural principles is currently the most effective specialist treatment approach.

Cognitive behaviour therapy is a collaborative psychological rehabilitation that incorporates graded increases in activity but also pays greater attention to patients' beliefs and concerns.

Graded exercise therapy is a structured progressive exercise programme administered and carefully monitored by a therapist.

Both may be used in conjunction with antidepressant drugs. Both have been found to be effective in randomised trials of hospital referred cases of chronic fatigue syndrome. Some general practitioners are able to provide graded exercise or cognitive behaviour therapy in their practice or clinic. Others may wish to refer to a trained therapist.

Conclusion

Fatigue is a ubiquitous symptom that is important to patients and has a major impact on their quality of life. It remains poorly understood and has hitherto probably been not been given adequate attention by doctors. Early and active management of fatigue in primary care may prevent progression to chronicity. Patients who have developed a chronic fatigue syndrome can benefit from specific treatments. Paying more attention to the symptom of fatigue may help to avoid the distress and poor outcome that is associated with patients feeling that their problems are neither accepted nor understood. It may also reduce the numbers who turn to a variety of unproved, and even harmful, alternative approaches.

What is graded exercise therapy?

- Explanation of fatigue as a physiological consequence of inactivity, poor sleep, and disturbed circadian rhythms
- Discussion, agreement, and implementation of graded exercise plans
- Monitoring of progress and setting of appropriate new targets

Evidence based summary

- Chronic fatigue syndrome is a descriptive term for a disabling syndrome that probably has multiple causes (physical and psychological)
- Graded exercise and cognitive behaviour therapies are effective in treating chronic fatigue syndrome

Wessely S. Chronic fatigue: symptom and syndrome. *Ann Intern Med* 2001;134:838-43

Whiting P, Bagnall AM, Sowden AJ, Cornell JE, Mulrow CD, Ramirez G. Interventions for the treatment and management of chronic fatigue syndrome: a systematic review. *JAMA* 2001;286:1360-8

Further reading

- Wessely S, Hotopf M, Sharpe M. *Chronic fatigue and its syndromes.* Oxford: Oxford University Press, 1998
- Campling F, Sharpe M. *Chronic fatigue syndrome: the facts.* Oxford: Oxford University Press, 2000
- Reid S, Chalder T, Cleare A, Hotopf M, Wessely S. Chronic fatigue syndrome. *Clinical Evidence* 2001 (Nov)

The painting *Weary* is held at Russell-Cotes Art Gallery and Museum, Bournemouth, and is reproduced with permission of Bridgeman Art Library. The graph of distribution of fatigue in the population is adapted from Pawlikowska T, et al *BMJ* 1994;308:763-6. The box of diagnostic criteria for chronic fatigue syndrome is adapted from Fukuda K, et al *Ann Intern Med* 1994;121:953-9. The print of "Mrs Bradbury's establishment for the recovery of ladies nervously affected" is reproduced with permission of Wellcome Library. The graph showing efficacy of cognitive behaviour therapy is adapted from Prins JB, et al *Lancet* 2001;357:841-7.

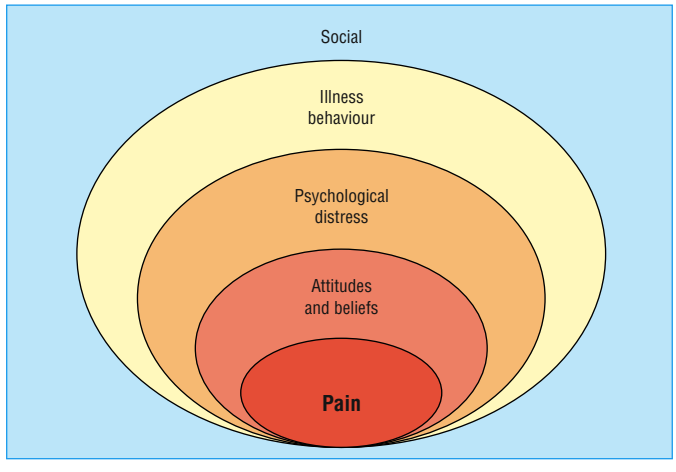
11 Musculoskeletal pain

Chris J Main, Amanda C de C Williams

Musculoskeletal symptoms of various types (neck pain, limb pain, low back pain, joint pain, chronic widespread pain) are a major reason for consultation in primary care. This article uses the example of low back pain because it is particularly common and there is a substantial evidence base for its management. The principles of management outlined are also applicable to non-specific musculoskeletal symptoms in general.

The increasing prevalence of musculoskeletal pain, including back pain, has been described as an epidemic. Pain complaints are usually self limiting, but if they become chronic the consequences are serious. These include the distress of patients and their families and consequences for employers in terms of sickness absence and for society as a whole in terms of welfare benefits and lost productivity. Many causes for musculoskeletal pain have been identified. Psychological and social factors have been shown to play a major role in exacerbating the biological substrate of pain by influencing pain perception and the development of chronic disability. This new understanding has led to a "biopsychosocial" model of back pain.

Research has also shown that there are many different reasons for patients to consult their doctor with pain—seeking cure or symptomatic relief, diagnostic clarification, reassurance, "legitimisation" of symptoms, or medical certification for work absence or to express distress, frustration, or anger. Doctors need to clarify which of these reasons apply to an individual and to respond appropriately.



Biopsychosocial model of the clinical presentation and assessment of low back pain and disability at a point in time

Managing acute back pain

Most patients can be effectively managed with a combination of brief assessment and giving information, advice, analgesia, and appropriate reassurance. Minimal rest and an early return to work should be encouraged. Explanation and advice can be usefully supplemented with written material.

Doctors' tasks include not only the traditional provision of diagnosis, investigation, prescriptions, and sickness certificates but also giving accurate advice, information, and reassurance. Primary care and emergency department doctors are potentially powerful therapeutic agents and can provide effective immediate care, but they may also unintentionally promote progression to chronic pain. The risk of chronicity is reduced by

- Paying attention to the psychological aspects of symptom presentation
- Avoiding unnecessary, excessive, or inappropriate investigation
- Avoiding inconsistent care (which may cause patients to become overcautious)
- Giving advice on preventing recurrence (such as by sensible lifting and avoiding excessive loads).

Research evidence supports a change of emphasis from treating symptoms to early prevention of factors that result in progression to chronicity. This has led to the development of new back pain management guidelines for both medical management and occupational health. The shift in emphasis from rest and immobilisation to active self management requires broadening the focus of the consultation from examination of symptoms alone to assessment, which includes

Excerpt from information booklet *The Back Book

It's your back

Backache is not a serious disease and it should not cripple you unless you let it. We have tried to show you the best way to deal with it. The important thing now is for you to get on with your life. How your backache affects you depends on how you react to the pain and what you do about it yourself.

There is no instant answer. You will have your ups and downs for a while—that is normal. But look at it this way

There are two types of sufferer

One who avoids activity, and one who copes

- The *avoider* gets frightened by the pain and worries about the future
- The *avoider* is afraid that hurting always means further damage—it doesn't
- The *avoider* rests a lot and waits for the pain to get better
- The *coper* knows that the pain will get better and does not fear the future
- The *coper* carries on as normally as possible
- The *coper* deals with the pain by being positive, staying active, or staying at work

*Roland M et al, Stationery Office, 2002.

patients' understanding of their pain and how they behave in response to it. The shift towards self directed pain management recasts the role of primary care doctor to the more rewarding one of guide or coach rather than a mere "mechanic."

Identify risk factors for chronicity

Guidelines for primary care management of acute back pain highlight the identification of risk factors for chronicity. A useful approach has been developed in New Zealand. It aims to involve all interested parties—patient, the patient's family, healthcare professionals, and, importantly, the patient's employer. Four groups of risk factors or "flags" for chronicity are accompanied by recommended assessment strategies, which include the use of screening questionnaires, a set of structured interview prompts, and a guide to behavioural management. The focus is on key psychological factors or "yellow flags" that favour chronicity:

- The belief that back pain is due to progressive pathology
- The belief that back pain is harmful or severely disabling
- The belief that avoidance of activity will help recovery
- A tendency to low mood and withdrawal from social interaction
- The expectation that passive treatments rather than active self management will help.

The assessment of "red flags" will identify the small number of patients who need referral for an urgent surgical opinion. Similarly, patients with declared suicidal intent require immediate psychiatric referral. These two groups of patients need to be managed separately.

For the vast majority of patients, however, the identification of contributory psychological and social factors should be seen as an investigation of the normal range of reactions to pain rather than the seeking of psychopathology. Questions in the form of interview prompts have been designed to elicit potential psychosocial barriers to recovery in the "yellow flags" system. They can be used at the time of initial presentation by the general practitioner.

Establish collaboration

Recent studies of miscommunications between doctors and patients with pain show that adequate assessment and collaborative management cannot be achieved without good communication between doctors and patients: only then will patients fully disclose their concerns.

The essence of good communication is to work toward understanding a patient's problem from his or her own perspective. In order to do this, the doctor must first gain the patient's confidence. A patient who has been convinced that the doctor takes the pain seriously will give credence to what the doctor says. Unfortunately, the converse is more common, and patients who feel that a doctor has dismissed or under-rated their pain are unlikely to reveal key information or to adhere to treatment advice.

Enhance accurate beliefs and self management strategies

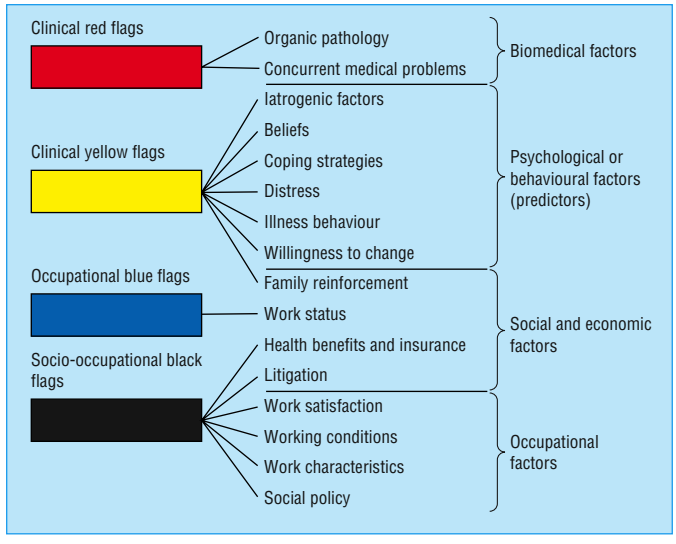
It is easy to overlook the value of simple measures. Many patients respond positively to clear and simple advice, which enables them to manage and control their own symptoms.

Factors associated with chronicity and outcome

Distress
- Symptom awareness and concern
- Depressive reactions; helplessness

Beliefs about pain and disability
- Significance and controllability
- Fears and misunderstandings about pain

Behavioural factors
- Guarded movements and avoidance patterns
- Coping style and strategies



The clinical flags approach to obstacles to recovery from back pain and aspects of assessment

Structured interview prompts
- What do you understand is the cause of your back pain?
- What are you expecting will help you?
- How are others responding to your back pain (employer, coworkers, and family)?
- What are you doing to cope with back pain?
- Have you had time off work in the past with back pain?
- Do you think that you will ever return to work? When?

Guidelines for collaborative management of patients with pain
- Listen carefully to the patient
- Carefully observe the patient's behaviour
- Attend not only to what is said but also how it is said
- Attempt to understand how the patient feels
- Offer encouragement to disclose fears and feelings
- Offer reassurance that you accept the reality of the pain
- Correct misunderstandings or miscommunications about the consultation
- Offer appropriate challenges to unhelpful thoughts and biases (such as catastrophising)
- Understand the patient's general social and economic circumstances

Examples of simple management strategies
- Explain the difference between "hurt" and "harm"
- Reassure patients about the future and the benign nature of their symptoms
- Help patients regain control over pain
- Get patients to "pace" activities—that is, perform activities in manageable, graded stages
- Advise that analgesic drugs be taken on a regular rather than a pain contingent basis
- Set realistic goals such as small increases in activity
- Suggest rewards for successful achievement (such as listening to some favourite music)

Some of these strategies may seem self evident or even trivial, but they are not. Only by building confidence slowly is it possible to prevent the development of invalidity. Occasionally patients will seem to "get stuck" and become demoralised or distressed. Suggesting ways to enhance positive self management can help maintain progress towards a more satisfactory lifestyle.

The success of the cognitive and behavioural approach described below has stimulated the development of secondary prevention programmes designed to prevent those with low back pain from becoming chronically incapacitated by it. Intervention programmes based on cognitive behaviour therapy have also been shown to be effective in reducing disability.

Manage distress and anger

If patients show evidence of distress or anger, find out why. Various strategies for dealing with distress and anger have been developed.

Ways of enhancing positive self management

Get patients to

- Identify when they are thinking in unrealistic, unhelpful ways about their pain (such as "It will keep getting worse") and to change to making a more balanced positive evaluation
- Notice when they are becoming tense or angry and then take steps to interrupt their thoughts and to use relaxation strategies
- Change how they respond when the pain gets bad (such as pause and take a break)
- Document their progress
- Elicit and use the help of others to establish and maintain successful coping strategies

Managing disabling chronic back pain

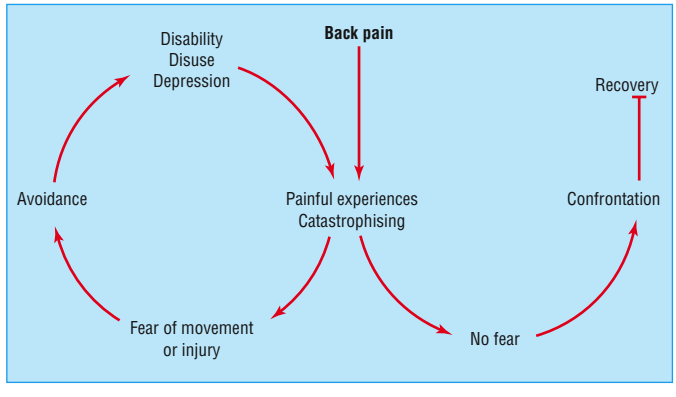
A minority of patients become increasingly incapacitated and require more detailed management of what has become a chronic pain problem. Research has shown that the most important influences on the development of chronicity are psychological rather than biomechanical. The psychological factors are high levels of distress, misunderstandings about pain and its implications, and avoidance of activities associated with a fear of making pain worse.

For patients with established chronic disabling pain specialist referral is required. The treatment of choice is an interdisciplinary pain management programme (IPMP). In these programmes the focus is changed from pain to function, with particular emphasis on perceived obstacles to recovery.

These pain management programmes address the clinical flags. The most commonly used therapeutic approach is a cognitive-behavioural perspective with emphasis on self management. Treatment approaches based on cognitive and behavioural principles have been found to be more effective than traditional biomedical or biomechanically oriented interventions.

Key strategies for assessing and managing distress and anger associated with pain

- Distinguish distress associated with pain and disability from more general distress
- Identify iatrogenic misunderstandings
- Identify mistaken beliefs and fears
- Try to correct misunderstandings
- Identify iatrogenic distress and anger
- Listen and empathise
- Above all, don't get angry yourself



Effects of confrontation or avoidance of pain on outcome of episode of low back pain: fear of movement and re-injury can determine how some people recover from back pain while others develop chronic pain and disability

Specific chronic pain syndromes

Many specific and more widespread pain syndromes have been described—such as "chronic pain," late whiplash syndrome, chronic widespread pain, fibromyalgia, somatoform pain disorder, repetitive strain disorder. It seems unlikely that these are distinct entities, and they are best seen as overlapping descriptive terms that do not have specific aetiological significance. Multidisciplinary treatment that includes psychological, behavioural, and psychiatric assessment and interventions is usually required.

Defining characteristics of modern pain management programmes

- Focus on function rather than disease
- Focus on management rather than cure
- Integration of specific therapeutic ingredients
- Multidisciplinary management
- Emphasis on active rather than passive methods
- Emphasis on self care rather than simply receiving treatment

Conclusion

There needs to be a revolution in the day to day management of musculoskeletal pain. Not only do we need to abandon prolonged rest and enforced inactivity as a form of treatment, but we also need to appreciate that addressing patients' beliefs, distress, and coping strategies must be an integral part of management if it is to be effective.

Lessons learnt in the management of chronic low back pain have direct relevance to the early and specialist management of musculoskeletal pain in general.

The photograph of a man with back pain is reproduced with permission of John Powell/Rex. The figure showing the biopsychosocial model of low back pain is adapted from Waddell G, *The back pain revolution*, Edinburgh: Churchill Livingstone, 1998. The figure showing the clinical flags approach to assessing back pain and the box of defining characteristics of modern pain management programmes are adapted from Main CJ and Spanswick CC, *Pain management: an interdisciplinary approach*, Edinburgh: Churchill-Livingstone, 2000. The boxes of guidelines for collaborative management of patients with pain, of key strategies for managing distress and anger associated with pain, of structured interview prompts, and of ways to enhance positive self management are adapted from Main CJ and Watson PJ, in Gifford L, ed, *Topical issues in pain*, vol 3, Falmouth: CNS Press (in press). The figure showing effects of confrontation or avoidance of pain on outcome of episode of low back pain is adapted from Vlaeyen JWS et al, *J Occup Rehabil* 1995;5:235-52.

Evidence based summary

- Acute back pain is best treated with minimal rest and rapid return to work and normal activity
- Psychological and behavioural responses to pain and social factors are the main determinants of chronic pain disability
- Specialist psychological treatments and pain management programmes are effective in treating chronic pain

Burton AK, Waddell G, Tillotson KM, Summerton N. Information and advice to patients with back pain can have a positive effect. A randomised controlled trial of a novel educational booklet in primary care. *Spine* 1999;24:2484-91

Linton SJ. A Review of psychological risk factors in back and neck pain. *Spine* 2000;25:1148-56

Morley SJ, Eccleston C, Williams A. Systematic review and meta-analysis of randomised controlled trials of cognitive behaviour therapy and behaviour therapy for chronic pain in adults, excluding headache. *Pain* 1999;80:1-13

Further reading

- Clinical Standards Advisory Group. *Clinical Standards Advisory Group report on back pain*. London: HMSO, 1994
- Kendall NAS, Linton SJ, Main CJ. *Guide to assessing psychosocial yellow flags in acute low back pain: risk factors for long term disability and work loss*. Wellington, NZ: Accident Rehabilitation and Compensation Insurance Corporation of New Zealand and the National Health Committee, 1997
- Royal College of General Practitioners. *Clinical guidelines for the management of acute low back pain*. London: RCGP, 1996
- Waddell G, Burton K. *Occupational health guidelines for the management of low back pain at work—evidence review*. London: Faculty of Occupational Medicine, 2000
- Roland M, Waddell G, Klaber-Moffett J, Burton AK, Main CJ. *The back book*. 2nd ed. Norwich: Stationery Office, 2002

12 Abdominal pain and functional gastrointestinal disorders

Elspeth Guthrie, David Thompson

Various functional gastrointestinal pain syndromes have been defined, but there is substantial overlap between them. There is also substantial overlap with other functional disorders such as chronic fatigue syndrome, fibromyalgia, and chronic pelvic pain. The classification system for functional gastrointestinal disorders (FGID) therefore remains controversial and is seldom used outside specialist and research settings. Furthermore, the psychological management of these different syndromes is essentially similar.

In primary care about half of the patients seen with gut complaints have FGID, the most common disorder being irritable bowel syndrome. A UK general practitioner is estimated to see eight patients with irritable bowel syndrome every week, one of whom will be presenting for the first time.

The quality of life of patients with chronic FGID is far poorer than in the general population, and is even significantly lower than in patients with many other chronic illnesses. These patients are not merely the "worried well." It is also important to resist the temptation to think of FGID as exclusively psychological disorders. A biopsychosocial approach is preferable. Physiological studies have suggested that patients with FGID have abnormal visceral sensation and abnormal patterns of bowel motility. Both psychological and physiological factors are involved, with the relative contribution of these varying among patients.

Aetiological factors include physiological and psychological predisposition, early life experience, and current social stresses. It has been shown that a combination of psychological factors and sensitisation of the gut after infection can trigger irritable bowel syndrome in adults.

Emotional distress—The degree of associated emotional distress with FGID depends on the treatment setting. In the community and general practice the prevalence of psychological distress in patients with functional abdominal pain is about 10-20%, whereas in clinic and outpatient settings it is 30-40%, and is even higher for patients who are "treatment resistant."

Abuse—Women with severe FGID often have a history of sexual and emotional abuse. This is as high as 30% in those attending gastroenterology clinics.

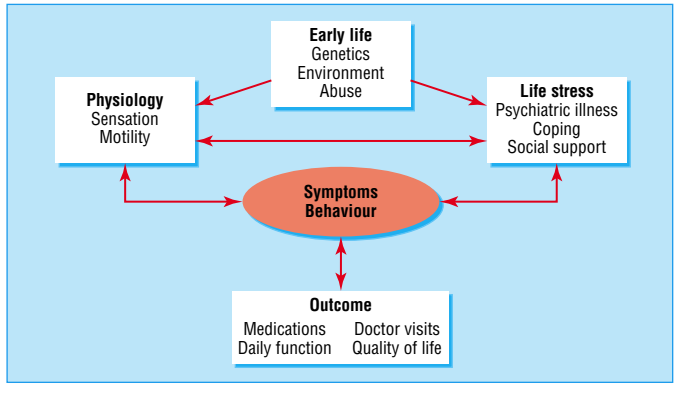
Functional gastrointestinal disorders

- Functional dyspepsia
- Ulcer-like dyspepsia
- Dysmotility-like dyspepsia
- Unspecified dyspepsia
- Functional diarrhoea
- Functional constipation
- Irritable bowel syndrome
- Functional abdominal bloating
- Unspecified functional bowel disorder
- Functional abdominal pain syndrome
- Unspecified functional abdominal pain

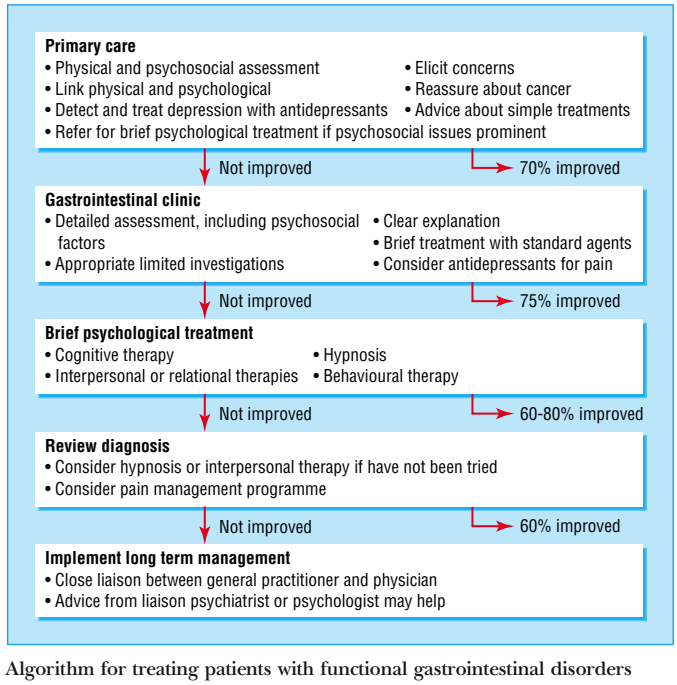
Diagnostic criteria for irritable bowel syndrome

In preceding 12 months at least 12 weeks of abdominal discomfort with 2 of 3 features: relieved with defecation, onset associated with change in frequency of stool, onset associated with change in form of stool
Supportive symptoms include

- Fewer than 3 bowel movements a week
- More than 3 bowel movements a day
- Straining during bowel movement
- Urgent bowel movements
- Feeling of incomplete bowel movement
- Hard or lumpy stools
- Loose or watery stools
- Passing mucus
- Abdominal fullness, bloating, or swelling



Biopsychosocial model for functional abdominal pain



Algorithm for treating patients with functional gastrointestinal disorders

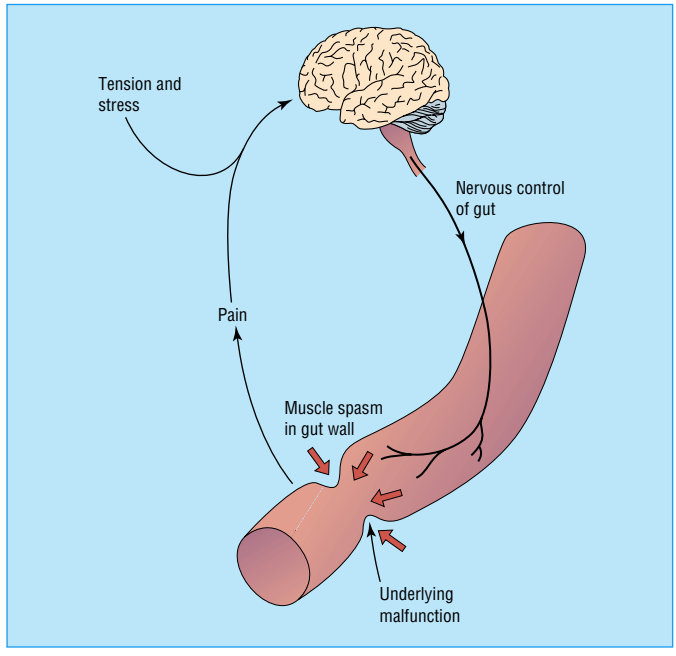
Initial management

Most patients with FGID have relatively mild symptoms and can be managed effectively in primary care. Only a third of patients seen in primary care with irritable bowel syndrome are referred to gastrointestinal specialists for further assessment and treatment.

Symptomatic treatment—Drug treatments for FGID are aimed at improving the predominant symptoms, such as constipation, diarrhoea, abdominal pain, or upper gastrointestinal symptoms. Standard treatments for lower bowel symptoms, depending on the predominant symptom, include dietary fibre, laxatives, antispasmodic agents (including anticholinergics and direct smooth muscle relaxants), and antidiarrhoeals. Treatment for upper gastrointestinal symptoms include H_2 receptor antagonists and prokinetics. There are several useful reviews of the efficacy of these agents in FGID (see further reading).

Psychological management—Initial management can be enhanced by incorporating brief psychological management strategies. Many patients with FGID are afraid that they have a serious underlying disease such as cancer, and attempts should be made to elicit such fears and address them. It is also important to provide a positive and credible explanation for the symptoms. The explanation should include both physiological and psychological factors. One way of explaining symptoms is to describe how the bowel is a segmented tube in which food is propelled down by the sequential squeezing of each segment. The nervous control of this system is delicate and complicated, and disruption of it consequently produces muscle spasm in the bowel wall, which results in pain and gas. Stress and other psychological factors such as anxiety cause bowel symptoms by affecting this nervous control.

Antidepressants—A recent meta-analysis of 12 randomised controlled trials of antidepressants for treating FGID concluded that they are moderately effective. On average, 3.2 patients need to be treated to substantially improve one patient's symptoms. Antidepressants should therefore be considered if there is clear evidence of a depressive disorder, but they may also help to reduce pain in the absence of depression.



Explanation of how physiological and psychological factors combine to produce abdominal pain

Management of chronic problems

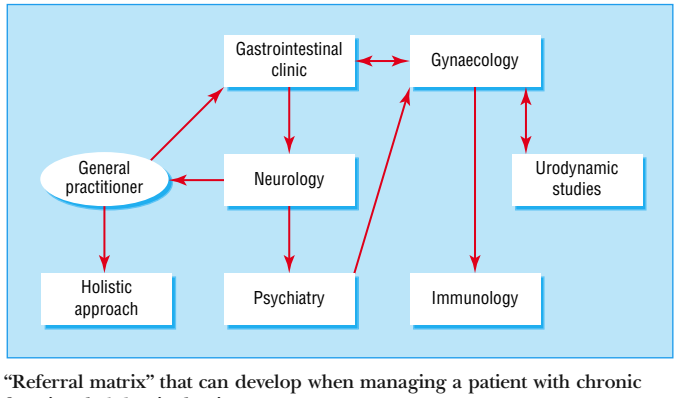
In the case of patients with chronic symptoms that have not responded to treatment, psychological factors are likely to be important. Doctors should try to elicit patients' concerns, seek evidence of emotional distress, and, over several consultations if necessary, help them to make tentative connections between psychological factors, life stresses, and the pain.

The following strategies are suggested:

- Set aside an appointment that is longer than usual, so there is time to deal with a patient's concerns. This is better than several fruitless, rushed consultations focusing only on symptoms
- Make sure that any investigations are based on the patient's history and examination. Do not allow yourself to be pushed into ordering investigations that are not clinically indicated. Try to avoid setting up a "referral matrix," with the patient being referred on from one specialty to another
- Emphasise the role that patients can play in improving or relieving pain by carrying out agreed strategies or exercises. Include the patient in decision about treatment options. Encourage membership of self help groups and organisations. The International Federation for Functional Gastrointestinal Disorders is a well respected organisation that provides useful information for patients. For patients with irritable bowel syndrome, the IBS Network is UK based and is also helpful
- Avoid changing treatments too often; improvement will be slow. Patients are likely to raise concerns about their condition at every consultation, so be prepared to give an explanation of the symptoms more than once. Make a note in the records of what you have said so that you don't contradict yourself
- Be prepared for patients to continually question your approach and think about ways to address this before each consultation. It may be helpful to discuss your management with a psychologist or psychiatrist with a special interest in somatic problems, even if patients do not wish to be referred for psychological treatment
- If you are concerned about a potential complaint, keep a detailed record of consultations, including any requests for investigations and the medical reasons for not ordering them. Repeated investigations that are not medically indicated can be unhelpful in increasing a patient's illness concerns. If you are worried about possible litigation, discuss the situation with a colleague and ask him or her to review the notes

Management of chronic functional abdominal pain

- Set the agenda
- Provide unambiguous information about findings
- Time planning: a longer planned session may save time in long run
- Identify psychosocial factors
- Set limits for investigations
- Encourage patient to take responsibility
- Don't treat what patient doesn't have



"Referral matrix" that can develop when managing a patient with chronic functional abdominal pain

Helpful patient organisations

- International Foundation for Functional Gastrointestinal Disorders. www.iffgd.org/
- IBS Network. http://homepages.uel.ac.uk/C.P.Dancey/ibs.html

- The aim of treatment should be to improve patients' symptoms and functioning rather than to abolish them. Although some patients may remain chronically disabled despite treatment, appropriate and consistent management can prevent deterioration and protect patients from unnecessary surgery.

Referral for psychological treatment

For patients who have not responded to initial management, four different kinds of psychological treatment have been evaluated in FGID. They are cognitive therapies, behavioural therapies, interpersonal therapies, and hypnosis. Each therapy has a different mechanism of action, but they have the common aims of reducing symptoms and improving functioning. Most treatments are delivered on a one to one basis, once weekly, over a period of two to four months.

Although most trials indicate a positive outcome for psychological treatment, many have methodological flaws and further studies are required before definitive recommendations about treatment can be given. The most convincing evidence for the efficacy of specific psychological treatments is for patients with chronic or refractory abdominal symptoms. However, there may also be an important role for earlier intervention in order to prevent such long term difficulties.

Psychological treatments are not always available. As in any other specialty, therapists need to have experience of treating chronic abdominal pain or chronic bowel disorders to develop and retain competence. Psychological services based in primary care are an option for patients with mild to moderate symptoms, but counsellors are unlikely to develop the expertise to enable them to treat patients with severe or refractory abdominal symptoms. Similarly, referral to a psychiatrist or psychologist who is more used to managing severe mental illness is unlikely to be fruitful. Dedicated medical liaison services with experience of somatic problems are more likely to be effective. If these do not exist consideration should be given to establishing a hospital based psychological medicine service.

The diagram of a biopsychosocial model for functional abdominal pain is adapted from Drossman DA et al, *Gut* 1999;45(suppl):II25-30.

Psychological treatments

Cognitive therapy
- Modifies patients' maladaptive beliefs about their pain and symptoms
- Encourages associated behaviour changes
- Patients keep diaries to monitor pain and other symptoms, associated thoughts, and behaviour
- As therapy progresses, it may be possible to identify underlying beliefs or fears about pain that drive preoccupation and worry
- Therapeutic work directed at activating three change mechanisms:
  1 Rational self analysis or self understanding (patients explore idiosyncratic beliefs and fears and connect these to their pain)
  2 Decentring (patients gain distance from their selves by identifying their self talk and labelling it)
  3 Experiential disconfirmation (patients challenge their fears or irrational beliefs through planned behavioural experiments)

Behavioural therapies
- Focus on changing behaviour; they do not address motives or fears
- Patterns that reinforce abnormal behaviour are identified and reversed
- Activity is gradually increased, particularly for functional activities such as social recreation and physical exercise
- Pain behaviours are ignored and activity related behaviours are reinforced
- Patients usually receive educational packages to increase their understanding of the condition
- Anxiety management strategies often included in treatment
- Biofeedback can be used to teach patients to reduce tension in affected muscles and to promote relaxation as a coping strategy

Interpersonal therapies
- Focus on resolving difficulties in interpersonal relationships that underlie or exacerbate abdominal symptoms
- Key problem areas include unresolved grief or loss, role transitions, and relationship discord
- Initial focus is on the patient's abdominal symptoms, which are explored in great detail
- Emotional distress and abnormal feeling states arising from or linked to physical symptoms are identified
- Key problem areas in relationships and their link to physical and psychological symptoms are understood
- Maladaptive relationship patterns, which may have developed after key childhood experiences (such as sexual abuse) are identified
- Solutions to interpersonal difficulties are tested out in therapy and implemented in real world

Hypnosis
- Directed at general relaxation
- Hypnosis is induced using an arm levitation technique, which is followed by deepening procedures
- General positive comments about health and wellbeing are made
- Patients are asked to place their hand on abdomen, feel a sense of warmth, and relate this to asserting control over gut function
- This is reinforced with visualisation (if patient has ability to do this)
- Sessions are concluded with positive, ego strengthening suggestions
- After third session patients are given a tape for daily autohypnosis

Further reading

- Thompson WG, Heaton KW, Smyth GT, Smyth C. Irritable bowel syndrome in general practice: prevalence, characteristics, and referral. *Gut* 2000;46:78-82
- Drossman DA, Creed FH, Fava GA, Olden KW, Patrick DL, Toner BB, et al. Psychosocial aspects of the functional gastrointestinal disorders. *Gastroenterol Int* 1995.1995:8:47-90
- Jailwala J, Imperiale TF, Kroenke K. Pharmacologic treatment of the irritable bowel syndrome: a systematic review of randomized, controlled trials. *Ann Intern Med* 2000;133:135-47
- Akehusrt R, Kaltenthaler E. Treatment of irritable bowel syndrome: a review of randomized controlled trials. *Gut* 2001;48:272
- Bytzer P. H_2 receptor antagonists and prokinetics in dyspepsia: a critical review. *Gut* 2002;50(suppl IV):58-62

Evidence based summary

- Treating functional gastrointestinal disorders with antidepressants is effective even in the absence of depression
- Although several psychological treatments show promise in treating functional bowel disorders, no trial has yet provided unequivocal evidence of effectiveness

Jackson J, O'Malley PG, Tomkins G, Balden E, Santoro J, Kroenke K. Treatment of functional gastrointestinal disorders with antidepressant medications: a meta-analysis. *Am J Med* 2000;108:65-72

Talley NJ, Owen BKO, Boyce P, Paterson K. Psychological treatments for irritable bowel syndrome: a critique of controlled treatment trials. *Am J Gastroenterol* 1996;91:277-86

13 Chest pain

Christopher Bass, Richard Mayou

Chest pain is one of the commonest reasons for consultation in primary care. Chest pain is usually mild and transient, but further management is required in some cases. These are of two main types—acute severe pain and persistent pain associated with distress and functional limitation. Acute central chest pain accounts for 20-30% of emergency medical admissions. Chronic chest pain is the commonest reason for referral to cardiac outpatient clinics.

Management of chest pain

The improved diagnosis and early treatment of ischaemic heart disease have not been accompanied by similar advances either in the delivery of long term rehabilitation of patients with ischaemic heart disease or in the management of non-cardiac causes of chest pain. Since at least half of those referred to cardiac outpatient clinics and about two thirds of emergency admissions have a non-cardiac cause for their chest pain, there is a pressing need to address this problem.

Primary care

Primary care doctors have a major responsibility for the continuing care of patients with angina and those with chronic non-cardiac chest pain, as well as a role in secondary prevention. They therefore need good communication with specialist cardiac services and access to appropriate resources, including psychological treatments.

Patients with a low risk of coronary disease (such as young women with no cardiac risk factors and atypical pain) do not usually need cardiac investigation. Some, however, especially those with chest pain who have a family history of heart disease or other risk factors, may need investigation. In such cases it is important that the possibility of a non-cardiac cause of the chest pain is explained before referral. If investigation reveals no cardiac cause for the pain patients need their worries to be fully discussed, need advice about coping with symptoms, and should be encouraged to maintain activity.

Patients with an intermediate or high risk (such as middle aged male smokers) often require investigations even if the chest pain is "not typical" of ischaemic pain. This will usually be achieved by referral to a cardiology outpatient clinic or to an emergency assessment service. When referring patients in whom the cause of chest pain is uncertain it is important to avoid giving them the impression that the diagnosis of ischaemic heart disease is already established (such as by prescribing anti-anginal drugs). This is because, if patients come to believe that they have ischaemic heart disease, such beliefs can be difficult to change even if they are subsequently disproved by investigation.

Secondary care

The best way to organise emergency care remains uncertain. A long wait for specialist investigations such as angiography is likely to increase anxiety and disability, as has been shown in patients waiting for coronary artery surgery. Quicker access to assessment (such as by rapid access clinics and observation units) can be helpful but needs to be accompanied by a greater emphasis on aftercare for all patients assessed, not only those who have had infarction or are undergoing cardiac surgery.

British soldier admitted for observation with the diagnosis of "disordered action of the heart"—a post-combat syndrome in the first world war characterised by rapid heartbeat, shortness of breath, fatigue, and dizziness. (From Lewis T. The tolerance of physical exertion, as shown by soldiers suffering from so-called 'irritable heart.' *BMJ* 1918;i:363-5)

Assessment and management of chest pain in primary care

- History of pain, other symptoms and risk factors
- If at high risk of heart disease, refer for specialist assessment
- If at low risk:
  Identify non-cardiac causes
  Give a positive explanation
  Advise how to cope with symptoms and return to normal activity
  Discuss worries
  Offer review if symptoms are persistent

Clinical priorities in managing patients with chest pain

Primary care
- Recognise and refer possible heart disease
- Reassure minor chest pain
- Basic treatment of persistent non-cardiac pain
- Reassess chronic pain as required, monitor and coordinate continuing care
- Advise on secondary prevention need

Hospital emergency care
- Immediate diagnosis and treatment plus initiating continuing care of angina
- Make a positive diagnosis; reassure if non-cardiac and arrange follow up to determine investigation and treatment needs
- Full and rapid communication with primary care

Cardiac outpatient care
- Initiate immediate and continuing care of angina
- Reassure and advise if non-cardiac; plan treatment or review

Other specialist care
- Cardiac rehabilitation or aftercare
- Psychological or psychiatric referral

Types of chest pain

Angina
The English national service framework for coronary heart disease recognises that patients' beliefs, attitudes, emotions, and behaviour are powerful determinants of clinical outcomes and suggests not only routine psychosocial assessment but also the integration of psychological approaches into cardiac rehabilitation programmes. Self help behavioural treatment programmes have also been shown to be of benefit. The general principles of treatment described below for non-cardiac chest pain are also applicable to angina.

Myocardial infarction and depression
About one in six patients who have a myocardial infarction develop major depression. The occurrence of depression has been found to be independently associated with poor outcome, including poor quality of life, increased heart disease, and probably increased mortality. There is some evidence that those who have the severest heart disease are at greatest risk of an adverse outcome attributable to depression. It is in just these patients that depression is most likely to be missed because both doctor and patient understandably focus their attention on the heart disease and its treatment, rather than on psychological factors.

Myocardial infarction, angina, and non-cardiac chest pain
Patients who have had a myocardial infarction or who have proved angina often report other chest pains that are clearly non-cardiac. Inevitably, they tend to misinterpret these symptoms as evidence of heart disease. The consequence is often greater disability and distress and a high and inappropriate use of medical care.

Non-cardiac chest pain
Fewer than half of the patients referred to emergency departments and cardiac outpatient clinics have heart disease. Over two thirds of these continue to be disabled by symptoms in the long term, and many also remain dissatisfied with their medical care. Some continue to take cardiac drugs and to attend emergency departments, primary care, and outpatient clinics. Hence, although these patients have a good outcome in terms of mortality, they suffer considerable morbidity.

It is desirable to make an early and confident diagnosis of non-cardiac chest pain because appropriate management of this condition in primary care can reduce subsequent morbidity.

Causes of non-cardiac chest pain
Explanations in terms of a single cause are rarely helpful. Instead, the cause is often best understood as an interaction of biological, psychological, and social factors. In many cases there is an interaction between normal or abnormal physiological processes (such as extrasystoles, oesophageal spasm or reflux, and costochondral discomfort), psychological factors (such as how somatic sensations are perceived, interpreted, and acted on), and the behaviour and reactions of other people, including doctors.

Establishing a positive diagnosis of non-cardiac chest pain

The key to establishing a positive diagnosis of non-cardiac chest pain, both in primary care and cardiac clinics, is, first, to consider the pattern of chest pain symptoms and, second, to seek evidence for non-cardiac causes.

Main components of cardiac rehabilitation treatment programme for patients with myocardial infarctions

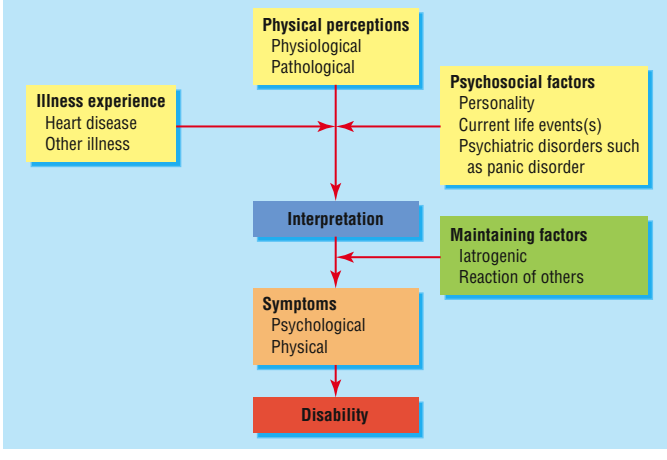
- Provide education about heart attacks and secondary prevention and correct misconceptions
- Agree and record goals for exercise, return to work, and everyday activities; provide copies for patients, medical notes, and primary care
- Offer home exercise programme or community group exercise, or both
- Routine early review of symptoms, activity, and progress with rehabilitation and secondary prevention goals
- Menu of specific interventions, including stopping smoking, diet, and identification and treatment of psychological and behavioural difficulties

Non-cardiac pain in patients with diagnosis of angina

Diagnostic uncertainty may result in
- Non-cardiac pain being wrongly attributed to angina
- Increased antianginal medication
- Increased iatrogenic distress and disability
- Unnecessary investigations
- Unnecessary admissions and consultations

Common causes of non-cardiac chest pain

- Oesophageal disorders—Gastro-oesophageal reflux, oesophageal dysmotility
- Musculoskeletal—Costochondritis, increased muscular tension
- Referred pain from thoracic spine
- Hyperventilation
- Psychological—Panic attacks, depression



Interaction of biological, psychological, and social factors to cause non-cardiac chest pain and subsequent disability

Iatrogenic factors maintaining symptoms and disabilities

- Giving probable diagnosis of angina before investigation
- Immediate prescription of antianginal drugs without explanation of possible causes before investigation
- Lack of explanation for distressing and continuing symptoms
- Inconsistent or ambiguous information
- Reassurance contradicted by continued antianginal drugs or other indications of uncertainty
- Lack of communication with all involved in care leading to contradictory and conflicting advice

Quality of chest pain
Attempts to identify certain characteristics of chest pain that can help to establish a positive diagnosis of non-cardiac chest pain have been encouraging. For example, as few as three questions can differentiate patients with chest pain but normal coronary arteries from those with coronary heart disease.

Evidence for common non-cardiac causes
Oesophageal disorders are often associated with chest pain, but chest pain is poor correlated with objective oesophageal abnormalities. Symptomatic treatment (such as proton pump inhibitors) can be useful. Psychological issues may need addressing whether or not there is oesophageal pathology. Gastro-oesophageal reflux is an important cause of atypical chest pain, but there is no convincing evidence that such chest pain is often related to disturbances of oesophageal motility.

Emotional disorders—Only a minority of patients who present to family doctors with non-cardiac chest pain are suffering from conspicuous anxiety or depressive disorders. The rate of such disorders is, however, higher among those referred for specialist assessment in cardiac clinics, especially those who undergo angiography and are shown to have normal coronary arteries. It is important to seek evidence of (*a*) the key symptoms of depression (which include hopelessness; lack of interest, pleasure, and concentration; poor sleep; and irritability as well as low mood) and (*b*) an association of the chest pain with anxiety and panic attacks.

Patients' beliefs and worries
Even if no definite psychiatric diagnosis can be made, it is essential to ask patients what goes through their mind when they experience chest pain.

Stressful life events
Distressing life events can precipitate not only anxiety and depressive disorders, but also functional symptoms such as chest pain. Events signifying loss, threat, and rejection are of particular importance. Open questions are most effective in eliciting these—such as: "Tell me about any changes or setbacks that occurred in the months before your chest pain began."

Treatment of non-cardiac pain

Early and effective intervention is crucial, but how can this best be provided? Because patients vary not only in the frequency and severity of symptoms and associated disability but also in their needs for explanation and treatment of their physical and psychological problems, management needs to be flexible.

Avoiding iatrogenic worries—A consultation for chest pain is inherently worrying. Inevitably, many patients assume that they have severe heart disease, which will have major adverse effects on their life. These concerns may be greatly increased by delays in investigation, by comments or behaviours by doctors, and by contradictory and inconsistent comments.

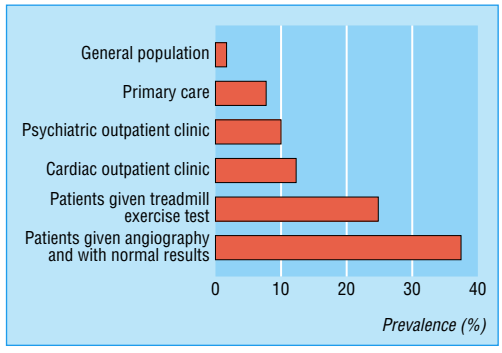
Symptomatic treatment—In some patients the pain is obviously musculoskeletal in origin and can be treated with non-steroidal anti-inflammatory drugs. Proton pump inhibitors provide effective relief from the symptoms typical of gastro-oesophageal reflux, even in those with an essentially normal oesophageal mucosa. In some cases oesophageal function testing may reveal a motility disorder or acid reflux unresponsive to first line drugs. These patients may require specialist gastroenterological referral.

Communication—Problems in the care of patients with chest pain often arise from failures in communication between primary and secondary care. Lack of information and contradictory or inconsistent advice makes it less likely that patients and their

Questions to differentiate patients with non-cardiac chest pain from those with coronary heart disease

Question	Response: Typical	Response: Atypical
If you go up a hill (or other stressor) on 10 separate occasions on how many do you get the pain?	10/10	<10/10
Of 10 pains in a row, how many occur at rest?	<2/10	≥2/10
How many minutes does the pain usually last?	<5	≥5

When answers to all three questions are "atypical" the chance of coronary disease is only 2% in patients aged <55 years and 12% in those aged ≥55



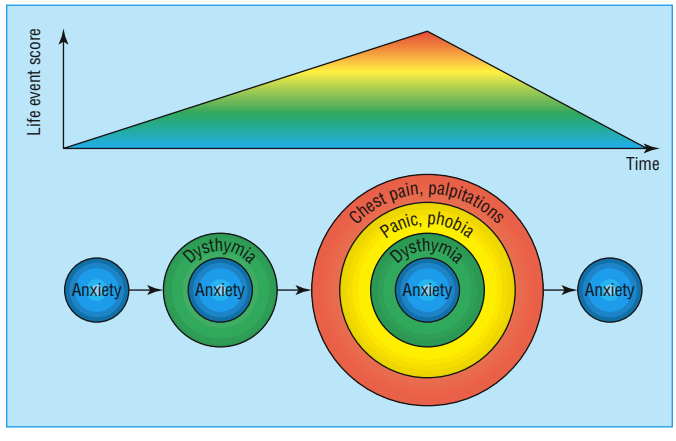
Prevalence of panic disorder in different medical settings

Screening questions for panic attacks

- In the past six months have you ever had a spell or an attack when you suddenly felt frightened, anxious, or very uneasy?
- In the past six months have you ever had a spell or an attack when for no reason your heart suddenly began to race, you felt faint, or you couldn't catch your breath?

If the answer is yes to either question then continue

- Obtain description
- Did any of these spells happen when you were not in danger or the centre of attention, such as in a crowd or when travelling?
- How many times have you had these spells in the past month?



Life events and symptom reporting. Stress of adverse life events may result in increases in reporting of psychological and physical symptoms

Management of non-cardiac chest pain

General management

- Explanation of the diagnosis
- Reassurance that it is a real, common, and well recognised problem
- Advice on specific treatments
- Advice on behaviour—such as not avoiding exercise
- Discussion of concerns
- Provision of written information
- Involvement of relatives
- Follow up to review

Specialist treatments

- Cognitive behaviour therapy
- Antidepressant drugs
- Psychosocial intervention for associated psychological, family, and social difficulties

families will gain a clear understanding of the diagnosis and of treatment plans. The increasing use of computerised exchange of key information may reduce this problem, although it remains important to ensure that the information is passed on to and understood by patients and relatives.

Effective reassurance—Those with mild or brief symptoms may improve after negative investigation and simple reassurance. Further hospital attendance may then be unnecessary. Others with more severe symptoms and illness concerns will benefit from a follow up visit four to six weeks after the cardiac clinic visit (or emergency room visit), which allows time for more discussion and explanation. This may be with either a cardiac nurse in the cardiac clinic or a doctor in primary care. It also provides a valuable opportunity to identify patients with recurrent or persistent symptoms who may require further help.

Specialist treatments—Psychological and psychopharmacological treatment should be considered for patients with continuing symptoms and disability, especially if these are associated with abnormal health beliefs, depressed mood, panic attacks, or other symptoms such as fatigue or palpitations. Both cognitive behaviour therapy and selective serotonin reuptake inhibitors have been shown to be effective. Tricyclic antidepressants are helpful in reducing reports of pain in patients with chest pain and normal coronary arteries, especially if there are accompanying depressive symptoms.

Organising care

Because of the heterogeneity of the needs of patients who present with chest pain, we propose a "stepped" approach to management. A cardiologist working in a busy outpatient clinic may require access to additional resources if he or she is to provide adequate management for large numbers of patients with angina or non-cardiac chest pain. One way of doing this is to employ a specialist cardiac nurse who has received additional training in the management of these problems. The nurse can provide patient education, simple psychological intervention, and routine follow up in a separate part of the cardiac outpatient clinic. For those patients who require more specialist psychological care, it is important for the cardiac department (possibly the cardiac nurse) to collaborate with the local psychology or liaison psychiatry service.

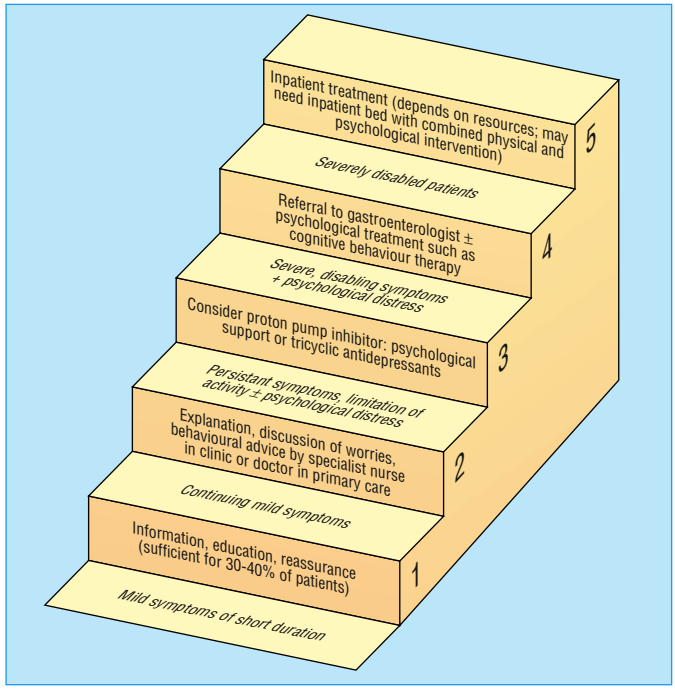
Conclusion

The management of coronary heart disease has received much attention in recent years, whereas non-cardiac chest pain has been relatively neglected. The structuring of cardiac care for both angina and non-cardiac chest pain to incorporate a greater focus on psychological aspects of medical management would be likely to produce considerable health gains.

The picture of a soldier with "disordered action of the heart" is reproduced with permission of Wellcome Trust. The box of questions to identify patients with non-cardiac chest pain is adapted from Cooke R et al, *Heart* 1997;78:142-6. The figure showing link between life events and range of psychological and physical complications is adapted from Tyrer P, *Lancet* 1985;i:685-8. The figure of stepped care for managing non-cardiac chest pain is adapted from Chambers J et al, *Heart* 2000;84:101-5.

Effective reassurance

- Accept reality of symptoms
- Give explanation of causes
- Explain that symptoms are common, well recognised, and have a good prognosis
- Understand patient's and family's beliefs and worries
- Plan and agree simple self help
- Provide written information and plans
- Offer to see patient's partner or other close relative
- Offer follow up if required



"Stepped" care in the management of non-cardiac chest pain

Evidence based summary

- Half of patients referred from general practice to a cardiac clinic with chest pain or palpitations do not have cardiac disease, but, despite the absence of disease, their symptoms tend to persist
- Psychological treatment and antidepressant drugs can be effective in treating non-cardiac chest pain

Mayou R, Bryant B, Forfar C, Clark D. Non-cardiac chest pain and benign palpitations in the cardiac clinic. *Br Heart J* 1994;72:548-53

Mayou R, Bryant B, Sanders D, Bass C, Klimes I, Forfar C. A controlled trial of cognitive behavioural therapy for non-cardiac chest pain. *Psychol Med* 1997; 27:21-31

Cannon RO 3rd, Quyyumi AA, Mincemoyer R, Stine AM, Gracely RH, Smith WB, et al. Imipramine in patients with chest pain despite normal coronary angiograms. *N Engl J Med* 1994;330:1411-7

Suggested reading

- Mayou RA, Bass C, Hart G, Tyndel S, Bryant B. Can clinical assessment of chest pain be made more therapeutic? *Q J Med* 2000;93:805-11
- Cooke R, Smeeton M, Chambers JB. Comparative study of chest pain characteristics in patients with normal and abnormal coronary angiograms. *Heart* 1997;78:142-6
- Creed F. The importance of depression following myocardial infarction. *Heart* 1999;82:406-8
- Jain D, Fluck D, Sayer JW, Ray S, Paul EA, Timmis AD. One-stop chest pain clinic can identify high cardiac risk. *J R Coll Physicians Lond* 1997;31:401-4
- Thompson DR, Lewin RJ. Management of the post-myocardial infarction patient: rehabilitation and cardiac neurosis. *Heart* 2000;84:101-5

14 Delirium

Tom Brown, Michael Boyle

Delirium is a common cause of disturbed behaviour in medically ill people and is often undetected and poorly managed. It is a condition at the interface of medicine and psychiatry that is all too often owned by neither. Although various terms have been used to describe it—including acute confusional state, acute brain syndrome, and acute organic reaction—delirium is the term used in the current psychiatric diagnostic classifications and the one we will use here.

Clinical features

Delirium usually develops over hours to days. Typically, the symptoms fluctuate and are worse at night. The fluctuation can be a diagnostic trap, with nurses or relatives reporting that patients had disturbed behaviour at night whereas doctors find patients lucid the next day.

Impaired cognitive functioning is central and affects memory, orientation, attention, and planning skills. Impaired consciousness, with a marked variability in alertness and in awareness of the environment is invariably present. A mistaken idea of the time of day, date, place, and identity of other people (disorientation) is common. Poor attention, and disturbed thought processes may be reflected in incoherent speech. This can make assessment difficult and highlights the need to obtain a history from a third party. Relatives or other informants may report a rapid and drastic decline from premorbid functioning that is useful in distinguishing delirium from dementia.

Disturbed perception is common and includes illusions (misperceptions) and hallucinations (false perceptions). Visual hallucinations are characteristic and strongly suggest delirium. However, hallucinations in auditory and other sensory modalities can also occur.

Delusions are typically fleeting, often persecutory and usually related to the disorientation. For example, an elderly person may believe that the year is 1944, that he or she is in a prisoner of war camp, and that the medical staff are the enemy. Such delusions can be the basis of aggressive behaviour,

Delirium can have a profound effect on affect and mood. A patient's affect can range from apathy and lack of interest to anxiety, perplexity, and fearfulness that may sometimes amount to terror. A casual assessment can result in an erroneous diagnosis of depression or anxiety disorder.

Disturbances of the sleep-wake cycle and activity are common. A behaviourally disturbed patient with night time agitation wandering around the ward is usually easy to recognise. However, presentations where a patient is hypo-alert and lethargic may go unrecognised.

Detection of delirium

Delirium often goes undiagnosed. Non-detection rates as high as 66% have been reported. Detection and diagnosis are important because of the associated morbidity and mortality: although most patients with delirium recover, some progress to stupor, coma, seizures, or death. Patients may die because of failure to treat the associated medical condition or from the associated behaviour—inactivity may cause pneumonia and decubital ulcers, and wandering may lead to fractures from falls.

Sensory misperceptions, including hallucinations and illusions, are common in delirium. (*Don Quixote and the Windmill* by Gustave Doré, 1832-1883)

Diagnostic criteria for delirium*

- Disturbance in consciousness with reduced ability to focus, sustain, or shift attention
- Change in cognition (such as memory, disorientation, speech, disturbance) or development of perceptual disturbance not better accounted for by pre-existing or evolving dementia
- Disturbance develops over hours to days and fluctuates in severity

*Adapted from *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (DSM-IV)

Alcohol addiction often goes undetected at the time of admission to hospital. All admitted patients should be asked about their alcohol consumption

Differential diagnosis

The main differential diagnosis of delirium is from a functional psychosis (such as schizophrenia and manic depression) and from dementia. Functional psychoses are not associated with obvious cognitive impairment, and visual hallucinations are more common in delirium. Dementia lacks the acute onset and markedly fluctuating course of delirium. Fleeting hallucinations and delusions are less common in dementia. It is important to note that delirium is commonly superimposed on a pre-existing dementia.

Prevalence

Most prevalence studies of delirium have been carried out in hospitalised medically ill patients, in whom the prevalence is about 25%. Most at risk are elderly patients, postoperative patients, and those who are terminally ill. The epidemiology of delirium in primary care and the community is unknown, but, with shorter length of stay in hospital and more surgery on a day case basis, it is likely to be increasingly common in the community and in residential care homes. It has been estimated that, among hospital inpatients with delirium, less than half have fully recovered by the time of discharge.

Aetiology

Delirium has a large number of possible causes. Many of these are life threatening, and delirium should therefore be regarded as a potential medical emergency. It is increasingly recognised that most patients have multiple causes for delirium, and consequently there may be several factors to be considered in diagnosis and management. Causes of delirium may be classified as

- Underlying general medical conditions and their treatment
- Substance use or withdrawal
- Of multiple aetiology
- Of unknown aetiology.

Prescribed drugs and acute infections are perhaps the commonest causes, particularly in elderly people. Prescribed drugs are implicated in up to 40% of cases and should always be considered as a cause. Many prescribed drugs can cause delirium, particularly those with anticholinergic properties, sedating drugs like benzodiazepines, and narcotic analgesics.

Withdrawal from alcohol or from sedative hypnotic drugs is a common cause of delirium in hospitalised patients separated from their usual supply of these substances. Delirium tremens is a form of delirium associated with alcohol withdrawal and requires special attention.

In addition to looking for precipitating causes of delirium, it is important to consider risk factors. These include age (with children and elderly people at particular risk), comorbid physical illness or dementia, and environmental factors such as visual or hearing impairment, social isolation, sensory deprivation, and being moved to a new environment.

Management

There are four main aspects to managing delirium:

- Identifying and treating the underlying causes
- Providing environmental and supportive measures
- Prescribing drugs aimed at managing symptoms
- Regular clinical review and follow up.

Good management of delirium goes beyond mere control of the most florid and obvious symptoms.

Distinguishing delirium from dementia

	Delirium	Dementia
Onset	Acute or subacute	Insidious
Course	Fluctuating, usually revolves over days to weeks	Progressive
Conscious level	Often impaired, can fluctuate rapidly	Clear until later stages
Cognitive defects	Poor short term memory, poor attention span	Poor short term memory, attention less affected until severe
Hallucinations	Common, especially visual	Often absent
Delusions	Fleeting, non-systematised	Often absent
Psychomotor activity	Increased, reduced, or unpredictable	Can be normal

Prevalence of delirium

Setting	% with delirium
Hospitalised medically ill patients*	10-30%
Hospitalised elderly patients	10-40%
Hospitalised cancer patients	25%
Hospitalised AIDS patients	30-40%
Terminally ill patients	80%

*High risk conditions and procedures include cardiotomy, hip surgery, transplant surgery, burns, renal dialysis, and lesions of the central nervous system

Causes of delirium due to underlying medical conditions

- Intoxication with drugs—Many drugs implicated especially anticholinergic agents, anticonvulsants, anti-parkinsonism agents, steroids, cimetidine, opiates, sedative hypnotics. Don't forget alcohol and illicit drugs
- Withdrawal syndromes—Alcohol, sedative hypnotics, barbiturates
- Metabolic causes
  Hypoxia, hypoglycaemia, hepatic, renal or pulmonary insufficiency
  Endocrinopathies (such as hypothyroidism, hyperthyroidism, hypopituitarism, hypoparathyroidism or hyperparathyroidism)
  Disorders of fluid and electrolyte balance
  Rare causes (such as porphyria, carcinoid syndrome)
- Infections
- Head trauma
- Epilepsy—Ictal, interictal, or postictal
- Neoplastic disease
- Vascular disorders
  Cerebrovascular (such as transient ischaemic attacks, thrombosis, embolism, migraine)
  Cardiovascular (such as myocardial infarction, cardiac failure)

Features of delirium tremens

- Associated with alcohol withdrawal
- Delirium with prominent anxiety and autonomic hyperactivity
- There may be associated metabolic disturbance and fits
- Chronic alcoholics are at risk of Wernicke's encephalopathy, in which delirium becomes complicated by ataxia and ophthalmoplegia. Urgent treatment with parenteral thiamine is required to prevent permanent memory damage

Making the diagnosis

Most patients with delirium are identified only because of marked behavioural disturbance. It would be preferable for all older patients to be screened for risk factors at admission to hospital. These would include substance misuse (particularly alcohol) and pre-existing cognitive impairment (assessed with the Hodkinson mental test or similar). Although such screening questions are part of the admission form in many hospitals, in our experience junior doctors seldom complete them. Once patients are admitted, minor episodes of confusion, behavioural disturbance, or increasing agitation should be taken seriously and investigated as appropriate. They should not be simply dismissed as "old age" or psychological reactions to hospitalisation.

Identifying and treating the cause

Delirium, by definition, is secondary to one or more underlying cause. Identifying such causes is often difficult, especially when patients are unable to give a coherent history or cooperate with physical examination. On occasions, it can be necessary to sedate a patient before conducting an adequate assessment. The interviewing of third parties is often helpful. Once a cause is found, appropriate treatment should be started without delay.

The environment

The aims of environmental interventions are, firstly, to create an environment that places minimum demands on a patient's impaired cognitive function and, secondly, to limit the risk of harm to the patient and others that may result from disturbed behaviour. Nursing should, as far as possible, be done by the same member of staff (preferably one trusted by the patient). This consistency should be supported with other strategies such as clear and if necessary repeated communication, adequate lighting, and the provision of clocks as aids to maintaining orientation. Visits from family and friends and provision of familiar objects from home can also be helpful. The correction of sensory impairments (such as by providing glasses or hearing aids) to help patients' grip on reality is sometimes overlooked.

It is also be important to minimise any risk to a delirious patient, other patients on the ward, and staff by ensuring that the patient is in a safe and separate area and that potentially dangerous objects are removed.

Drug treatment

Drug treatment of delirium should only be used when essential and then with care. This is because drugs such as antipsychotics and benzodiazepines can make the delirium worse and can exacerbate underlying causes (for example, benzodiazepines may worsen respiratory failure).

Antipsychotic drugs

Antipsychotics are the most commonly used drugs. Their onset of action is usually rapid, with improvement seen in hours to days. Haloperidol is often used because it has few anticholinergic side effects, minimal cardiovascular side effects, and no active metabolites. As it is a high potency drug it is less sedating than phenothiazines and therefore less likely to exacerbate delirium. It is, however, prone to causing parkinsonism, which may exacerbate a patient's tendency to fall. Low dose haloperidol (1-10 mg/day) is adequate for most patients. In severe behavioural disturbance haloperidol may be given intramuscularly or intravenously.

It is preferable to use a fixed dose that is frequently reviewed from the time of diagnosis rather than always giving the drug "as required" in response to disturbed behaviour. It is essential,

Hodkinson mental test

Score one point for each question answered correctly and give total score out of 10

Question

- Patient's age
- Time (to nearest hour)
- Address given, for recall at end of test (42 West Street)
- Name of hospital (or area of town if at home)
- Current year
- Patient's date of birth
- Current month
- Years of the first world war
- Name of monarch (or president)
- Count backwards from 20 to 1 (no errors allowed but may correct self)

Environmental and supportive measures in delirium

- Education of all who interact with patient (doctors, nurses, ancillary and portering staff, friends, family)
- Reality orientation techniques
  Firm clear communication—preferably by same member of staff
  Use of clocks and calendars
- Creating an environment that optimises stimulation (adequate lighting, reducing unnecessary noise, mobilising patient whenever possible)
- Correcting sensory impairments (providing hearing aids, glasses, etc)
- Ensuring adequate warmth and nutrition
- Making environment safe (removing objects with which patient could harm self or others)

Simple measures to help orientation (such as glasses, hearing aids, and clocks) are effective in the management of delirium

In postoperative patients judicious use of oxygen can treat delirium effectively

yet often forgotten, to monitor patients for both adequate response and unacceptable side effects. While a patient is in hospital this consists of at least a daily assessment of symptoms, level of sedation, and examination for extrapyramidal and other unwanted drug effects.

Preliminary experience with new antipsychotics suggest they may also be effective in delirium, but their advantages remain unestablished.

Benzodiazepines

Benzodiazepines are usually preferred when delirium is associated with withdrawal from alcohol or sedatives. They may also be used as an alternative or adjuvant to antipsychotics when these are ineffective or cause unacceptable side effects. Intravenous or intramuscular lorazepam may be given up to once every four hours. In patients with delirium due to hepatic insufficiency, lorazepam is preferred to haloperidol. Excessive sedation or respiratory depression from benzodiazepines is reversible with flumazenil.

Review

One of the most consistent failings in the management of delirium is lack of review. The acute symptoms are usually dealt with "out of hours" by junior staff and are forgotten by the next day. It is essential to review management of delirium and of the underlying causes for the duration of the hospital stay.

Patients' capacity and consent

Increasingly issues of capacity and informed consent may be raised in relation to the treatment of delirium. Urgent interventions needed to prevent serious deterioration or death or necessary in the interests of a patient's safety are deemed to be covered by common law in the United Kingdom. Although opinions differ, most agree that (*a*) if medical colleagues would deem a treatment appropriate and (*b*) if reasonable people would want the treatment themselves, then it can be given if urgently necessary.

Explaining the diagnosis

Effective management requires that not only the doctors and nurses caring for a patient understand the condition, but that the patient's family and friends appreciate the reasons for the dramatic change in the person's behaviour and that it is usually a reversible condition.

Aftercare

Many patients with delirium still have residual symptoms at the time of discharge from hospital. There is therefore a need for continued vigilance about medication, environmental change, and sensory problems during discharge planning and aftercare. Close liaison between hospital and primary care is an essential part of discharge planning.

Patients or their families will often need reassurance that an episode of delirium is not the start of an inevitable progression to dementia and that a full recovery can usually be expected. Delirious patients may erroneously be placed in long term care as "demented": decisions to place patients in care should be made only after an adequate assessment that differentiates delirium from dementia.

Excessive use of sedative drugs often causes more problems that it solves

Key medicolegal judgments about patients' capacity and consent (English Law)

Re c (mental patients: refusal of treatment) [1994] 1 WLR 290
An adult has the capacity to consent to or refuse treatment if he or she can

- Understand and retain the information relevant to the decision in question
- Believe that information
- Weigh the information in the balance to arrive at an informed choice

Re f (mental health sterilisation) v West Berkshire Health Authority (1989) 2 WLR 1025: (1989) All ER 673
"not only (1) must there be a necessity to act when it is not practicable to communicate with the assisted person but also (2) the action taken must be such as a reasonable person would in all circumstances take, acting in the best interests of the assisted person."
"Action properly taken to preserve life, health or wellbeing of the assisted person (which) may well transcend such measures as surgical operations or substantial treatment and may extend to include such humdrum matters as routine medical or dental treatment, even such simple care as dressing and undressing and putting to bed."

Evidence based summary

- A quarter of hospitalised elderly patients will have delirium
- Occurrence of delirium predicts poorer outcome and greater length of stay even after controlling for other variables, including severity of illness
- Positive identification and management of risk factors can reduce incidence and severity of delirium in elderly patients

Francis J, Martin D, Kapoor WN. A prospective study of delirium in hospitalized elderly. *JAMA* 1990;263:1097-101

O'Keeffe S,.Lavan J. The prognostic significance of delirium in older hospital patients. *J Am Geriatr Soc* 1997;45:174-8

Cole MG, Primeau FJ, Elie LM. Delirium: prevention, treatment, and outcome studies. *J Geriatr Psychiatry Neurol* 1998;11:126-37

Further reading

- American Psychiatric Association. *Practice guideline for the treatment of patients with delirium.* Washington, DC: APA, 1999
- Meagher DS. Delirium—optimising management. *BMJ* 2001;322:144-9
- Meagher DS, O'Hanlon D, O'Mahony E, Casey PR. The use of environmental strategies and psychotropic medication in the management of delirium. *Br J Psychiatry* 1996;168:512-5
- Taylor D, Lewis S. Delirium. *J Neurol Neurosurg Psychiatry* 1993;56:742-51

Index

Page numbers in **bold** refer to figures in the text; those in *italics* refer to tables or boxed material

ABC OF THE FIRST YEAR

Fifth edition

BMJ Books

ABC OF THE FIRST YEAR

Fifth edition

H B VALMAN
Consultant paediatrician,
Northwick Park Hospital,
Harrow, Middlesex
Honorary Senior Lecturer
Imperial College, London

R M THOMAS
Consultant paediatrician,
Northwick Park Hospital,
Harrow, Middlesex
Honorary Senior Lecturer,
Imperial College, London

BMJ Books

BMJ Books is an imprint of the BMJ Publishing Group

First edition 1980
Second impression 1982
Second edition 1984
Second impression 1987
Third impression 1988
Third edition 1989
Fourth edition 1995
Second impression 2000
Fifth edition 2002
by BMJ Books, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0-7279-1681-5

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed and bound in Spain by Graphy Cems, Navarra

Contents

Preface to the fifth edition

The First Year of Life was the first series of ABC articles commissioned and published weekly by the *BMJ* and later collected and published as a book. This new fifth edition is called *ABC of the First Year* to identify it with the ABC series. The ABC series is continuing to tackle new subjects in all branches of medicine and surgery: 46 series have been published and 43 books are still in print. The first edition of *The First Year of Life* was printed in black and white and some colour was introduced in the last edition. The majority of the pages in this new edition are in colour and nearly 100 new photographs have been added. Each chapter has been thoroughly revised with considerable changes in the chapters on prenatal assessment, infants of low birthweight, breathing difficulties in the newborn, and diarrhoea. The charts for developmental review have incorporated the latest recommendations of the Department of Health. New sections have been added on the infant of low birthweight at home, advice on travelling abroad with an infant, and paediatric HIV infection.

The book was written for family doctors, GP vocational trainees, medical students, midwives, and nurses. It has become the standard textbook for several undergraduate and postgraduate courses. The emphasis has been on the practical aspects of management, based on clinical experience, but theory is introduced where it is essential for understanding the basis of management. No previous experience of paediatrics is assumed.

I welcome Dr Roslyn Thomas who has joined me as co-author of this edition. I wish to thank the staff of the BMJ, and especially Alex Stibbe who has acted as midwife for the fifth edition of the book, and my wife who has constantly supported me and encouraged me to write.

For ease of reading and simplicity a single pronoun has been used for both feminine and masculine subjects; a specific gender is not implied.

Bernard Valman

Foreword to the first edition

The care of infants and their mothers has changed rapidly in the past 10 years and it is often difficult to identify those advances that will prove of lasting value to the clinician.

Dr Bernard Valman's articles on the first year of life, published recently in the *BMJ* and collected in book form, aim at providing the clinician in the community and in hospital with generally accepted views on the medical management of infants.

The main difference between paediatrics and general medicine is the range of normality, which changes with age. The greatest changes occur in the first year of life. Dr Valman's articles provide an account of normal development during this year, with particular emphasis on its assessment, so that deviations may be easily recognised. These articles have been collected together to provide a practical guide for general practitioners and the many other staff who care for the new born and young infants.

Stephen Lock
Editor, *BMJ* 1980

1 Prenatal assessment

Recent advances in ultrasound technique, equipment, and training together with rapid advances in molecular biology have increased the range of antenatal diagnoses. Some methods are available only at specialised centres. This chapter will give a background to successful techniques. An anomaly may be detected during routine examination of the fetus which is carried out by ultrasound between 18 and 24 weeks of gestation. Maternal serum screening for Down's syndrome is performed as a double, triple or two stage integrated test (see below).

After the birth of an abnormal baby or the detection of genetic disease in an older child, a paediatrician or geneticist may recommend a specific test at a particular week in the subsequent pregnancy. Some tests are at an early stage in development and the false positive and negative rates have not been assessed. Some genetic tests are not yet sufficiently precise to enable an accurate prognosis to be given to every family with that disease.

At the first antenatal visit it is still important to carry out a full blood count and haemoglobin electrophoresis, blood grouping, rhesus antibody titre, tests for rubella and hepatitis B, human immunodeficiency virus (HIV), and syphilis. The haemoglobin electrophoresis may show that the mother has β-thalassaemia trait or sickle cell trait and the father's red cell investigations may suggest that further studies of the fetus are needed.

Box 1.1 Screening

- Ultrasound at 12 weeks
- Ultrasound at 18–24 weeks
- Maternal serum screening for Down's syndrome

Box 1.2 First antenatal visit

- Full blood count and haemoglobin electrophoresis
- Blood group
- Rhesus antibody titre
- Rubella
- Hepatitis B
- HIV
- Syphilis

Ultrasound studies

The first routine examination of the fetus by ultrasound is usually performed at the gestational age of 12 weeks. The gestational age is confirmed and anomalies of the central nervous system or cystic hygromas may be detected. A further scan at 18–24 weeks may detect anomalies of the central nervous system, heart, kidneys, intestinal tract, and skeleton. Signs which suggest a chromosome abnormality include choroid plexus cysts, echogenic cardiac foci, renal pelvic dilatation or echogenic bowel. They occur in approximately 1: 250 pregnancies and are associated with a 1: 300 risk of a chromosome abnormality. These isolated signs in the presence of normal serum screening probably do not merit the fetal risks of amniocentesis but full discussion is necessary and the patient may still opt for karyotyping to be performed. Mothers with a family history of congenital heart disease should be offered a detailed cardiac ultrasound scan at 18–24 weeks as the risk of the fetus having a heart problem is 3–5%. The consultant obstetrician, ideally with the paediatrician, should discuss the diagnosis and prognosis of an anomaly with both parents. Termination of the pregnancy may need to be considered or serial ultrasound examination performed during the pregnancy and in the neonatal period.

Ultrasound guidance is used in taking samples of the amniotic fluid (amniocentesis) and in selected centres it has been used to take blood samples from the umbilical cord (cordocentesis) and to give blood transfusion by that route.

The samples can be used in gene probe techniques, enzyme estimation, and chromosome studies. In rhesus incompatibility a low haematocrit in the cord blood indicates the need for fetal transfusion.

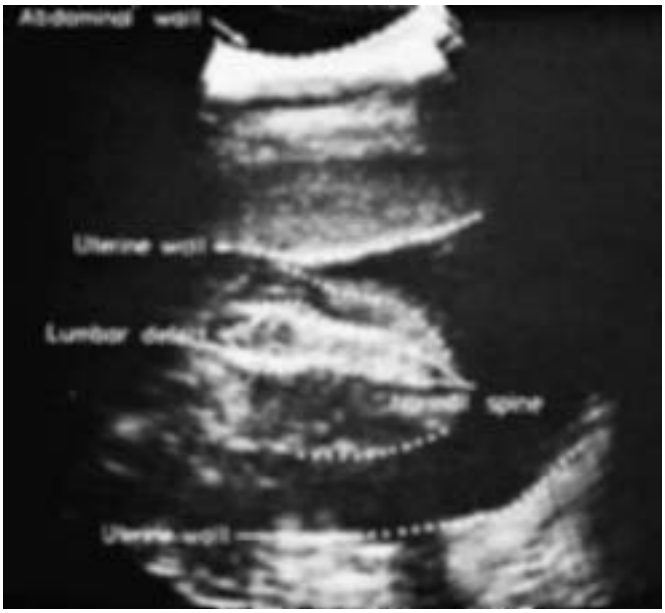


Figure 1.1 Ultrasound showing lumbar spine defect.

Amniocentesis

Amniotic fluid is removed by passing a needle into the amniotic cavity through the mother's abdominal wall and uterus under

ultrasound guidance. Amniocentesis yields amniotic fluid containing cells that have been shed from the skin of the fetus. Examination of the cultured cells reveals the chromosome constitution of the fetus, including sex. Specific enzymes can be sought and deoxyribonucleic acid (DNA) probes used. Women who are found to be at higher risk for Down's syndrome on serum screening are offered amniocentesis. In high risk women fluorescent *in situ* hybridisation (FISH test) may be offered which uses the polymerase chain reaction (PCR) to detect chromosome abnormalities such as the common trisomies, 21, 18, and 13 – Down's, Edward's, and Patau syndrome respectively – the results of which are available within a few working days.

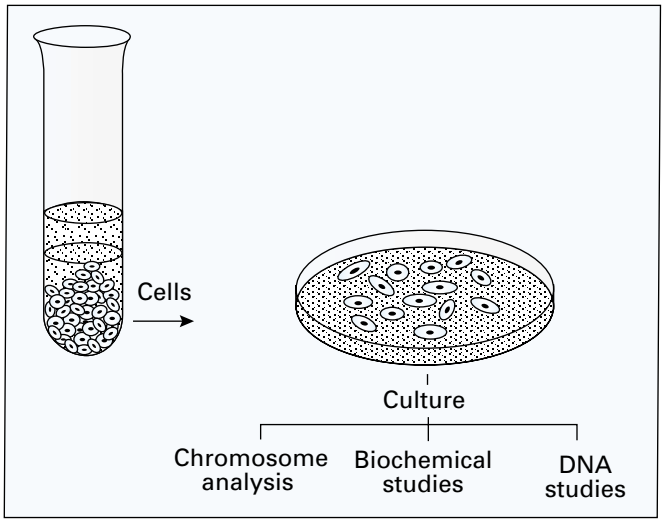


Figure 1.2 Examination of amniotic fluid.

Chorionic villus biopsy

Chorionic villus biopsy is carried out mainly by the transabdominal route under ultrasound guidance after 10 weeks gestation. The main indications have been maternal age, previous chromosome anomaly, fetal sexing, enzyme assay, and gene probe assessment. Gene probes have been developed for several diseases including cystic fibrosis, Duchenne muscular dystrophy, and the haemoglobinopathies. DNA is extracted from the chorionic villus sample and the probe is used to determine whether a specific part of a particular gene is present or absent.

There is a higher miscarriage rate with chorionic villus biopsy compared to amniocentesis. As there is a risk of limb reduction deformities and facial anomalies when it is performed early, it should be carried out after the 10th week of gestation.

Autosomal dominant
- Huntington's chorea
- Myotonic dystrophy
- Adult polycystic kidneys
- Tuberous sclerosis
- Von Recklinghausen's disease

X linked
- Duchenne muscular dystrophy
- Haemophilia A and B
- Fragile X

Autosomal recessive
- α and β-thalassaemia
- Sickle cell disease
- Cystic fibrosis
- Phenylketonuria
- α1-antitrypsin deficiency
- Congenital adrenal hyperplasia

Figure 1.3 Examples in which DNA gene probes are available.

Maternal serum screening for Down's syndrome

The majority of babies with Down's syndrome are born to mothers under the age of 37 years because they form the largest proportion of mothers. The serum screening for Down's syndrome should be offered to all mothers irrespective of maternal age as it provides an assessment of the risk but not a definite diagnosis of Down's syndrome. The double or triple test is offered at the initial antenatal visit to all mothers. Three biochemical parameters (serum α-fetoprotein, β-human chorionic gonadotrophin (HCG), and oestriol) are taken with an accurate gestational age assessed by ultrasound and maternal age to assess the risk of Down's syndrome. If all those mothers identified as being at risk (screen positive = risk greater than 1: 250) using the triple test had an amniocentesis, then it is thought that 60–65% of Down's syndrome babies would be detected. It is hoped that with the addition of ultrasound to detect signs such as an increase in nuchal thickness and other biochemical tests, it may be possible to improve the risk assessment even more, using a two-stage integrated test (see Box 1.3). It is hoped that in future, with improved techniques of DNA gene replication, it might be possible to karyotype a fetus from fetal cells in the maternal circulation.

Box 1.3 Two-stage integrated test

Stage one at 10–13 weeks
1. Ultrasound scan to determine gestational age and nuchal thickness
2. Blood level of plasma protein A (PAPP-A)

Stage two at 15–22 weeks
1. Blood levels of:
   - α-fetoprotein (**AFP**)
   - free β-human chorionic gonadotrophin (**β-hCG**)
   - unconjugated oestriol (**uE_3**)
   - inhibin-A (**inhibin**)
2. Integration of results from the two stages to estimate risk of Down's syndrome or a neural tube defect

Risks

The risk to a particular fetus depends on the gestational age of the fetus, the indication for the procedure, and the experience of the operator. The incidence of complications has fallen as skill in the newer techniques has increased. The abortion rates

are difficult to assess but Table 1.1 opposite has been compiled from expert advice on the available evidence. The risk of abortion after amniocentesis at 15 weeks is about 1%, which is about twice the spontaneous incidence in normal pregnancies. Fetal or maternal bleeding has been considerably reduced by the use of ultrasound, but a slight risk of infection remains and the incidence of respiratory distress syndrome and orthopaedic problems, such as talipes, is probably slightly increased in fetuses who have undergone early amniocentesis. Chorionic villus biopsy has a higher risk of abortion of about 5% against a background of spontaneous abortion of 3%. Chorionic villus biopsy carried out at about 10 weeks gestation provides a result early in pregnancy, when termination of the pregnancy is less traumatic and more acceptable for many mothers. Some tests are slightly more accurate when the sample is obtained by amniocentesis. Some investigations can be performed only on a specific sample.

Table 1.1 Risk of abortion

Procedure	Gestational age performed (weeks)	Spontaneous abortion (%)	Risk of abortion after procedure (%)
Amniocentesis	14–18	0.5	1
Chorionic villus biopsy	>10	2–3	3–5
Cordocentesis	18–20	<1	1–2

Screening for bacterial vaginosis

Preterm birth is the major cause of death and disability in babies. The aetiology of preterm labour is multifactorial but there is increasing evidence to implicate infection as a possible cause in up to 40% of cases. This information may not help once a woman is admitted in preterm labour, since by that time there may be irreversible changes in the cervix. Where the information may be useful is in the prediction and prevention of preterm labour. A few recent studies have reported that abnormal colonisation of the vagina in the form of bacterial vaginosis carries a risk of up to fivefold for the subsequent development of preterm labour and late miscarriage. Whether by reversing this condition it is possible to reduce the incidence of preterm labour and delivery is currently being tested.

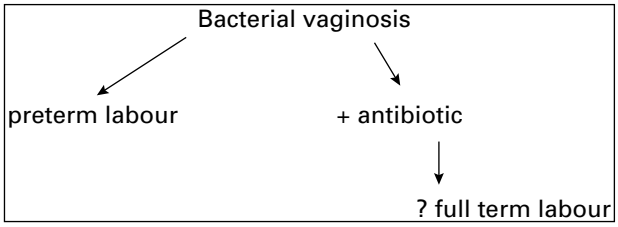


Figure 1.4 Bacterial vaginosis.

Follow-up of fetal renal tract anomalies

Mild dilatation (<10 mm) of the fetal renal pelvis is often found on the routine antenatal ultrasound scan done at 18–20 weeks gestation. Serial scans at 2–4 weekly intervals will establish whether there is any progressive change before birth. The finding of reduced liquor, a distended thick-walled bladder or progressive dilatation >20 mm may be suggestive of an obstructive uropathy. Preterm delivery or antenatal surgical intervention is rarely indicated, except very occasionally in a male fetus where posterior urethral valves are causing renal compromise at <34 weeks gestation.

For most infants, postnatal investigation with several ultrasound scans over the first few months of life and sometimes a micturating cystourethrogram (MCUG) or renal isotope scan will be necessary. Until the results of these investigations are known, most infants will be given a small daily dose of prophylactic oral antibiotics (usually trimethoprim 1–2 mg/kg). This is to prevent urinary tract infections in those infants who may be at risk because they have vesicoureteric reflux. The radiological investigations are rarely urgent and some are more meaningful when the infant is a little older (for example, isotope scans).

All infants should be followed up postnatally as it is not easy to predict which infants will have significant ongoing dilatation, but most antenatally diagnosed fetal renal tract dilatation is found to be benign or transient on serial postnatal follow-up. A small number of infants will be diagnosed as having pelviureteric obstruction, multicystic dysplastic kidney or bladder outlet obstruction, but only the latter requires urgent diagnosis and surgery in the neonatal period.

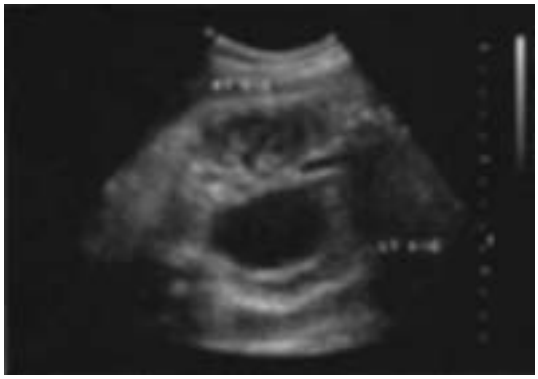


Figure 1.5 Ultrasound scan showing dilatation of left fetal renal pelvis.

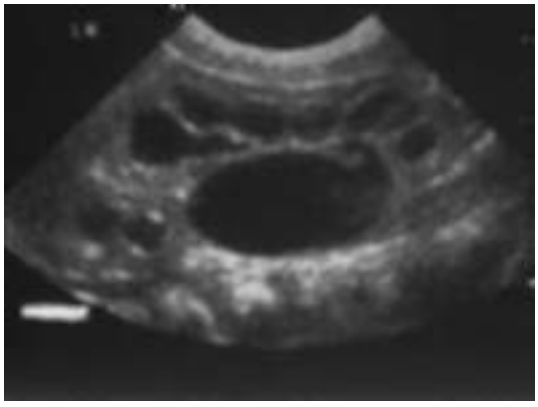


Figure 1.6 Postnatal scan showing dilation of renal pelvis.

2 Resuscitation of the newborn

Wherever babies are delivered there should be a person with adequate skill and experience in resuscitation immediately available throughout the 24 hours. The majority of babies can be resuscitated with a closely fitting mask and an inflatable bag with a valve. The equipment is cheap, simple to use, and can be carried in a small case. Some infants cannot be resuscitated by this method but require intubation, which to be successful should be done by a doctor or midwife with continual experience of the procedure.

Babies who have developmental brain abnormalities before labour may develop fetal distress during the stress of labour and may have difficulty in establishing spontaneous respiration. For this reason, the contributions of brain development and perinatal management in the causation of later cerebral palsy are often difficult to resolve.

Figure 2.1 Resuscitation kit in case.

Assessment

The following high risk factors indicate that resuscitation may be needed:

- fetal distress
- caesarean section
- preterm infant
- breech delivery
- forceps delivery
- twins
- maternal anaesthetic
- maternal diabetes
- rhesus incompatibility.

These factors predict about 70% of the babies needing resuscitation. The remainder arise unexpectedly. The APGAR scoring system is used to assess the infant's condition one minute and five minutes after birth. A numerical score is given for each of five features. The heart rate and respiratory effort determine the action to be taken.

Table 2.1 APGAR scoring system

	0	1	2
Appearance (colour)	Blue, pale	Body pink, extremities blue	Completely pink
Pulse (heart rate)	Absent	Below 100	Over 100
Grimace (response to stimulation)	No response	Grimace	Cry
Activity (muscle tone)	Limp	Some flexion in extremities	Active movements
Respiration (respiratory effort)	Absent	Slow irregular	Strong cry

Procedure

Suctioning the oropharynx

The rare indications for suction of the oropharynx are meconium aspiration and blood in the mouth. It is best not to use mucus extractors, as there is a risk of the operator swallowing or inhaling infectious material. Use a suction catheter (size FG 8) connected to the Resuscitaire or directly to a wall suction unit. The mouth can safely be suctioned but care must be taken in the oropharynx. This should be done under direct vision and is usually part of tracheal intubation. Do not blindly push the catheter as far as it will go since this can cause a vagally mediated bradycardia and apnoea and is invariably associated with a fall in oxygen saturation.

Administering facial oxygen

Set the oxygen flow rate to 5 l/min and hold the funnel-shaped mask just in front of the baby's face. The oxygen may be connected either to the funnel-shaped mask or to the bag and mask apparatus, but in the case of the latter, it is prevented from flowing out of the mask by the valve unless the bag is

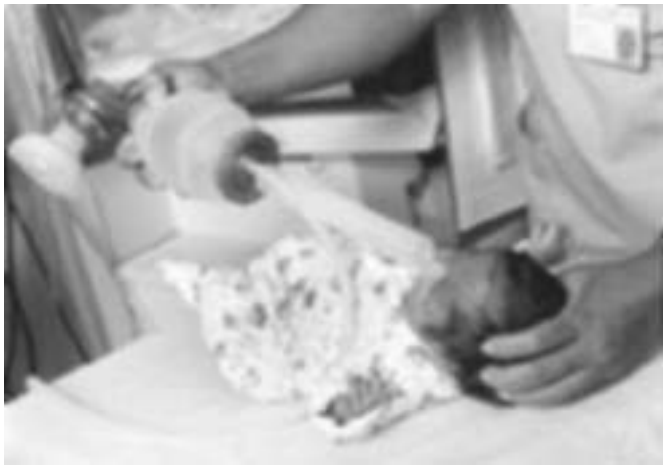


Figure 2.2 Giving oxygen.

compressed. However, it will come out of the corrugated tube that is attached to the other end of the bag, so turn it round and hold the end of this tube to the baby's face.

Using the bag and mask

If the infant does not breathe by 30 seconds after birth, the closely fitted mask is applied to the face with the head in the neutral position. For a right handed person, the left hand is used to hold the mask to the baby's face while the right hand squeezes the bag. Place the little and ring fingers of your left hand under the baby's chin, taking care not to push too hard. Alternatively the jaw is elevated with two fingers on the angle of the mandible. This prevents the head from moving around and straightens the upper airways, ensuring their patency. With the other fingers and thumb, apply the mask firmly to the baby's face to ensure a tight seal. A proper seal is confirmed when you squeeze the bag, as there is a characteristic rasping noise as the valve opens. If the seal is inadequate, the valve makes no noise and you will not feel any resistance when squeezing the bag. This can be practised with the mask against the palm of your hand. Use only the thumb and two fingers, rather than your whole hand, to squeeze the bag. Do not empty the bag but gently depress it to a few centimetres only. This will safeguard against a pneumothorax. The smaller the baby, the more gentle you must be. The rate should be maintained at 40/min with an inflation time of approximately 1 second. The first five inflations should be slightly prolonged as lung fluid is still present in the airway. Check you are producing an adequate chest expansion. Air can be used but oxygen should be used if it can be introduced into a side arm.

Figure 2.3 Operator holding mask with *right* hand to show how to place the infant's head in the neutral position.

Intubation

If there are no spontaneous respiratory movements at the end of one minute after birth or if the heart rate is less than 100 beats/min at any time the infant should be placed supine on a flat surface. A special resuscitation trolley is ideal. The laryngoscope is held in the left hand and passed over the infant's tongue as far as the epiglottis. The tip of the blade is advanced over the epiglottis about another 0·5 cm and is then withdrawn slightly. This presses the epiglottis against the root of the tongue, revealing the glottis. In the newborn the glottis is a slit in the centre of a small pink mound and the slit may expand into a triangular opening during a gasp. Gentle backward pressure on the infant's larynx by an assistant may help to bring the glottis into view. Secretions in the pharynx should be aspirated with a large catheter – for example, FG 9. The endotracheal tube held in the right hand is then guided through the larynx about 1–2 cm into the trachea. A metal introducer inside the endotracheal tube makes introduction easier but it is essential to ensure that it does not extend beyond the end of the tube.

Intermittent positive pressure should be applied at a rate of 40 times per minute with an inflatable bag with a valve.

The positive pressure applied should not usually be higher than 30 cmH_2O; otherwise there is a danger of rupturing the lung and producing a pneumothorax or pneumomediastinum. These low pressures are enough to induce a gasp reflex, which is then followed by normal respiratory movements of the chest. Occasionally in an infant with a severe lung problem, such as severe meconium aspiration or diaphragmatic hernia, higher pressures are needed. A return to a normal heart rate is a good sign that resuscitation is satisfactory. If the endotracheal tube has to be left in place for a short period the tube should be fixed to the cheek by adhesive tape or a special tube holder.

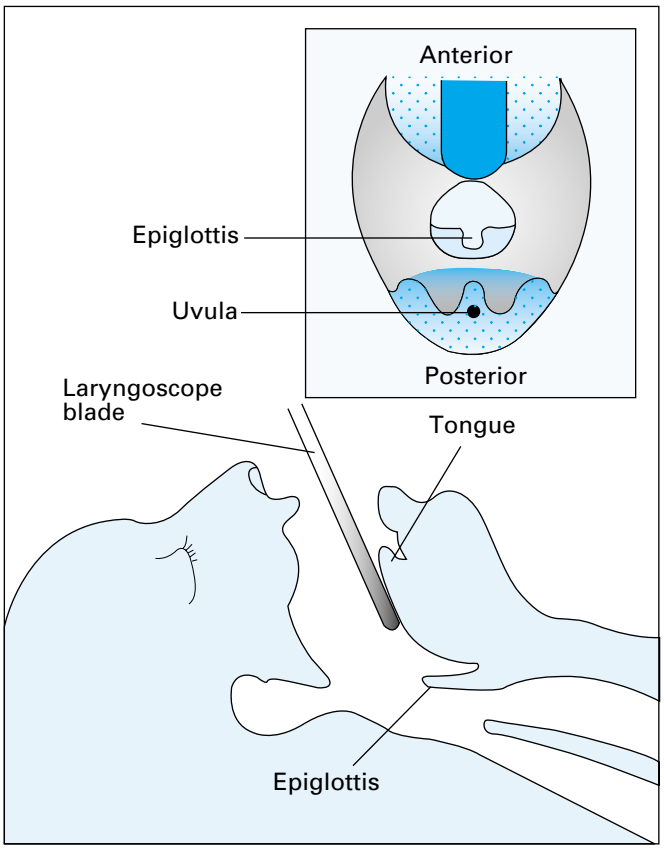


Figure 2.4 Passing the laryngoscope over the tongue.

Aspiration of secretions in the endotracheal tube and in the trachea can be carried out with a fine catheter (suction catheter FG 6). Either a straight or a shouldered tube is satisfactory for emergency resuscitation of the newborn. The shoulder on the endotracheal tube is designed to prevent the tube going too far down through the vocal cords and therefore into the right main bronchus, but this may still occur. If breath sounds are heard equally on both sides of the chest the tube is probably in the trachea and not beyond the bifurcation.

Intermittent positive pressure should be stopped every three minutes for about 15 seconds to determine whether spontaneous respiratory movements will start.

During prolonged apnoea the blood pressure is maintained initially but later falls. If the heart rate is less than 100/min a short period of cardiac massage should be given at the same time as efforts to start respiratory movements. Cardiac massage is carried out by applying firm pressure with two fingers over the lower sternum one finger's breadth below an imaginary line between the two nipples.

Hypothermia is a special hazard for infants who have been resuscitated and exposed during these procedures. Rapid initial drying of the infant with a warm towel is the most important preventive factor, but resuscitation should also be carried out under a radiant heater of at least 400 watts. After resuscitation, infants should be wrapped up and handed to their mothers for at least a minute or two even if they have to be placed in a transport incubator and taken to the special care unit. Most full term infants who have required resuscitation do not need to be admitted to the special care unit.

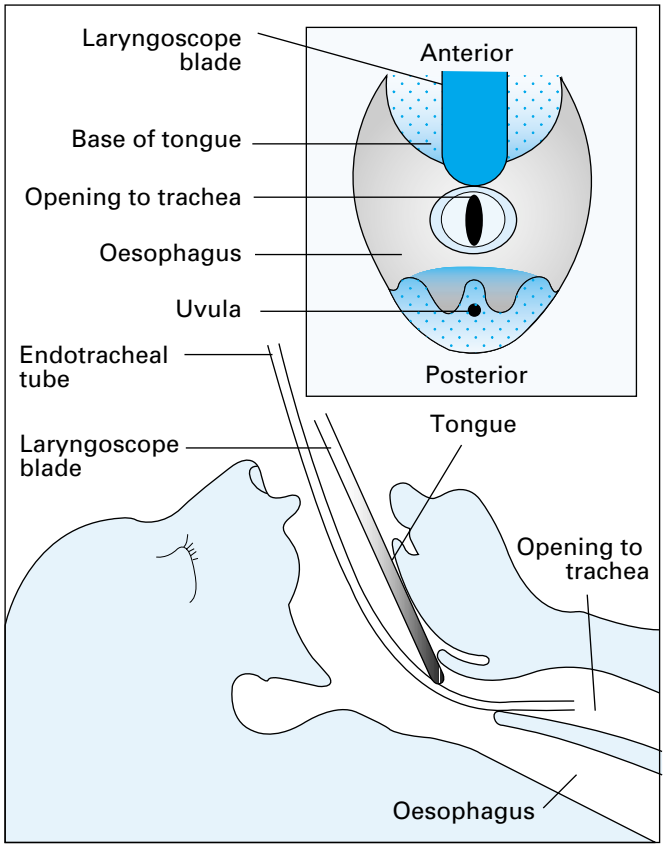


Figure 2.5 Pressing the epiglottis against the root of the tongue.

Failure to improve

The best sign that resuscitation has been successful is an increase in heart rate. If this does not occur within about 15 seconds the following should be considered. (1) The gas cylinders may be empty or the pipe line disconnected. (2) The endotracheal tube may have been misplaced into the oesophagus or have slipped out of the trachea during extension of the neck. If there is any doubt the tube should be removed and a fresh tube inserted immediately. (3) The endotracheal tube may be in the right main bronchus. After these possibilities have been excluded other diagnoses to be considered are pneumothorax, pulmonary hypoplasia associated with Potter's syndrome (renal agenesis with a squashed facial appearance and large, low set floppy ears), and diaphragmatic hernia.

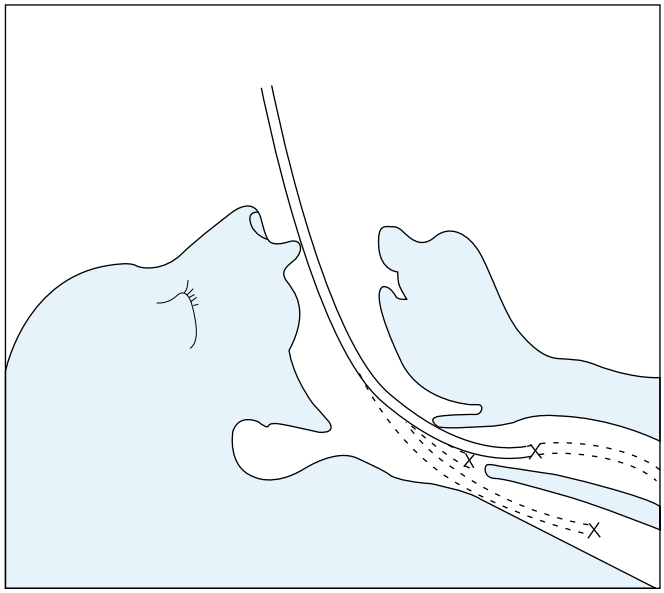


Figure 2.6 Misplacement of the endotracheal tube into the pharynx or oesophagus.

Drugs

Ventilation using bag and mask or intubation is usually effective in resuscitation of the newborn and drugs are rarely necessary.

If the mother has recently received pethidine or morphine, a chemical antagonist can be given to the infant. If the infant needs both intubation and drugs, intubation should always be performed first. The only chemical antagonist available is naloxone, but its period of action is short. The manufacturer's current recommended dose is 10–20 micrograms per kilogram body weight, which can be given intramuscularly or intravenously and may be repeated at 2–3 minute intervals. Alternatively, a single dose of 60 micrograms/kg body weight may be given intramuscularly at birth.

An adequate supply of oxygen quickly reverses acidosis and it is rarely necessary to consider giving intravenous sodium bicarbonate or glucose solution.

Table 2.2 Recommended doses

Drug	Concentration	Route	Dose (ml/kg)	Dose for 3 kg baby (ml)
Adrenaline	1 in 10 000 (0·1 mg/ml)	IT*, IV, IC	0·1	0·3
Sodium bicarbonate	8·4% (1 mmol/ml)	IV	2–4 (diluted 1 in 2 with water)	6–12
Glucose	10%	IV	1–2	3–6
Albumin	5%	IV	10–20	30–60

* Double the dose and add 2 ml of 0·9% sodium chloride solution for intratracheal route.

If there are no spontaneous respiratory efforts by three minutes after starting resuscitation, blood is taken for urgent pH and bicarbonate estimations. Without waiting for the result, the probable acidosis is partially reversed by giving 2–4 ml/kg of 4.2% sodium bicarbonate solution slowly at a rate that does not exceed 2 ml/min. The standard 8·4% sodium bicarbonate solution must be diluted 1 in 2 with sterile water for intravenous use. The solution should be given by a peripheral vein if possible, as the solution is hypertonic and may cause local vascular damage. If this is not possible, an umbilical vein catheter can be used in an emergency.

Ten percent glucose solution can be given if a glucometer shows hypoglycaemia.

Adrenaline can be given if there is asystole or there is persistent severe bradycardia. If there is no response, the adrenaline can be repeated. The use of volume expanders, albumin or 0·9% sodium chloride solution is not recommended routinely in the resuscitation of the newborn. Shock in the newborn is usually related to hypoxaemia and responds to administered oxygen but a volume expander may be needed where there is volume loss. The infant is transferred to the neonatal intensive care unit.

There is no specific treatment for the hypoxic-ischaemic encephalopathy that may follow perinatal asphyxia. The infant may be apnoeic and need continuous positive pressure ventilation, have fits, episodes of bradycardia, lethargy, or be reluctant to suck. Mannitol, frusemide, steroids, and phenobarbitone in high doses have been used to prevent or treat possible cerebral oedema but there is no evidence that they are effective.

These infants will also need long term follow-up to assess neurological development. The survival rate of full term newborn infants who have taken 20 minutes to breathe spontaneously is about 50% and about 75% of the survivors are neurologically intact.

When to stop

Poor outcome can be predicted when spontaneous respirations are not established by 30 minutes. If, in addition, there is no cardiac output, then survival cannot be expected. It is at this stage that attempts at resuscitation should cease.

When not to start

This can be an extremely difficult decision and should not be made by the most junior paediatrician, so begin resuscitation and call for help. If the heart rate has been recorded at any time during the second stage of labour, resuscitation should be attempted even if there is no heartbeat at birth. With fetal

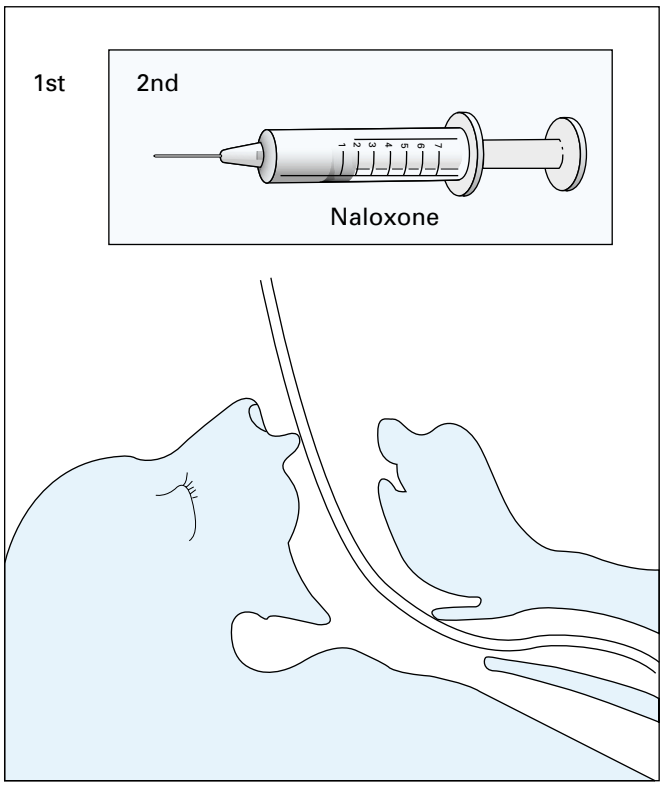
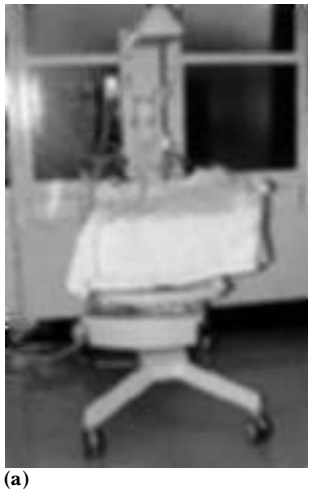


Figure 2.7 Intubation is performed before giving drugs.



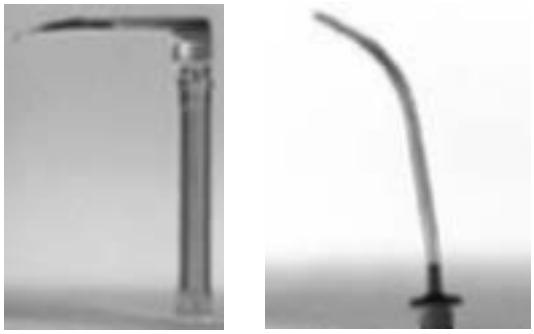
(a)

(b)

Figure 2.8 **(a)** Resuscitation trolley. **(b)** Bag with face mask.

monitoring it is uncommon for babies to die during labour and stillbirths are usually expected. Babies who have been dead for longer than 12 hours have an obvious "macerated" appearance with gross peeling of the skin.

A baby born at less than 22 weeks gestation cannot survive and often a paediatrician will not be called if the obstetrician is sure of the dates. Certain conditions are non-viable, such as anencephaly or gross hydrocephalus, but fortunately these are usually detected prior to birth and a decision is reached with the parents before delivery.



(a) **(b)**

Figure 2.9 **(a)** Small laryngoscope with straight 10 cm blade. **(b)** Neonatal endotracheal tube.

Vitamin K

Vitamin K is given to all babies as prophylaxis against haemorrhagic disease of the newborn. More predictable blood levels are produced by intramuscular compared to oral vitamin K. A single intramuscular dose protects against the early form of haemorrhagic disease, which occurs between the second and fourth day, and the late form, which occurs after three or four weeks. Following controversy on the safety of the intramuscular route, some paediatricians recommend that breastfed babies receive oral vitamin K in the neonatal period and during the subsequent two months. Infants receiving a commercial cows' milk preparation exclusively have an adequate intake of vitamin K.

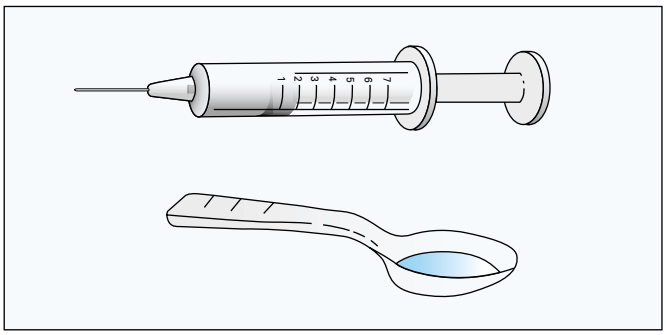


Figure 2.10 Vitamin K may be given intramuscularly or orally.

Discussion with parents

An infant who is breathing normally should be given to the mother after a rapid examination. She should be told that the infant seems to be normal and that it is common for infants to need help with breathing at birth. It is important to emphasise that the infant's progress will be no different from that of other infants who have not needed resuscitation. Most infants who have required intubation should go with their mothers to the postnatal wards and receive routine observation. Only if there has been a prolonged period before the establishment of spontaneous respiration should the infant be taken to the special care unit.

Figure 2.11 Discussion with parents.

When an infant dies

If resuscitation of an infant has not been successful, the most senior clinician involved, preferably accompanied by a midwife or nurse, should see the parents immediately to impart the news in a sensitive manner. The parents should be given as much factual information as is known to the clinicians at the time about the likely cause of the death. It should also be made clear that further discussion will need to take place, particularly if the death is sudden or unexpected. The term "birth asphyxia" should not be used as it may be misleading and there is increasing evidence that neonatal deaths and morbidity are more likely to be due to developmental brain problems or antenatal insults rather than problems occurring during birth. The family doctor should be informed of the death by telephone as soon as possible.

Parents should be encouraged to spend as much time as they want with their infant after death and should be supported by an experienced nurse or midwife. An infant with severe congenital anomalies can be suitably wrapped for the parents to cuddle. Many parents appreciate being able to see the anomalies with a supportive clinician present to discuss any questions they may have at the time. Other family members,

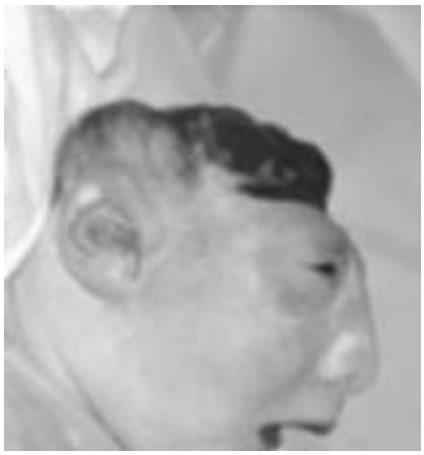


Figure 2.12 Anencephaly.

including siblings and grandparents, should also be encouraged to see the dead infant over the next few hours or days. It is now clear that the grieving process is helped by family participation in seeing and cuddling their dead infant. Many parents also welcome the opportunity to participate in the care of their dead infant by bathing and dressing the infant in baby clothes of their choice with the assistance of a sympathetic and unhurried nurse. However, health professionals should also be aware that there may be cultural and religious differences that should be respected if parents decline involvement in such procedures after the death of their baby.

Photographs of the dead infant and the family group together with the infant should be encouraged. Photographs should be kept in a safe place if the family do not want to take them away with them immediately, and they may still request them, sometimes even many years later. Hand and footprints and a lock of hair can also be taken as lasting and tangible mementoes for the family.

Discussion with parents will help to decide who will be best able to help the grieving family. There may be a supportive network of family and friends and for some, religious advisers or bereavement counsellors may provide support. If the death occurs in hospital, the family doctor and health visitor should be informed as soon as possible as they will have an important role in supporting the bereaved family in the long term. Siblings should be able to see the dead baby with their parents and should be reassured that they are loved, safe, and not in any way to blame for the distress of the parents. Children can accept the death of a baby in a very matter of fact way if they are allowed to be involved and are not excluded from the family at the time of the death. Expert help is rarely necessary for siblings with whom there has been communication and involvement in a manner that is understandable to the child in the bereaved family.

The potential benefits of a postmortem examination should be discussed with the parents by a senior doctor, preferably a consultant. The nature of the autopsy process should be explained in a sympathetic but factual manner and parents should be aware of the possibility of a limited examination of certain body cavities or organs if they do not want to contemplate a full examination. They should be made aware that the face of the baby will not usually be disfigured by the autopsy and that the body can be viewed again later, suitably clothed.

It is important that a health professional with knowledge of the local arrangements for the burial and cremation of infants is available to meet with the parents within a few days of the death in order to assist them to decide how they wish to proceed with funeral arrangements. Although there is a legal requirement to register the death of a baby within five working days of the death in the UK, the funeral does not have to take place immediately unless it is a requirement of the religious beliefs of the family. Some families prefer to take some time to decide about arrangements, particularly when the mother of the infant may not be able to be present immediately after a difficult birth.

A follow-up bereavement appointment should be arranged with the parents some weeks later. The results of the postmortem examination should be conveyed to them and they should be given the opportunity to ask questions about the illness and death of their infant. It is usually helpful to parents if the clinicians, including obstetricians or midwives who were present at the birth or involved in the care of the infant, can attend this bereavement discussion. An assessment of whether the family are likely to need expert help and support to progress with their bereavement can also be made at this time. Some parents may also find it helpful to talk to other bereaved families or voluntary organisations providing support, for example SANDS (Stillbirth and Neonatal Death Society).

3 Infants of low birthweight

Low birthweight infants weigh 2500 g or less at birth. Infants may be small at birth owing to a short gestation period (born too early) or because of a restricted growth rate. When the period of gestation is less than 37 completed weeks the infant is called preterm. A baby with restricted growth rate is "light for dates" or small for gestational age (SGA) and may be either malnourished or, rarely, hypoplastic (for example, an infant with a chromosome abnormality). Some babies who are preterm are also light for dates.

The length of gestation can be calculated from the first day of the mother's last menstrual period, provided the periods are regular. The routine ultrasound examination carried out before 18 weeks is considered to be the most accurate method of assessing gestational age. Detecting restriction in growth of the fetus is difficult, but palpation of the fetus can be supplemented by serial ultrasound measurements of the fetal skull, abdominal girth, femur length and crown–rump length.

The gestational age of the infant can be assessed by a detailed neurological examination of the infant, as the development of the central nervous system is related to the gestational age. Scoring systems using both the neurological development and specific external features can be used to estimate the gestational age but require considerable experience for accuracy.

Most neonatal units consider that infants with a birthweight below the 10th centile for the gestational age are light for dates, although a more accurate definition is below the third centile with an adjustment for maternal size.

Ultrasound studies have shown that when intrauterine malnutrition starts early in pregnancy the head circumference and weight are in proportion with each other, but when malnutrition starts late in pregnancy the head is disproportionately large, owing to relatively normal growth of the brain.

The preterm infant is especially prone to developing hypothermia, the respiratory distress syndrome, infection, and intracranial haemorrhage. The light for dates infant is particularly prone to hypothermia, hypoglycaemia, and hypocalcaemia.

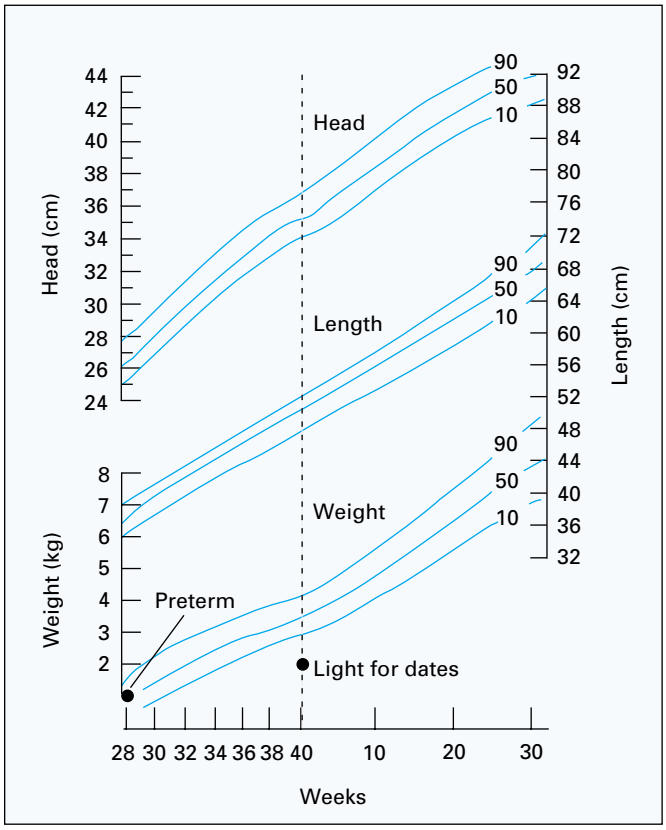


Figure 3.1 Growth chart showing birth weights of preterm and light for dates infants.

Temperature

Small infants become hypothermic quickly. Heat loss may be considerable because they have a large surface area in relation to body weight and they are also deficient in subcutaneous fat, which provides insulation. They also lack "brown fat", which is usually present in a full term baby and can be metabolised rapidly to produce heat.

Hypothermia is associated with a raised metabolic rate and increased energy consumption. To prevent hypothermia small infants should be kept in a high constant environmental temperature. Excessive heat loss by radiation can be minimised by an additional tunnel of Perspex, a heat shield, placed immediately over an infant in an incubator. The dangers of hypothermia can be reduced by carrying out resuscitation under a heat lamp or radiation heat canopy, bearing in mind the danger of burns if the lamp is too close. When an infant is transferred from one hospital to another for special care, there is a danger of hypothermia. A swaddler made of aluminium foil has been designed to prevent this and is commercially available, but wrapping the baby with Gamgee wool is also satisfactory.

Figure 3.2 Open incubator with infant receiving artificial ventilation.

The portable incubator must be kept warm continuously and ready for immediate use.

Infection

A separate plastic apron or waterproof gown should be used for nursing each infant, mainly to prevent soiling the nurses' clothes.

Scrupulous washing of the hands should be carried out before and after touching each infant. Washing with soap and water is adequate, although many units use a disinfectant soap solution. Scrubbing with a brush is unnecessary. Although this principle is simple, it may be difficult to ensure that it is carried out. Proper use of elbow taps is often difficult because of their poor design. Using disposable plastic gloves when changing infants' napkins may reduce cross-infection.

Ideally, any infant who develops an infection should be barrier nursed in a separate cubicle.

Feeding

Preterm infants have poor sucking and cough reflexes and methods of feeding must prevent aspiration into the lungs. Infants who are not able to suck are given frequent tube feeds of small volumes or continuous intragastric feeds to avoid sudden falls in arterial oxygen concentrations and apnoeic attacks, which are associated with abdominal distension. Intravenous feeding with glucose amino acid solutions and lipids is used for infants who do not tolerate tube feeding, but a scrupulous aseptic technique in managing the solutions is necessary to avoid septicaemia.

Early feeding, within two hours of birth, prevents hypoglycaemia and reduces the maximum plasma bilirubin concentrations. Asymptomatic hypoglycaemia can be detected early by performing regular blood glucose measurements using a glucometer on all babies of low birthweight in the first 24 hours. If the infant is feeding well and the blood glucose level is consistently above 2.6 mmol/l, the tests are stopped after 24 hours. From the second week onwards vitamin supplements are added so that infants receive an additional dose of 400 units of vitamin D and 50 mg of vitamin C daily. Vitamin preparations usually contain small amounts of vitamin B complex and vitamin A. Vitamin supplements should be given until the age of two years and additional iron supplements until the age of six months.

The low birthweight infant at home

As more extremely low birthweight infants are surviving, community health professionals need to be aware of a few specific needs and problems of these infants in the first year of life.

Home oxygen

Infants who still require a small amount of oxygen but who are feeding and are otherwise well can often be cared for at home by their parents. The family doctor may be asked to prescribe an oxygen concentrator for use at home as the continuous provision of oxygen cylinders is impractical in the long term. A small portable cylinder which can be carried on a pram or pushchair will be useful to enable the infant to accompany the family on brief trips outside the home environment.

Figure 3.3 Light for dates: 1800 g at 40 weeks.

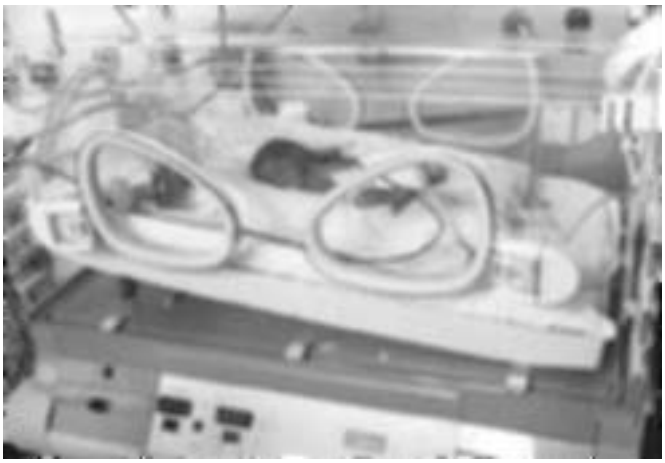


Figure 3.4 Closed incubator.

Figure 3.5 Hand washing prevents cross-infection.

Figure 3.6 Syringe pump for giving milk.

A community neonatal or paediatric nurse will support the family at home and will monitor the oxygen saturation of the infant intermittently until no additional oxygen is required, sometimes weeks or even months after discharge from hospital. Family carers will have been taught basic life support by the community nurse.

Figure 3.7 Oximeter for measuring blood oxygen saturation.

Respiratory problems

Preterm infants, particularly those who have required artificial ventilation, have an increased risk of respiratory illnesses in the first year of life. Wheezing is common in association with viral infections, but only those infants with a strong family history of atopy have an increased incidence of asthma. Smoking should be strongly discouraged in homes where there is a preterm infant.

Immunisations

Routine immunisation with DTP triple vaccine, Hib, meningococcal C and oral polio vaccine can be commenced at two months after birth, irrespective of the gestational age of the infant. Even infants of very early gestation have been shown to mount an appropriate immune response to childhood vaccines. In practice, many preterm infants will still be in the neonatal unit at two months of age and some will have had their immunisations commenced before discharge from hospital. The administration of oral polio vaccine will be delayed by three months if intravenous immunoglobulin has been given to the infant at risk of serious infection.

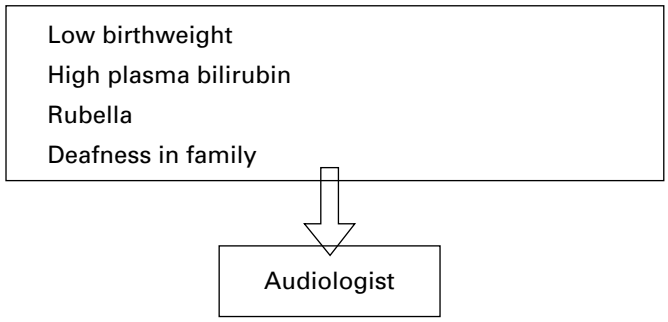


Figure 3.8 Indications for direct referral to the audiologist in the neonatal period.

Vitamins and iron supplementation

Preterm infants require vitamin A, B, C, and D supplementation until two years of age. Some extremely immature infants may also require additional phosphate for adequate bone mineralisation during the period of rapid growth in the first year of life.

Additional oral iron is usually given from four weeks to six months of age.

Box 3.1 Supplements for preterm infants

- Vitamins A, B, C, D until two years of age
- Phosphate in the first year for extremely preterm infants
- Iron from four weeks to six months

Growth and development

The growth and development of preterm infants born at <30 weeks gestation should be assessed according to their corrected age (chronological age − period of prematurity). Most will have caught up with their peers born at full term by around 18 months to two years of age. Some extremely low birthweight infants, particularly those born at <26 weeks gestation, remain <3rd centile in all growth parameters despite being otherwise healthy. Around 90% of infants <28 weeks gestation who survive to go home achieve normal developmental milestones for their corrected age, but many experience minor motor and educational difficulties by school age. Myopia is also common in children who were preterm.

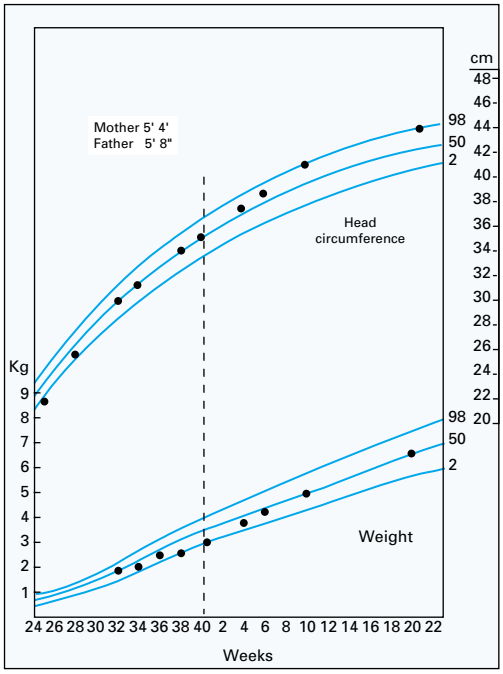


Figure 3.9 Growth chart of a preterm infant.

4 Breathing difficulties in the newborn

Breathing difficulties in the newborn are called respiratory distress. One or more of the following features are present:

- respiratory rate over 60 per minute
- an expiratory grunt
- subcostal or intercostal recession or sternal retraction
- cyanosis.

Respiratory distress syndrome

The respiratory distress syndrome (hyaline membrane disease) is the commonest cause of respiratory problems in the newborn and a common cause of death in preterm infants. Cerebral ischaemia and haemorrhage or lung damage may occur in the acute phase and cause death or long term morbidity. Hypoxia before, during, or after birth is a predisposing factor.

The cause is a deficiency of pulmonary surfactant, a substance normally present on the alveolar walls. Surfactant lowers the surface tension in the alveoli so that during the first few breaths the same pressure is required to inflate them all. This produces uniform inflation of all alveoli. Surfactant also prevents the alveolar walls from collapsing during expiration. Without surfactant the surface tension is great in the smaller alveoli, causing them to collapse, while large alveoli continue to expand easily. Thus there is uneven expansion, with increasingly widespread alveolar collapse. Surfactant production in the fetal lung increases with gestational age and reaches adequate levels for normal lung function by about the 36th week. At 27–31 weeks 35–50% of all infants are affected by the respiratory distress syndrome. The use of antenatal steroids given to the mother has reduced this prevalence in infants of less than 34 weeks gestation.

Microscopy of postmortem specimens often shows the presence of amorphous material lining the terminal bronchioles and alveoli, which is called hyaline membrane. There are also multiple areas of alveolar collapse.

When animal surfactant extract or artificial surfactant is placed in the trachea at birth or during the first day of life, the severity and mortality of the respiratory distress syndrome are reduced. Supplementary surfactant has been shown to be effective in babies of more than 26 weeks of gestation. Supplementary surfactant is given at birth to all babies less than 30 weeks gestation and to more mature infants who have developed specific features of surfactant deficiency that suggest that the course of the disease will be of moderate or greater severity (rescue treatment).

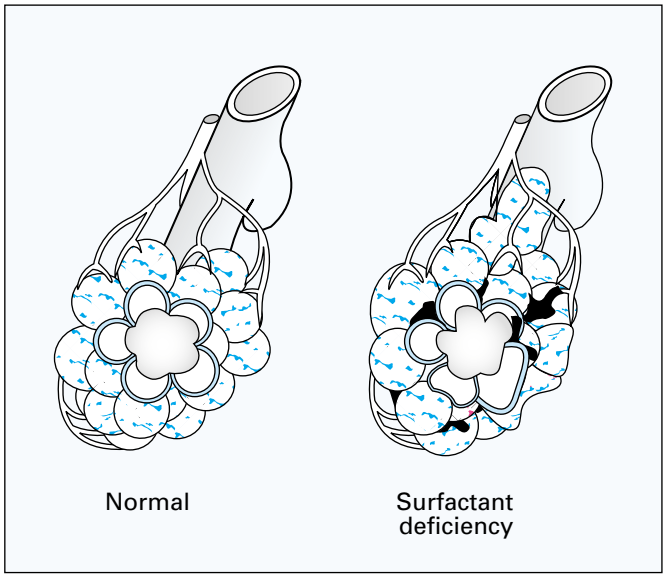


Figure 4.1 Cross-section of lungs showing uneven expansion with alveolar collapse in the surfactant deficient lung.

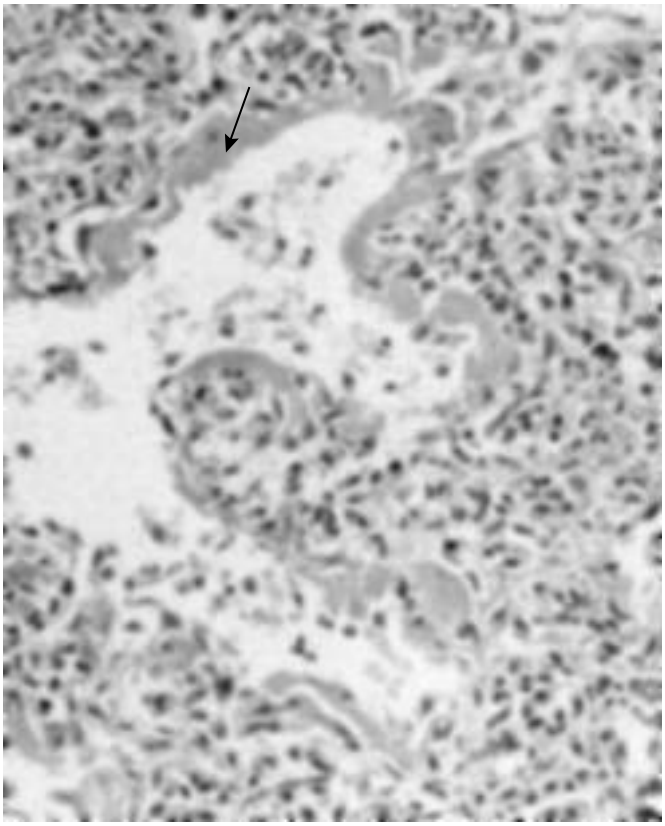


Figure 4.2 Arrow indicates hyaline membrane.

Clinical features and general management

An expiratory grunt, a respiratory rate over 60 per minute, and recession of the chest wall begin within four hours of birth. Central cyanosis may appear later. Usually auscultation of the lungs reveals only reduced breath sounds, but crepitations may be present. The present approach to intensive care is to monitor the blood arterial oxygen, carbon dioxide, and pH levels from birth and to start treatment early. This may prevent the appearance of the features described above, especially

where artificial ventilation is used electively from birth in infants weighing less than 1 kg.

In most infants the chest radiograph shows normal lung fields in the early stages of the disease, but later a fine reticular pattern or generalised loss of translucency (the ground glass appearance) may be present. The contrast between the air in the bronchi and the opaque lung fields produces an air bronchogram. The greatest value of the chest radiograph is in excluding other conditions (see below).

The object of management is to support infants until further surfactant is produced. Infants should be handled as little as possible, as their condition may deteriorate abruptly. Elective monitoring of blood oxygen saturation, heart rate, temperature and respiratory rate has reduced the need to handle the infant. Infants should be given oxygen, their arterial blood oxygen and carbon dioxide concentrations should be monitored, and body temperature, fluid, and electrolyte balance maintained.

Oxygen treatment and measurement

During oxygen treatment the arterial oxygen tension should not fall below 6·7 kPa (50 mmHg) or rise above 12 kPa (90 mmHg). Below 6·7 kPa the risk of cerebral palsy and mental retardation increases and above 12 kPa there is a possibility of retrolental fibroplasia and subsequent blindness. Some infants may need high concentrations of inhaled oxygen to prevent hypoxia. This can be given safely only if there are facilities for making frequent estimations of the blood oxygen concentrations.

The ideal arterial carbon dioxide level is 5–7 kPa (37·5–52 mmHg). Lower carbon dioxide levels may cause cerebral artery constriction and reduce cerebral blood flow. Higher carbon dioxide levels may lead to cerebral artery dilatation followed by systemic hypotension or intraventricular haemorrhage.

There are several methods of measuring oxygen concentrations in the blood. Firstly, samples can be taken from an umbilical artery catheter at regular intervals or they can be obtained by repeated puncture of one of the radial arteries.

Secondly, an oxygen sensitive electrode implanted into the tip of an umbilical artery catheter measures oxygen levels continuously, but the electrode must be calibrated at regular intervals with samples of blood aspirated through the lumen of the same catheter.

Thirdly, a heated transcutaneous oxygen and carbon dioxide electrode increases the blood flow to the skin, where the gas tensions are measured. The electrode must be moved every four hours to avoid burns. The method is not invasive and the electrode is calibrated with a gas mixture. The levels shown by the transcutaneous electrode must be compared with levels estimated in arterial blood at regular intervals. The blood samples are also used to measure pH, bicarbonate, and base deficit.

Artificial ventilation

Results of artificial ventilation have improved considerably and deaths are now usually due to complications rather than respiratory failure. Respiratory tract infection and displacement of the endotracheal tube are constant hazards and this form of treatment can only be undertaken where there are adequate nursing and medical staff and enough patients for an effective service to be provided.

Box 4.1 Management

- Temperature
- Arterial oxygen
- Carbon dioxide
- pH
- Blood pressure
- Plasma, Na, K, creatinine

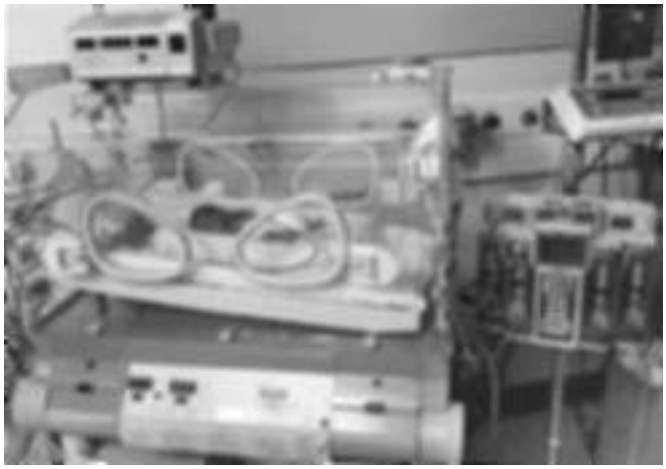


Figure 4.3 Intensive care in a closed incubator.

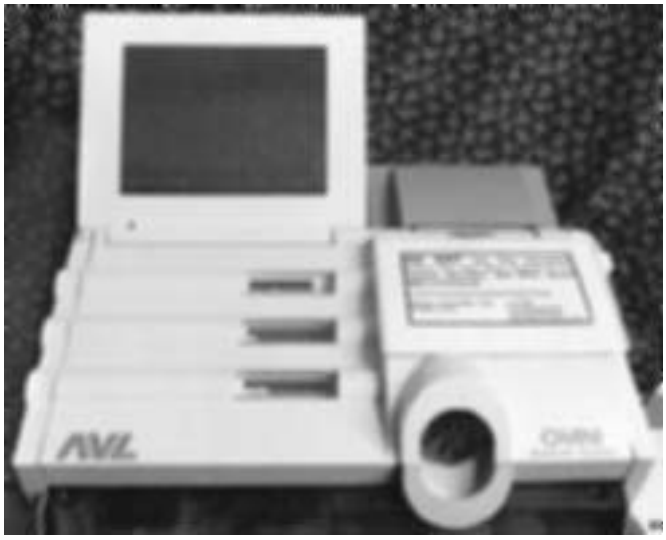


Figure 4.4 Blood oxygen analyser.

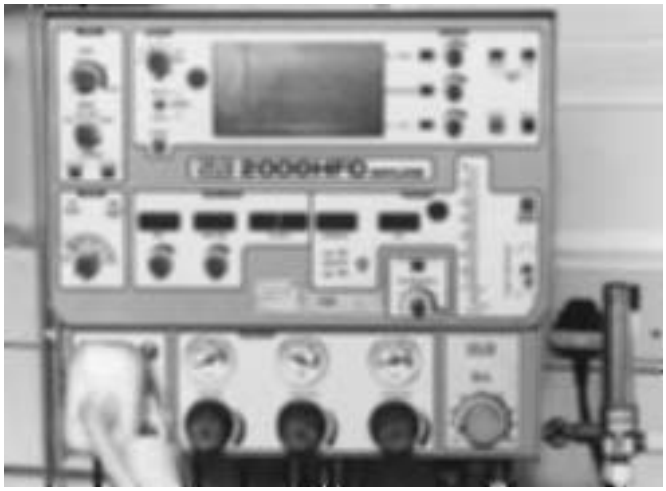


Figure 4.5 Ventilator.

Positive pressure ventilation is needed if one of the following is present:

- failure to establish effective spontaneous respiration at birth
- recurrent apnoeic attacks
- arterial oxygen tension (PaO_2) less than 7 kPa in 60% oxygen
- arterial carbon dioxide tension ($PaCO_2$) more than 7 kPa
- rapid deterioration in blood gas values associated with clinical deterioration.

Other aspects of treatment

Incubators are used to improve observation and to provide warmth and humidified air. There are two types: the conventional incubator and an open type. A perspex head box can be used to maintain an oxygen concentration above that in the air.

Adequate fluids and electrolytes must be given. The methods will vary with the severity of the respiratory problem. Milk can be given by continuous gastric infusion. Alternatively, small volumes of milk can be given intermittently at frequent intervals. Some infants need intravenous fluids which may include total parenteral nutrition (TPN) which contains glucose, amino acids, lipids, and vitamins.

The group B Streptococcus may produce a clinical picture similar to that of respiratory distress syndrome and infants with the respiratory distress syndrome are given antibiotics to cover this possibility. Intravenous antibiotics must be given until blood culture results are known, as the syndrome cannot be differentiated from group B streptococcal pneumonia in the early stages.

Complications

A pneumothorax should be suspected in any infant who deteriorates rapidly for no obvious reason. The diagnosis may be confirmed quickly by transilluminating each side of the chest with a powerful fibreoptic light source. A chest radiograph can be used to confirm the condition, but if symptoms are severe the pneumothorax can be drained before the radiograph has been obtained. A disposable cannula is inserted into the pleural space and connected to an under water seal. Pneumothorax may also follow intermittent positive pressure ventilation during resuscitation of the newborn or it may occur as a result of the initial vigorous spontaneous respiratory efforts of a normal infant.

Cerebral lesions

Periventricular haemorrhage is the most common form of haemorrhage in the newborn infant. Advances in neonatal care have led to increased survival of preterm infants who are susceptible to periventricular haemorrhage. When ultrasound scans have been performed routinely within the first few days of life about 10% of infants with a birthweight less than 1500 g have evidence of periventricular haemorrhage. The majority of these haemorrhages are small and localised and they are associated with a good prognosis but early mortality is high and long term handicap is common in infants with extensive haemorrhages. These haemorrhages originate in the subependymal germinal matrix and may spread to involve the ventricular system or may extend into the cerebral parenchyma adjacent to the lateral ventricle. Most of the infants with

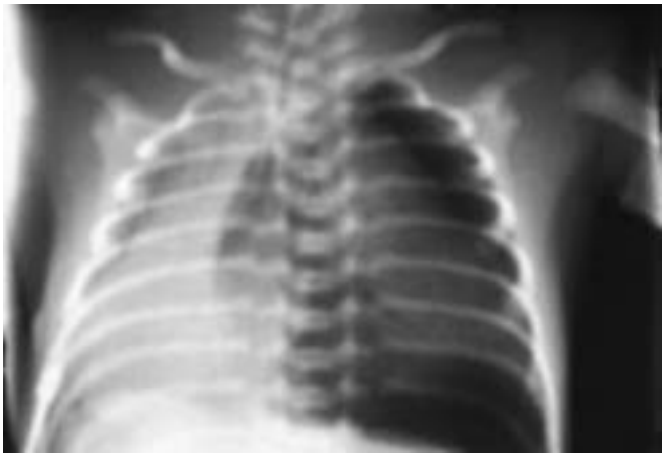


Figure 4.6 Pneumothorax.

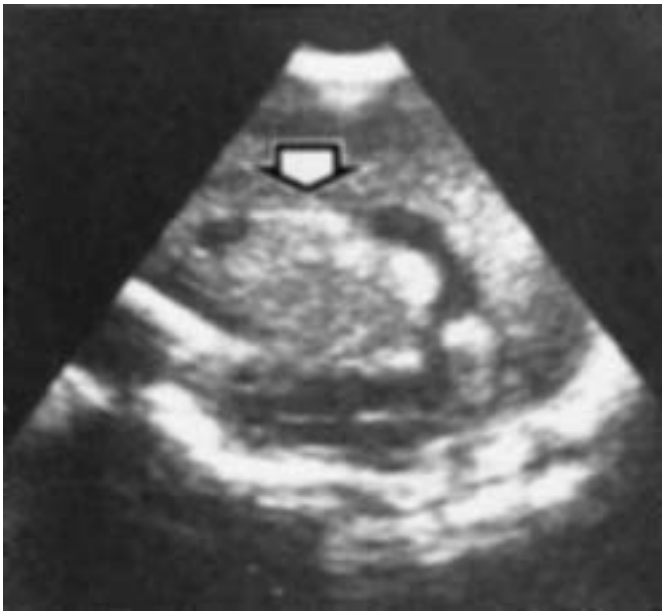


Figure 4.7 Arrow shows intraventricular haemorrhage.

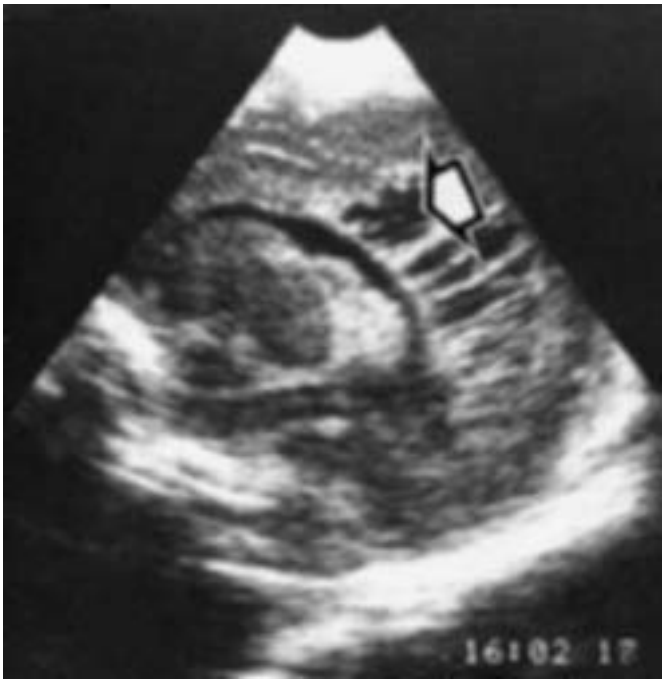


Figure 4.8 Arrow indicates periventricular leucomalacia.

periventricular haemorrhage have no specific symptoms but some have recurrent apnoeic attacks, impaired spontaneous limb movements, and hypotonia or sudden severe clinical deterioration.

Periventricular leucomalacia is considered to be related to hypoxia and ischaemia of the brain. Small cysts are shown in the periventricular region in ultrasound scans of the brain taken after the third week of life. These lesions are associated with later major handicaps such as cerebral palsy, blindness, or deafness.

About 15% of infants who survive the neonatal period have some degree of developmental problem but about 7% have a severe problem.

Bronchopulmonary dysplasia

This is a chronic respiratory disease occurring in preterm babies who are very preterm or have needed artificial ventilation. The causes include inflammation, barotrauma from artificial ventilation, recurrent aspiration of milk or postnatal infection. The infant requires increased inflation pressures or ambient oxygen concentrations as coarse streaking and areas of hyperinflation appear on the chest radiograph after the first two or three weeks of life. After weaning from the ventilator, supplementary oxygen is often required for several months and those with severe disease die usually about the third month of life. The survivors have a high prevalence of recurrent episodes of wheezing but usually have no further clinical respiratory symptoms after about 18 months of age. The duration of oxygen dependency may be reduced by dexamethasone but further studies are needed to determine the optimal time to start it. A transcutaneous oxygen saturation monitor is used to monitor the oxygen requirement and to assess the effects of treatment.

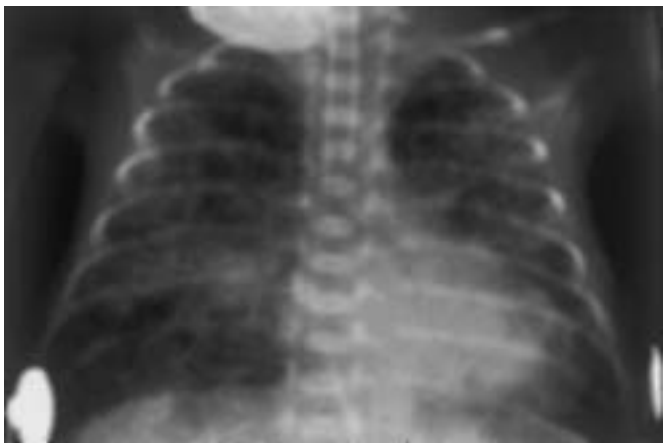


Figure 4.9 Bronchopulmonary dysplasia.

Other causes of respiratory problems

The main causes are pulmonary, cardiac, and central nervous system disorders.

Transient tachypnoea of the newborn is found in full term infants after a short labour or elective caesarean section, before or at term. It can occur in preterm infants but there may be difficulty in distinguishing it from mild respiratory distress syndrome. Transient tachypnoea usually resolves within 48 hours. The chest radiograph often shows a streaky appearance of the lung fields but may be normal. This syndrome may be due to failure of normal reabsorption of the lung fluid at birth or it may be a mild form of respiratory distress syndrome.

Diaphragmatic hernia may present with difficulty in resuscitating the baby or with a raised respiratory rate and an apex beat on the right side. The hernia is most commonly left sided and the heart is often displaced to the right. Most diaphragmatic hernias are detected by antenatal ultrasound scans. The diagnosis is confirmed by a chest radiograph that shows loops of small gut or solid organs in the thorax. It may take 12 hours from birth for air to reach the colon and produce the characteristic radiographic appearances.

The first symptoms of *congenital heart disease* are often noticed by the nurse or mother when the infant has dyspnoea during feeding or is reluctant to feed. There may be no murmurs with some lesions. The respiratory rate is raised and there may be recession of the chest wall. Excessive weight gain and enlargement of the liver are early confirmatory signs. The edge of the liver is normally about 2 cm below the costal margin in the right midclavicular line in the full term newborn.

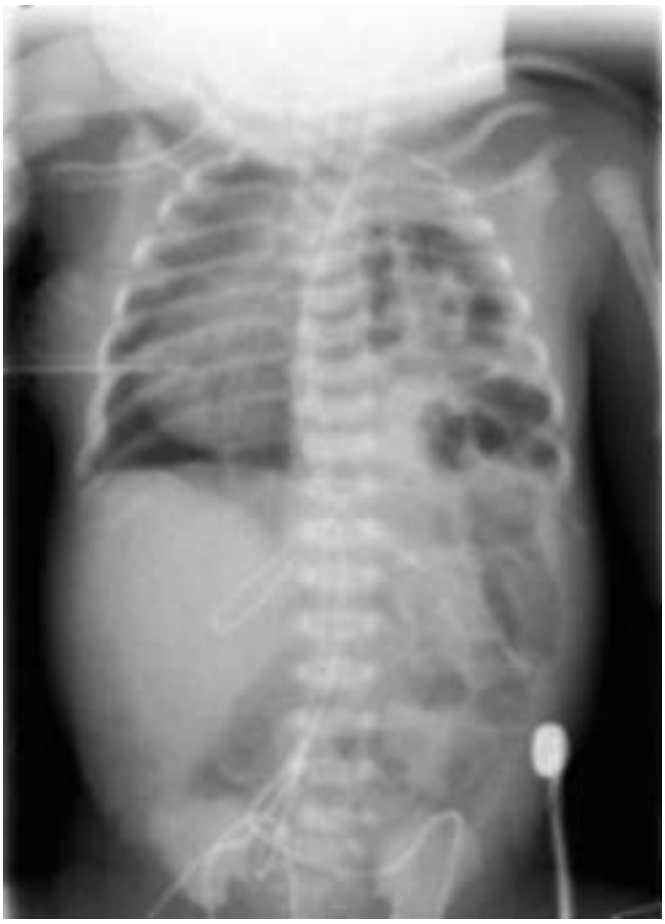


Figure 4.10 Diaphragmatic hernia.

Pneumonia may occur if there has been rupture of the membranes for longer than 24 hours: the infant may inhale infected liquor before birth and so develop pneumonia. If the mother has had ruptured membranes for over 24 hours before delivery antibiotics should be considered for the mother to prevent or treat pneumonia in the fetus.

Group B streptococcal infection may present with a raised respiratory rate and a chest radiograph may show extensive areas of consolidation in both lungs or may appear normal.

Meconium aspiration usually occurs in an infant who has become hypoxic before delivery. The infant may start respiratory movements before the mouth and pharynx have been cleared. Aspiration of meconium may cause bronchial obstruction, secondary collapse, and subsequent infection of the distal segments of the lungs. Oxygen and antibiotics are often needed and mechanical ventilation may be required for a few days.

Preterm infants have a poor cough reflex and material such as regurgitated milk in the pharynx is easily aspirated into the lungs and may cause pneumonia. Signs may be minimal and are often limited to a small increase in respiratory rate, but a chest radiograph may show extensive changes.

Severe anaemia may cause a raised respiratory rate with a metabolic acidosis.

Choanal atresia or stenosis – a congenital posterior nasal obstruction makes it difficult for newborn infants to breathe, as they depend on a clear nasal airway. An oropharyngeal airway produces immediate improvement and an ENT surgeon should be consulted. The diagnosis can be confirmed quickly by noting the absence of movement of a wisp of cotton wool placed just below the nares.

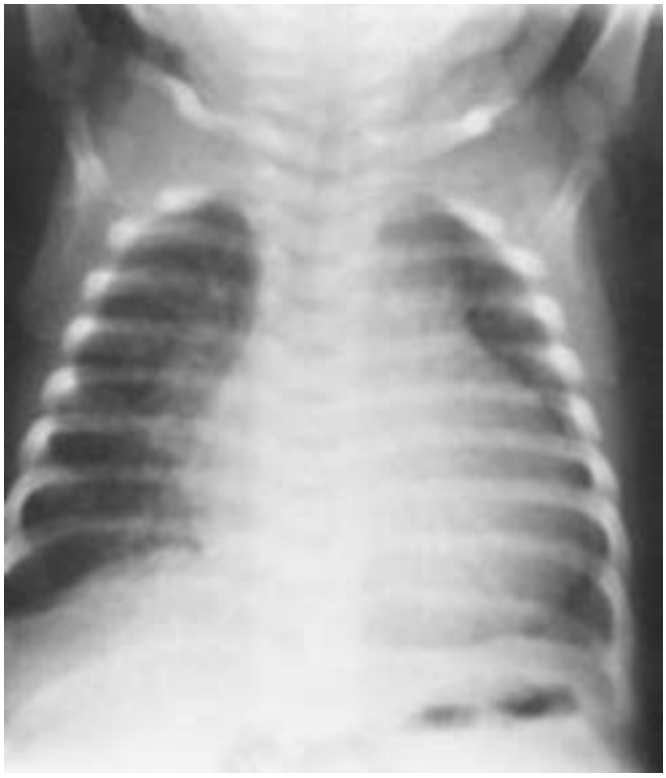


Figure 4.11 Large heart.

Apnoea

Apnoeic attacks are defined as episodes of cessation of respiratory movements for more than 10 seconds. They may be due to cerebral hypoxia during the perinatal period, hypoxaemia due to the respiratory distress syndrome, hypoglycaemia, or meningitis. An ultrasound scan of the brain may help in determining the presence and site of any intracranial haemorrhage.

Morphine or pethidine given to the mother before delivery may cause apnoea in the newborn (see p. 6).

Recurrent apnoeic episodes are common in preterm infants of less than 32 weeks gestation. The babies are otherwise well and the episodes start when the infants are a few days old. The apnoea may be accompanied by bradycardia and a fall in oxygen saturation levels. Recent research suggests that some episodes are of central and others of laryngeal origin. The diagnosis of apnoea of prematurity can be made only after excluding other causes; which are:

- pulmonary disease
- airway obstruction, for example due to aspiration of feeds
- hypoglycaemia
- hypocalcaemia
- intracranial haemorrhage
- convulsions, which may be misdiagnosed as apnoea
- cardiac disease.

Several attacks may occur daily for the first month of life. These may be difficult to manage and may be followed by cerebral palsy at a later date if the episodes are prolonged. The various forms of treatment include stimulation of a limb during attacks or prophylaxis with oral caffeine or continuous positive airways pressure.

Figure 4.12 Apnoea sensor with alarm.

Box 4.2 Causes of apnoea

- Prematurity
- Hypoxia
- Intracranial haemorrhage
- Hypoglycaemia
- Meningitis
- Drugs

5 Birth trauma

Cephalhaematoma

Cephalhaematoma is a subperiosteal haemorrhage that is limited by surrounding sutures. In most cases there is probably a hairline fracture of the underlying cranial bone, which may be difficult to demonstrate but is unimportant since it affects only the outer table. There is usually no brain damage. A surprisingly large amount of blood may be present and blood transfusion is occasionally required. The lesion may not be noticed until the third day.

During resolution a calcified rim may appear, which wrongly suggests a depressed fracture, or there may be a hard swelling that takes several months to disappear.

Cephalhaematoma must be distinguished from caput succedaneum, which is a soft tissue swelling due to oedema of the part of the head presenting at the cervix and which is not limited by the sutures.

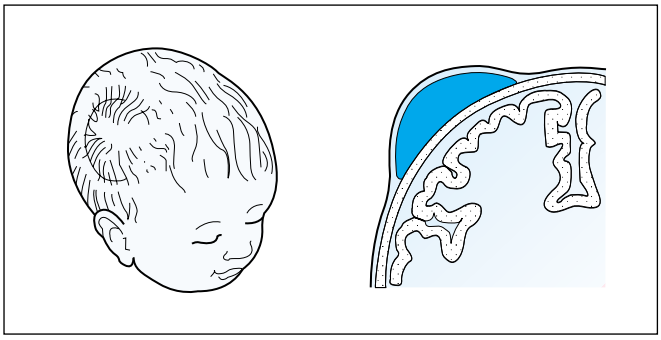


Figure 5.1 Cephalhaematoma.

Vacuum extraction injuries

Vacuum extraction is often associated with a "chignon", which is subcutaneous oedema where the cap has been applied. In some cases a large haematoma may form and occasionally the skin becomes necrotic. Nevertheless, the skin grows rapidly from the borders to cover the area within a few weeks.

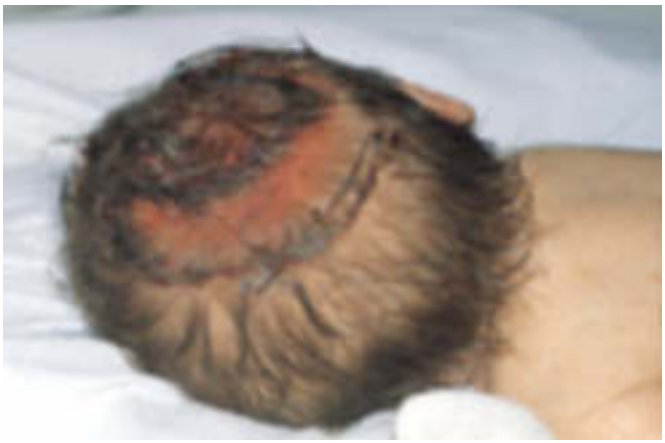


Figure 5.2 Chignon.

Reduced arm movements

Fractures of the clavicle or humerus and stretching of the brachial plexus present as lack of spontaneous movement and an absent Moro reflex in the arm on that side. The clavicle and humerus should be radiographed. Fracture of the clavicle needs no treatment, but if the humerus is fractured the arm should be splinted to the infant's body with a crêpe bandage. Injuries to the brachial plexus (Erb's palsy) will be recognised as lack of spontaneous movement of the shoulder and arm with the hand held in dorsiflexion at the wrist (waiter's tip position). Movements of the fingers are associated with good long term prognosis with spontaneous resolution. Early referral to a physiotherapist and orthopaedic surgeon is advised as injuries to the brachial plexus sometimes have a poor prognosis. Diaphragmatic paralysis often accompanies stretching of the brachial plexus and may cause a raised respiratory rate. Another condition that may be mistaken for injury of the brachial plexus is temporary paralysis of the dorsiflexors of the wrist caused by pressure on the radial nerve in the radial groove of the humerus. A fracture of the humerus may injure the radial nerve at this site.

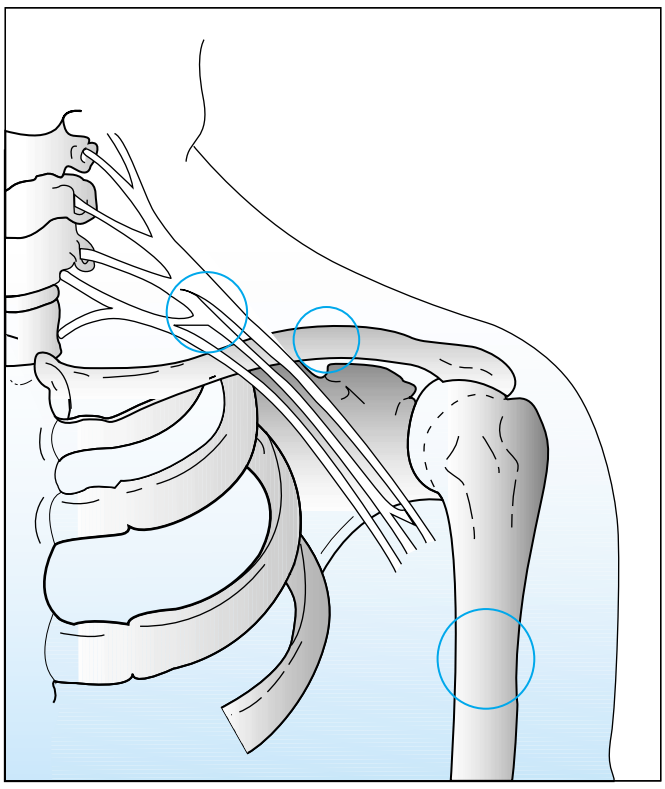


Figure 5.3 Three sites of injury causing reduced arm movements.

Intracranial injuries

When the fetal head is large in relation to the pelvic outlet or delivery is precipitate or by the breech, there is a risk of extracerebral haemorrhage due to laceration of the tentorium cerebelli or falx cerebri affecting a venous sinus.

Infants may be unduly lethargic or especially irritable shortly after birth and there are usually no helpful confirmatory signs then. Later there may be pallor, a high pitched cry, poor muscle tone or increased tone, convulsions,

reluctance to suck, vomiting, or apnoeic attacks, and fast or periodic breathing. Rarely the tension of the anterior fontanelle is raised, the heart rate is slow, or the pupils fail to react to light. A blood glucose test should be performed to exclude hypoglycaemia, a plasma calcium concentration estimated, and a lumbar puncture considered to exclude meningitis.

Infants should be nursed in an incubator for easy observation but if they weigh over 3000 g, the thermostat should be turned to the minimum level. This type of haemorrhage is now rare. The prognosis is poor.

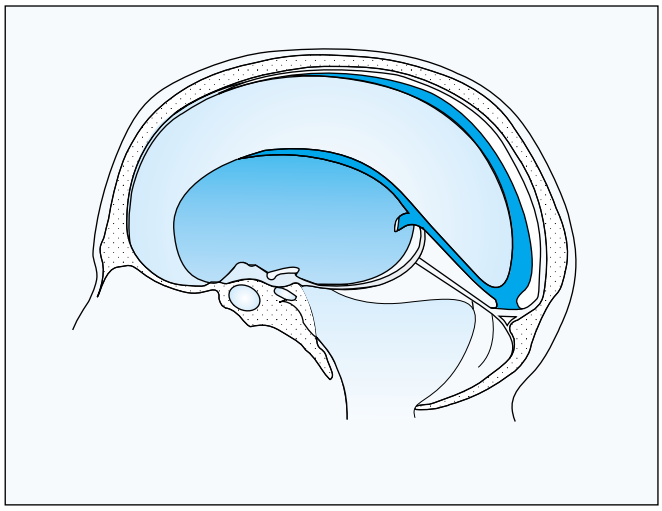


Figure 5.4 Extracerebral haemorrhage may be due to laceration of a venous sinus.

Other conditions

A sternomastoid "tumour" is not noticeable at birth but usually presents after the first week. A firm mass, 1–2 cm in diameter, is usually found in the middle or lower third of the sternomastoid muscle but it may be anywhere along its length. It will disappear within a year and usually the infant will then be perfectly normal. The mother is taught by a physiotherapist to move the infant's head passively through the whole range of normal movements daily until the lesion resolves. Without treatment about 10% of the infants develop torticollis in the second year of life and a further 10% when they are over five years old.

Subconjunctival haemorrhages, like petechiae on the head and neck, are common and unimportant. But if there are petechiae on the trunk further investigations are indicated.

Subcutaneous fat necrosis produces a firm subcutaneous area at the site of pressure, especially of the obstetric forceps. It may be red and tender. No treatment is needed. If the site is unusual and the swelling not noticed until a few days after birth, the area of necrosis may be confused with a pyogenic abscess.

Pressure on the facial nerve before birth or from obstetric forceps may cause a transient palsy which lasts up to two weeks. Care of the exposed cornea is essential.

Figure 5.5 Facial palsy.

6 Some congenital abnormalities

Myelomeningocoele

A myelomeningocoele is a flat or raised neural plaque partly devoid of skin in the midline over the spine due to abnormal development of the spinal cord and associated deficiency of the dorsal laminae and spines of the vertebrae. It is usually found in the lumbar region. The absence of the various coverings that normally protect the cord allows meningitis to occur easily. If there are no active movements in the legs and the anus is patulous, the infant will probably be incontinent of urine and faeces for life and never be able to walk unaided. Thoracic lesions and kyphosis are signs of poor prognosis.

Infants with a good prognosis need urgent treatment, so all affected infants should either be seen by a consultant paediatrician without delay or sent to a special centre, where selection for surgery can be made. About 30% of the infants have surgery as a result of this policy. During the first operation the lesion on the back is covered by skin.

Most of these infants develop progressive hydrocephalus later and those considered suitable for surgery require insertion of a catheter with a valve from a cerebral ventricle to the peritoneal cavity to reduce the cerebrospinal fluid pressure.

Hydrocephalus can be detected by ultrasound examination of the brain. Serial measurements show whether ventricular size is increasing rapidly. In addition, progressive hydrocephalus is confirmed by measuring the circumference of the head at its largest circumference (occipitofrontal) every three days with a disposable paper tape measure, plotting these values on a growth chart, and showing that the head is growing faster than normal.

Raised concentrations of α-fetoprotein are found in the amniotic fluid when the fetus has an open myelomeningocoele or anencephaly. In anencephaly there is absence of the cranial vault and most of the brain.

A routine anomaly scan at 18–24 weeks of gestation will detect most neural tube defects and the parents may decide that the pregnancy should be terminated in view of the poor prognosis. Anencephaly is a lethal condition but some infants survive for a few hours or days after birth.

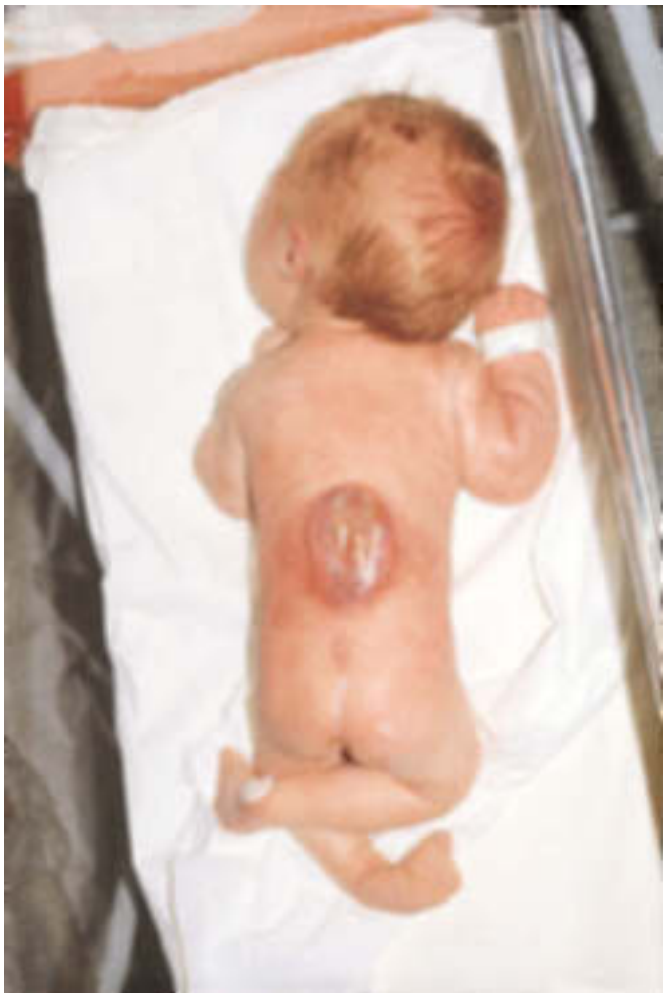


Figure 6.1 Myelomeningocoele.

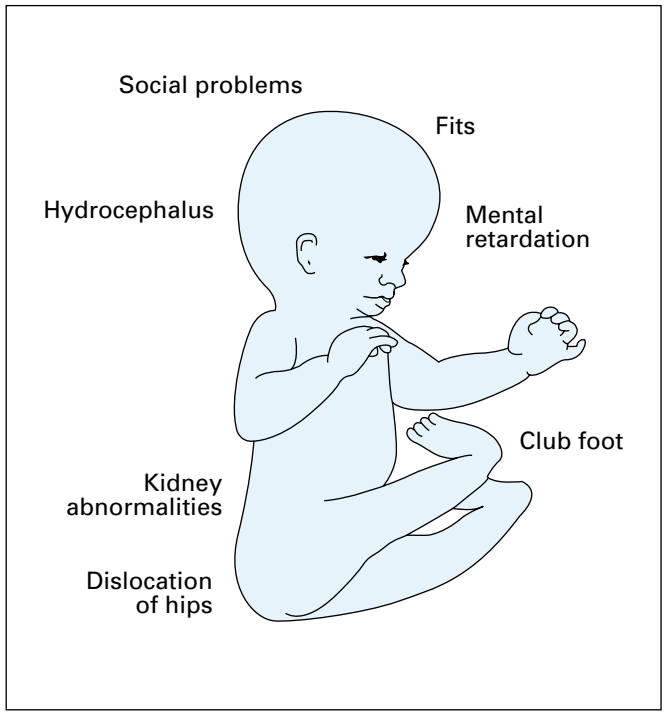


Figure 6.2 Associated features of myelomeningocoele.

Microcephaly

The signs of microcephaly are a small head and forehead that is particularly small in relation to the face. The diagnosis is confirmed by showing a small head circumference in relation to the baby' weight and gestational age. It is usually associated with mental impairment. Other congential abnormalities may be present. Evidence of an intrauterine infection such as toxoplasmosis should be sought and the parents should receive accurate genetic counselling. The diagnosis may be difficult at birth but the increasing discrepancy between the head circumference centile and the weight and length centiles with developmental delay clarifies the diagnosis.

Cleft lip and palate

Mothers are often severely disturbed by the appearance of these infants and may be reassured by seeing photographs of similar patients before and after their repair operations. Cleft lip and cleft palate are often associated. A cleft lip, caused by

failure of the maxillary process growing towards the midline to fuse with the premaxilla, may be unilateral or bilateral. Minor degrees of cleft palate may be easily missed if the posterior part of the palate is not seen and palpated. Most of these infants feed normally from the breast or bottle. If there are feeding difficulties, a special teat, a large normal teat, a special spade-like spoon, or an ordinary spoon may be tried. The lip is usually repaired at three months and the palate by the age of a year. The value of an obturator before operative closure of the palate is controversial. If an obturator is needed, it should be made and fitted within 24 hours of birth. Despite excellent operative results these children are prone to recurrent otitis media and problems with speech development.

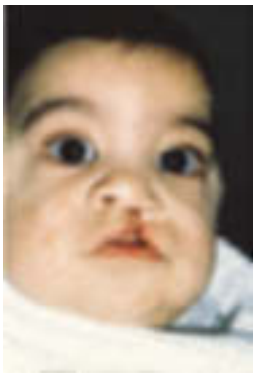


Figure 6.3 Cleft lip and palate.

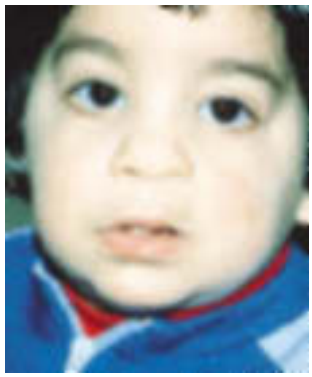


Figure 6.4 Same patient after operation.

Umbilical hernia

An umbilical hernia, usually containing omentum and gut, is most common in African infants or West Indians of African descent. About 30% of preterm infants who have received mechanical ventilation have an umbilical hernia. No treatment is needed, as the hernia usually disappears spontaneously by the age of three years, although in West Indian infants it may take a further three years.

In contrast to an umbilical hernia the sac of an omphalocoele is covered by peritoneum but incompletely by skin. An omphalocoele is a hernia into the base of the umbilical cord and contains gut and sometimes solid organs like the liver. Immediate transfer to a surgical unit is needed.

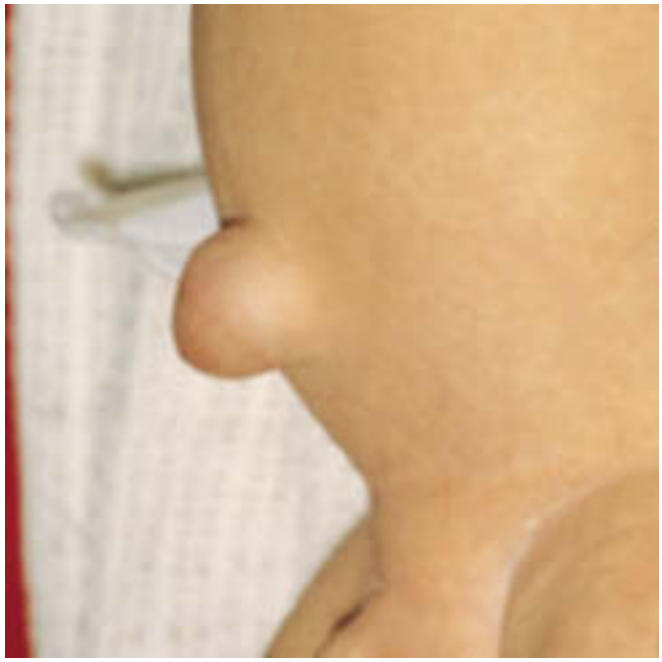


Figure 6.5 Umbilical hernia.

Oesophageal atresia

Oesophageal atresia will be suspected in any newborn baby who has a continual accumulation of frothy secretions in the mouth with drooling, sometimes with cyanotic attacks.

The diagnosis of oesophageal atresia is confirmed by attempting to pass a tube down the oesophagus. The tube should have a relatively wide lumen (FG 10), must be stiff enough to prevent coiling in the upper oesophageal pouch, and should have a radio-opaque line so that the position can be checked by a chest radiograph. The tube should be aspirated every few minutes to keep the upper pouch clear until the infant reaches a specialised surgical unit.

Oesophageal atresia will be suspected antenatally if there is polyhydramnios and difficulty in detecting a normal stomach bubble at the routine 18–24 weeks anomaly scan.

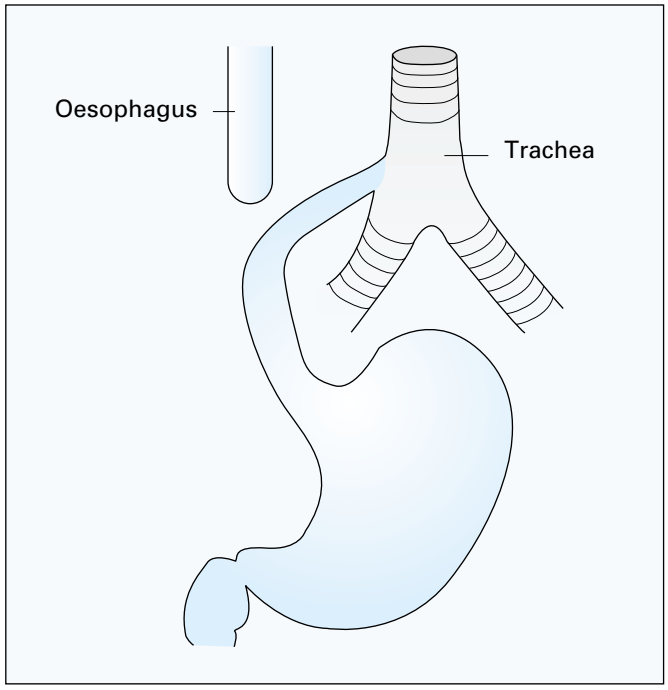


Figure 6.6 Oesophageal atresia.

Multiple abnormalities

Infants with multiple abnormalities should be examined and investigated by a paediatrician without delay to ensure correct management of the infant and to provide the information needed for genetic counselling. Some of the infants may have a recognisable syndrome and the investigations may show a chromosomal defect or evidence of intrauterine infection.

The most common problem is the infant with possible Down's syndrome (mongolism). Although a scoring system for a large number of clinical features has been used it the past, any infant with suspected Down's syndrome should have a blood chromosome analysis performed.

The most useful features suggesting the diagnosis are the facial appearance with palpebral fissures that slant upwards and outwards, prominent epicanthic folds, baggy cheeks, and white spots on the iris. A flat occiput, poor limb tone, and single palmar crease are suggestive but not specific signs. About 2% of normal infants have a single palmar crease in one hand and prominent epicanthic folds are even more common. If the infant has congenital heart disease, it may be detected by performing an ultrasound study, as a definite murmur may not be present until several weeks after birth. All infants with Down's syndrome have developmental delay, but it may be possible to reduce maternal depression and increase the infant's drive and ability to learn by introducing the parents to an occupational therapist in the neonatal period.

Figure 6.7 Down's syndrome.

7 Routine examination of the newborn

Immediately after birth all infants should be examined for the presence of gross congenital abnormalities or evidence of birth trauma. Later, preferably on the morning after delivery, every infant should be examined again in detail. The obstetric notes should be checked to determine whether the infant has been at special risk – for example, from maternal rubella or a difficult delivery. A systematic approach should be used so that abnormalities are not missed. The infant must be completely undressed and in a good light. The mother should be present during the examination so that the results of the examination can be discussed with her and so that she can voice her anxieties.

Skin

Diffuse capillary naevi on the face, eyelids, or occiput are common and resolve within a few months.

The "strawberry mark" starts as a tiny red spot and grows rapidly for several weeks until it has a raised red appearance with small white areas, suggesting the seeds of a strawberry. Such marks are common in preterm babies. They may occur anywhere on the body but cause no symptoms, except on the eyelids, where they may prevent easy opening of the eyes and need treatment. Strawberry naevi grow, often rapidly, for 3–9 months, but at least 90% resolve spontaneously, either completely or partially. Resolution usually begins at 6–12 months and is complete in half the children by the age of five and in 70% by the age of seven years. In 80% of cases these naevi resolve completely without trace.

The port wine stain is not raised and may be extensive. It does not resolve, but the skin texture remains normal. When the naevus occurs in the distribution of the trigeminal nerve, there may be an associated intracranial vascular anomaly.

Neonatal erythema (erythema "toxicum") consists of blotchy ill defined areas of bright erythema surrounding white or yellow wheals which may resemble septic spots. It usually appears on the second day of life and in most infants clears within 48 hours. The lesions contain many eosinophils and have no pathological importance. Neonatal erythema is more common in full term infants. By ringing individual lesions with a skin pencil they can be shown to disappear in a few hours, to be replaced by others elsewhere. This contrasts with septic lesions, which appear later and do not resolve so quickly.

Mongolian blue spots are patchy accumulations of pigment, especially over the buttocks and lower back in infants of races with pigmented skins. They are common in babies of African or Mongolian descent, but also occur in Italian and Greek babies. They may be mistaken for bruises and a wrong diagnosis of non-accidental injury made. They become less obvious as the skin darkens.

A midline pit over the spine is most commonly found over the coccyx, where it does not usually communicate with the spinal canal. A midline pit anywhere else along the spine may be connected with an underlying sinus, which may communicate with the spinal canal and requires excision to prevent the entry of bacteria and meningitis.

Head and neck

An infant with the Pierre Robin syndrome has a small lower jaw, glossoptosis, and cleft palate. The infant must be nursed prone to prevent his tongue falling backwards and occluding

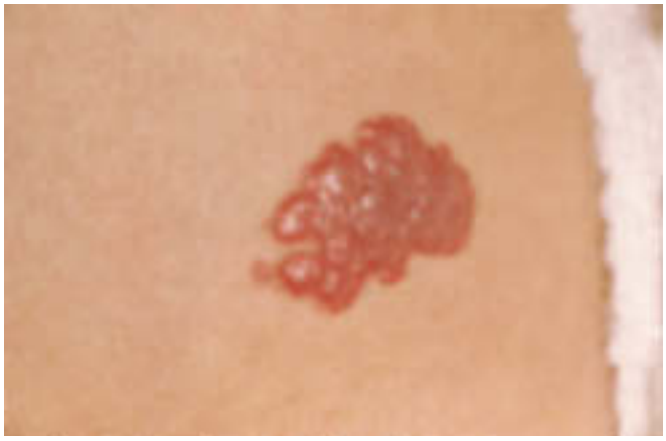


Figure 7.1 Strawberry naevus.

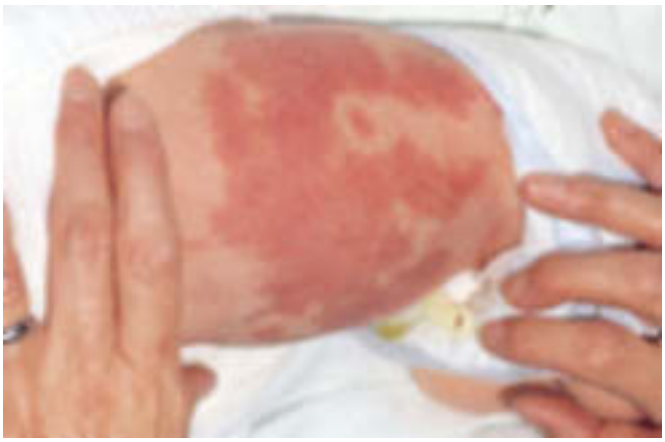


Figure 7.2 Port wine stain.

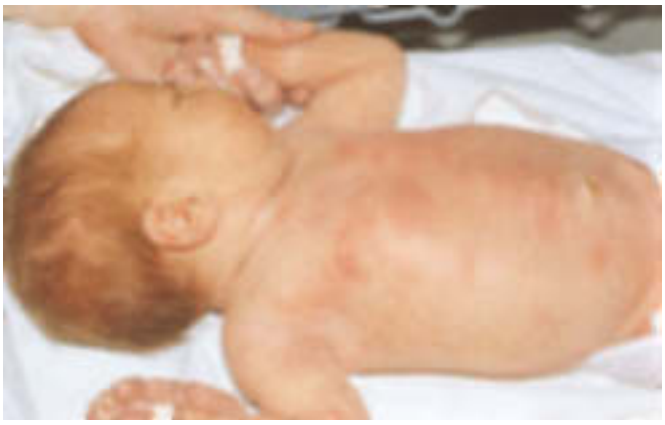


Figure 7.3 Neonatal erythema.

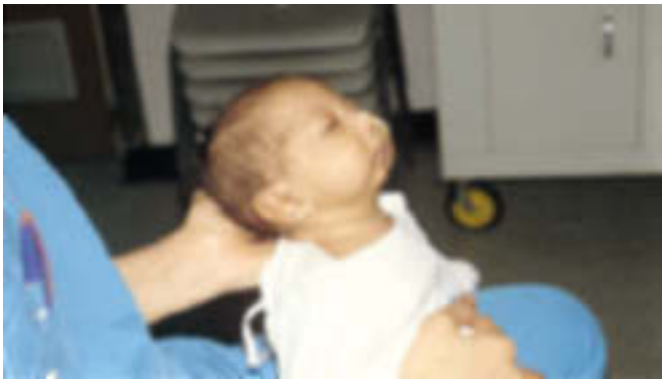


Figure 7.4 Pierre Robin syndrome.

the airway. A small jaw may occur alone or with other abnormalities. As the child grows the small size of the lower jaw becomes less obvious.

Nodules of epithelial cells resembling pearls (Epstein's pearls) just lateral to the midline on the hard palate are a normal finding. A cluster of these pearls is present at the junction of the hard and soft palates.

Heart murmurs

In the first two days of life one of the most common problems is an infant who is feeding normally but is found to have a short systolic murmur during a routine examination. The murmur is usually due to a patent ductus arteriosus and has usually disappeared by the time the infant is examined again at seven days. An electrocardiogram (ECG) or chest radiograph is unnecessary.

If a long murmur is first heard around the eighth day of life, a chest radiograph, ECG, and echocardiogram should be performed and the infant seen by a paediatrician. Most of these murmurs are due to a ventricular septal defect or mild pulmonary stenosis with no symptoms. Most of the ventricular septal defects (VSD) close spontaneously before the infant reaches the age of five years.

The normal respiratory rate at rest is less than 60/min and there should be no recession of the chest wall or below the mandible. If there is any murmur and the infant is feeding poorly or has a respiratory rate faster than 60/min at rest, a chest radiograph and ECG should be performed and the baby seen by a paediatrician urgently. These symptoms and signs indicate congestive heart failure usually due to multiple heart defects.

Cyanosis of the hands and feet is common in newborn infants and has no importance if the tongue is of normal colour and the infant is feeding normally.

Abdomen

The liver edge is normally 1–2 cm below the costal margin in the midclavicular line. In a thin, relaxed infant the kidneys may be palpable bimanually.

The inguinal areas are inspected for the presence of a hernia. Absence of the femoral pulses suggests that coarctation of the aorta is present. The blood pressure is measured in all four limbs and the advice of a cardiologist is obtained urgently.

Failure to pass urine in the first 36 hours or a poor urinary stream suggests posterior urethral valves in a boy. An enlarged bladder may be palpable. Diagnosis is confirmed by an ultrasound examination, which may show hydronephrosis, and cystourethrogram.

Patency of the anus is confirmed by inspection.

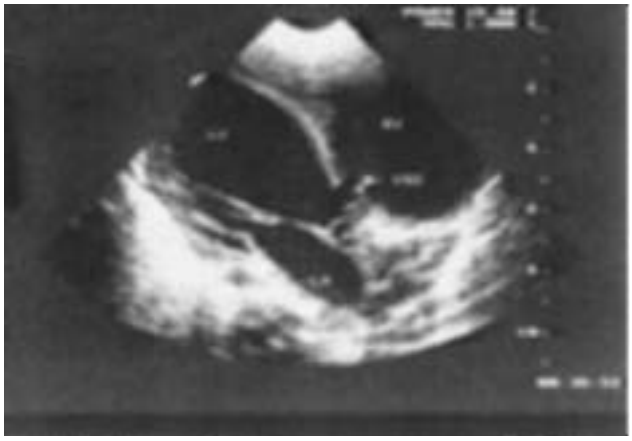
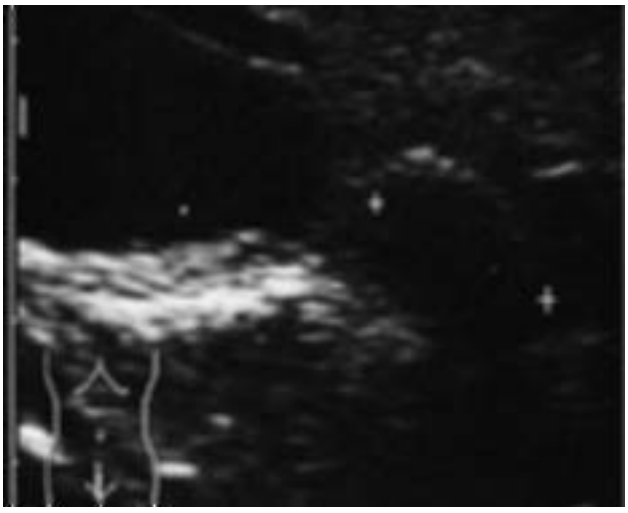
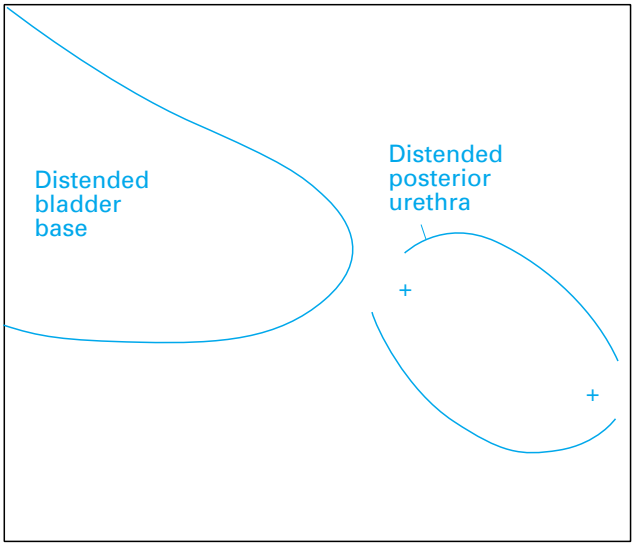


Figure 7.5 Echocardiogram showing ventricular septal defect (VSD).



(a)



(b)

Figure 7.6 Ultrasound scan (a) and diagram (b) showing distended bladder and distended posterior urethra due to urethral valves.

Genitalia

The foreskin is attached to the penis at birth and no attempt should be made to retract it.

A hooded prepuce suggests hypospadias with the urethral orifice at the base of the glans penis. If the urethral meatus is adequate, no immediate treatment is needed, but the infant must not be circumcised and should be referred to the paediatric urologist's next clinic.

If the urethral orifice is nearer the perineum, the adrenogenital syndrome should be considered. In the adrenogenital syndrome cortisol secretion fails and the adrenals produce excessive androgen. Girls become virilised, with enlargement of the clitoris and fusion of the labia, and may be mistaken for boys (see p. 52). In boys the genitalia are normal.

Intersex is a less common cause of ambiguity of the external genitalia and a paediatrician should be consulted without delay. An accurate diagnosis is an emotional and social emergency. The parents should be advised that the baby should not be named until the results of chromosome studies are available.

Small hydrocoeles usually disappear spontaneously during the first month but an associated inguinal hernia should be sought.

Poor development of the scrotum suggests that an undescended testis is present. Undescended testis is especially common in preterm infants, in whom the testes usually descend during the first three months after birth. An undescended testis is present in 30% of preterm and 3% of term infants at birth. An infant with an undescended testis should be seen again at the age of three months. At that time 5% of preterm infants and 1% of term babies will still have an undescended testis and they should be referred for surgery, which is performed at about the age of two years. If the baby develops an associated hernia, the operation will be needed as an emergency.

As the cremasteric reflex is usually absent at birth, the testis cannot be retractile. If there is any doubt about whether a testis is descended the examiner should (*a*) palpate the pubic tubercle with one hand, (*b*) hold the testis between the thumb and forefinger of the other hand and gently draw it down to its fullest extent then (*c*) measure the distance from the pubic tubercle to the centre of the firm globular testis. At term the testis, if fully descended, lies 4–7 cm from the tubercle. If the distance is under 4 cm the testis has not completely descended. In preterm babies, who are more likely to have undescended testes and whose testes are smaller, 2·5 cm has been arbitrarily chosen to divide descent from non-descent.

Spontaneous primary descent of the testes rarely occurs after the age of four months and never after one year. After the neonatal period an active cremasteric reflex can easily pull the testis out of the scrotum, especially if the examiner's hands are cold. The mother will often have noted whether the testes are both in the scrotum after a hot bath, and a retractile testis will descend into the scrotum when the thigh and knee are maximally flexed on that side.

In girls a small amount of vaginal bleeding is common, usually five to seven days after birth, and follows withdrawal of maternal or placental oestrogens, which are transmitted to the fetus before birth. White vaginal discharge or prolapse of the vaginal mucosa is normal.

In either sex physiological enlargement of the breast may occur towards the end of the first week. This enlargement, which may be unilateral, resolves within a few weeks, but if there is local redness, a breast abscess should be suspected.

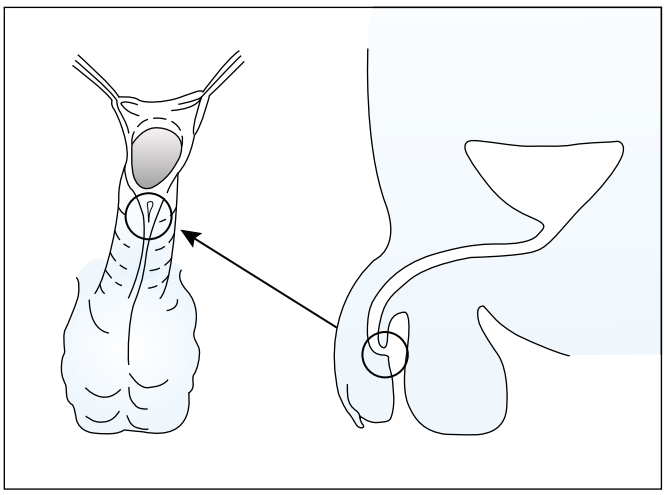


Figure 7.7 Hypospadias.

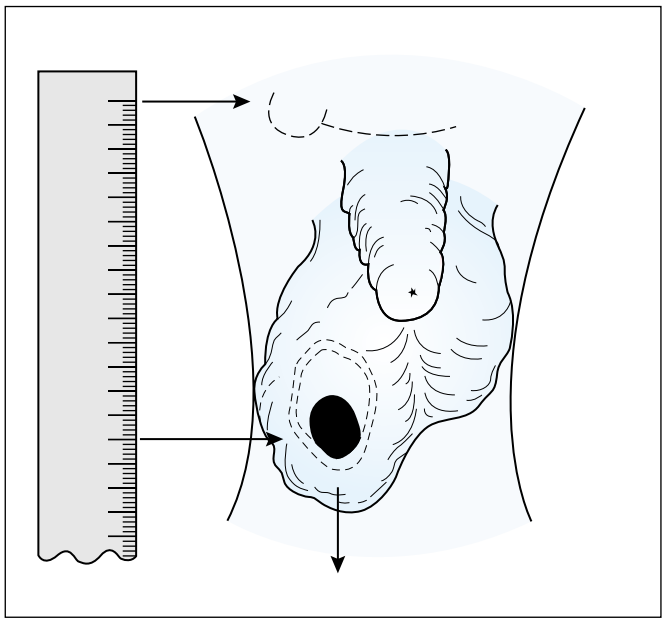


Figure 7.8 Measurement of testicular descent.

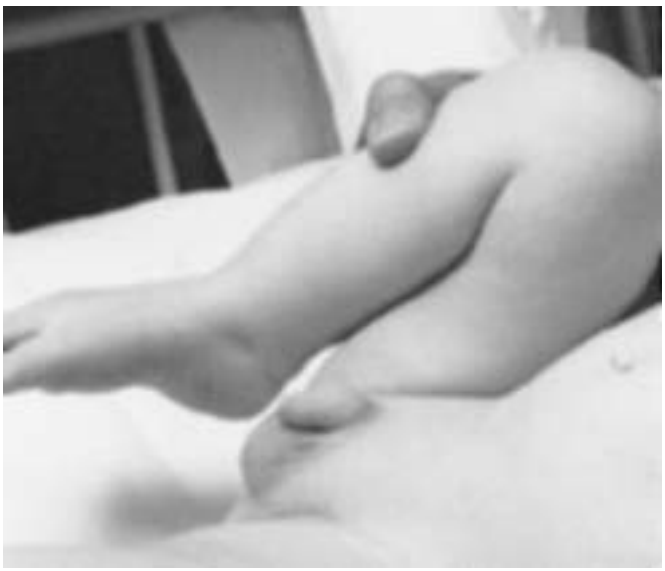


Figure 7.9 Flexion of thigh and knee to determine testicular descent.

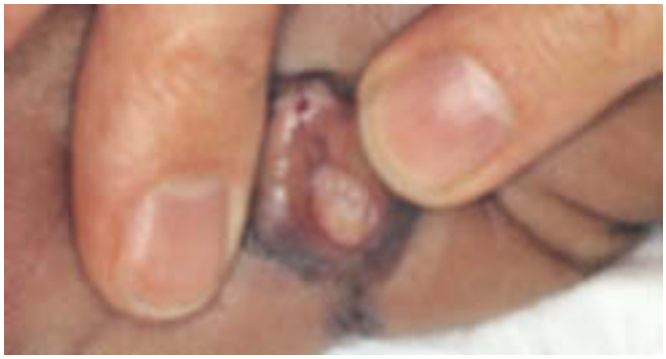


Figure 7.10 Prolapse of the vaginal mucosa.

Circumcision

In 95% of babies the foreskin and glans of the penis are still united at birth. It has been found that the foreskin can be retracted by the age of one year in about half the babies and by the age of three years in nine out of ten. But no attempt should be made to retract the foreskin until the baby is about four years old. Attempts to retract the foreskin earlier are likely to injure the mucosa, causing bleeding followed by adhesions, and circumcision may later become necessary. Mothers often request circumcision to be carried out because the prepucial orifice appears small. In most cases the adequacy of the orifice can be shown by gently stretching the foreskin *distally* and no attempt should be made to retract the foreskin. Before the age of four years the only medical indications for circumcision are recurrent purulent balanitis and ballooning of the foreskin at the beginning of micturition.

Most circumcisions are performed as a religious ritual, in Jewish families on the eighth day of life and in Muslim boys between the ages of three and 15 years.

Club foot and extra digits

Muscular imbalance due to the posture of the infant's feet *in utero* is the commonest cause of club foot. In postural club foot it should be possible to dorsiflex the foot fully and to obtain passive inversion to 90°. The mother should be taught to manipulate the foot through the whole range of movements after each feed for several days after birth, although the shape usually reverts to normal within a few weeks even without treatment. In contrast, in structural club foot the range of passive movements is restricted and orthopaedic advice on strapping, manipulation, or serial plasters is needed within 24 hours of birth.

To detect extra digits, the digits should be counted with the infant's palms open, or an extra thumb may be missed. Polydactyly requires the advice of a plastic surgeon to determine which digit should be removed at the age of 3–4 years for the best functional results. Extra fingers and toes are often familial and vary from an apparently normal digit to a skin tag. The latter can be tied off with a sterile silk thread and will separate by aseptic necrosis.

Central nervous system and eyes

To assess the central nervous system the alertness of the infant and the symmetry of spontaneous movements should be noted. The tension of the anterior fontanelle and the width of the fontanelle should be palpated while the infant is at rest and the head circumference should be measured. A more detailed

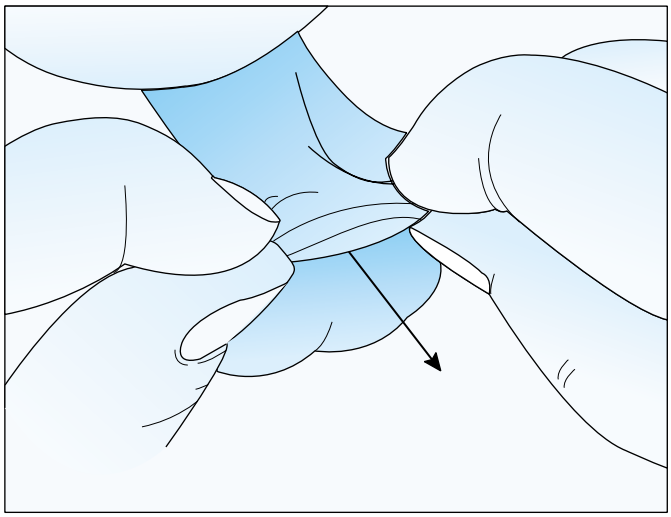


Figure 7.11 Stretching foreskin distally.

(a)

Figure 7.12 (a) Postural.

(b)

Figure 7.12 (b) Structural.

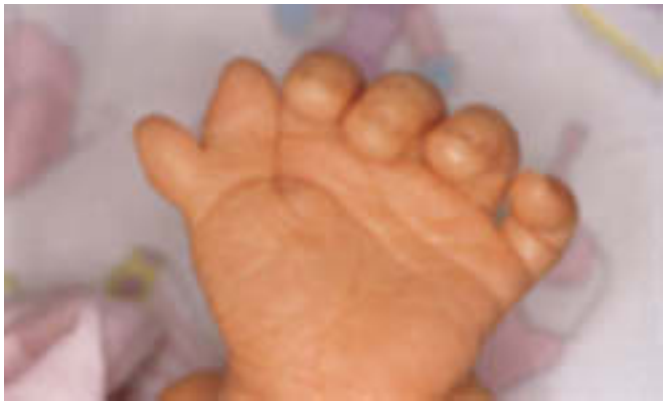


Figure 7.13 Extra thumb.

Figure 7.14 Eyes open when the infant is held over the mother's shoulder.

neurological examination is not required unless there is a special indication.

Babies will open their eyes when being fed or when held upright over their mother's shoulder. Babies will follow the movements of an examiner's face provided that the distance is 50 cm or more. The pupils of every baby should be examined with an ophthalmoscope at a distance of 50 cm. A bright red glow is seen, which is a reflection of light from the back of the retina. Absence of this "red reflex" is found in congenital cataract, which produces a dull grey appearance.

Watering eye

If there is persistent watering of the eye with clear fluid, the tear duct is probably blocked. No action is required until at least the age of one year, when the infant can be referred to an ophthalmic surgeon.

Rarely, the tear duct is probed, but most ophthalmic surgeons prefer to take no action because the condition resolves spontaneously in virtually all infants and probing may induce fibrosis of the tear duct. If there is repeated or persistent purulent discharge, the possibility of chlamydia infection should be considered and arrangements made with the laboratory for specimens to be taken.

Taking blood

Capillary blood taken from infants aged up to two years can be used for a variety of tests. The commonest is the Guthrie test, but capillary blood may also be used for blood glucose testing, haemoglobin estimations, and most biochemical tests.

The best site for taking capillary blood from an infant aged under six months is the heel. In older infants the thumb is a better site. The heel must be warm. If it is cold, the infant's foot should be dipped into hand hot water (40°C) for five minutes and dried thoroughly. The examiner then holds the infant's foot by encircling the ball of the heel with thumb and forefinger. The site selected for the heel prick must be on the side of the heel; if the ball or back of the heel is used a painful ulcer may form.

The site is wiped with isopropyl alcohol and allowed to dry. A spring loaded lancet introducer is used.

The initial drop of blood should be wiped away with a dry cotton swab and succeeding drops allowed to drip into the container, onto the Guthrie test card or onto the end of a glucometer strip. To milk blood into the heel the examiner should squeeze and release the fingers around the infant's calf and keep the heel below the rest of the leg. Up to 2 ml of blood can be obtained. When the required volume has been obtained, the heel is wiped and the wound pressed with a clean cotton wool ball. A small plaster should then be applied.

Phenylketonuria and hypothyroidism

The Guthrie test detects high blood concentrations of phenylalanine, which indicate that the infant probably has phenylketonuria. Drops of blood are taken by heel prick after six days' breast or bottle feeding (usually the seventh day of life) and placed on the special absorbent card provided. The blood is usually taken by the midwife if the infant is at home on the seventh day. If the test is positive, the infant should be admitted to hospital to confirm the diagnosis, as he or she may need a special low phenylalanine diet to prevent brain damage.

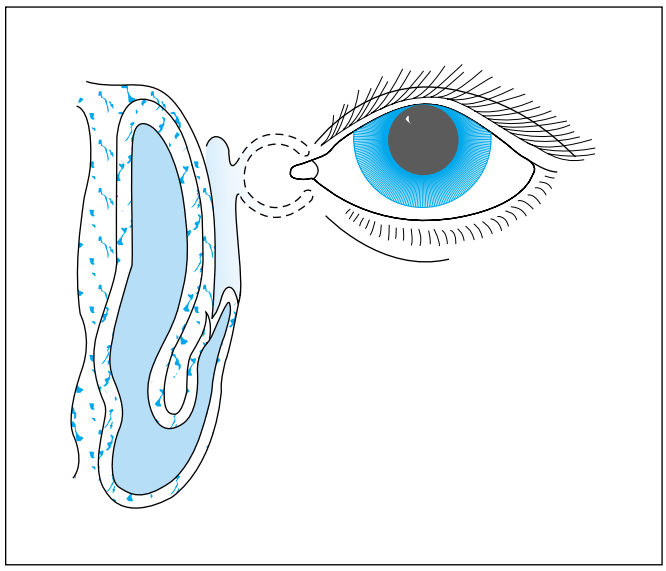


Figure 7.15 Tear duct.

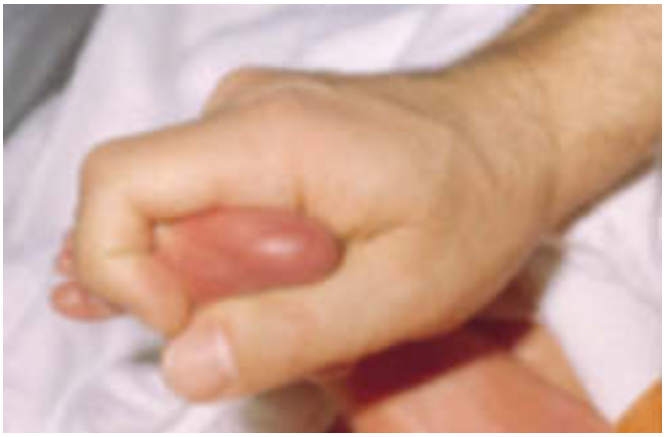


Figure 7.16 Holding foot for taking blood.

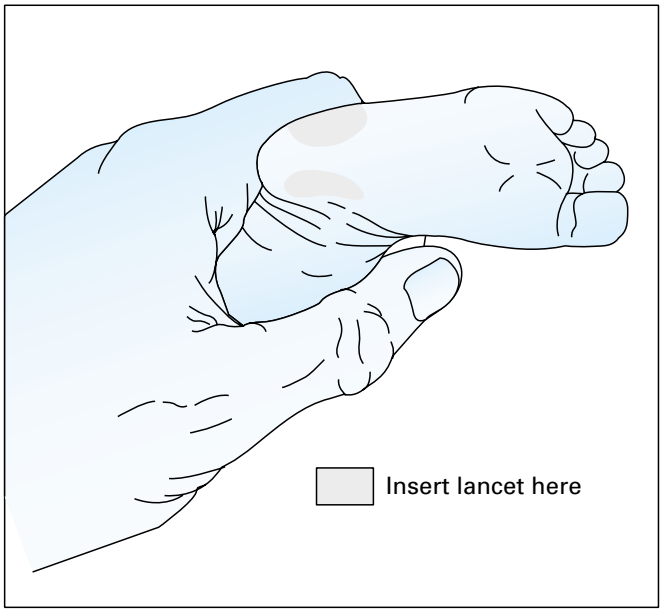


Figure 7.17 Sites for insertion of lancet.

Figure 7.18 Screening blood card.

A capillary tube of blood can be used (for the Scriver test) instead of the dried blood spot.

Some laboratories use the plasma thyroid stimulating hormone and others the plasma thyroxine value as the screening test for hypothyroidism. The screening blood card is used for both the Guthrie test and the test for hypothyroidism.

Talking to parents

Whenever possible both parents should be seen together. If an abnormality is found during the routine examination some indication must be given to the mother so that arrangements can be made to see both parents. Particularly if the problem is likely to be long term, the doctor who will continue the long term care should speak to the parents initially. Ideally, a nurse should be present as well since the parents will probably ask all the questions again as soon as the doctor has left.

The diagnosis should be explained in terms that the parents can understand and the positive aspects should be emphasised. For example, if the probable diagnosis is ventricular septal defect, it is not necessary to explain that there is a remote possibility of multiple cardiac defects. After a difficult forceps delivery it is better to say that most babies develop normally rather than say they may be slightly backward. An unnecessarily pessimistic prognosis based on out of date information may alter a mother's attitude towards her baby and impair her attachment to him or her.

In discussing management the help and support that will be given by social workers or physiotherapists should be emphasised and the parents advised to tell their relations and friends the diagnosis rather than hide it.

A doctor with only limited experience of a particular problem should tell the parents that he or she needs to seek further advice rather than guess at the answer. Parents often cling to the first opinion and may find it difficult to accept a more experienced view later. A statement of the facts – for example, that there is an abnormality of the spine – may be given with the assurance that a more experienced doctor will discuss them later.

Figure 7.19 Talking to parents.

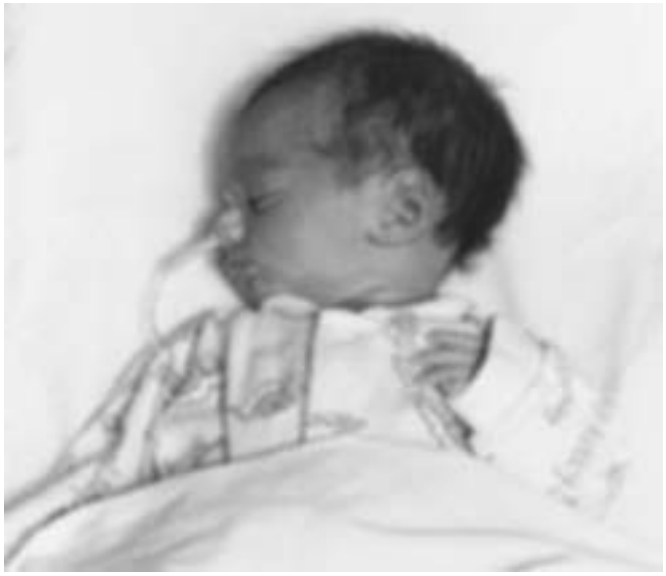


Figure 7.20 Position for sleeping.

Useful advice for new mothers

Sleeping position of infants

It is recommended that babies should be placed on their back to sleep as this has been shown to reduce the risk of cot death.

Healthy babies placed on their backs are not more likely to choke if they vomit. The mattress used in a baby's cot should be firm, flat, clean, and covered in waterproof material like PVC. Pillows should not be used. To prevent the baby wriggling down under the covers, a baby's feet should be placed at the foot of the cot or pram.

By 6–7 months of age, many babies will roll over onto their front during sleep. There is no need for concern about this, as the risk of cot death is markedly reduced by this age. There are, however, certain circumstances in which babies should be nursed on their side or prone (on their front). These include some babies in neonatal units, babies with severe gastro-oesophageal reflux, babies receiving treatment in splints for unstable hips, and babies with Pierre Robin syndrome.

Cellular blankets are ideal and duvets or loose bedding should be avoided. An infant should not sleep on a sofa or bed with a parent. It has been recommended that the infant should sleep in the same room as the parent for the first six months of life.

Avoid smoking near babies

Smoking in pregnancy by the mother will reduce the birthweight of the infant by approximately 200 g at term and increases the risk of cot death. Babies exposed to cigarette smoke after birth are also more likely to suffer wheezing and respiratory infections. It is best if everyone, especially parents, refrains from smoking in the same house or room as a baby.

Temperature

Overheating a baby may increase the risk of cot death. A room temperature of about 18°C (65°F) is comfortable for babies as well as adults.

Car seats

Whenever a baby is taken in a car, he should always be strapped into a baby seat or rear facing infant carrier. It is not safe to nurse a baby on the lap in a car, even if the adult is wearing a seat belt.

Cot deaths

About one baby in every 2000 born alive dies suddenly and unexpectedly between the ages of one week and two years. Typically, an apparently healthy baby (or, occasionally, one with only minor symptoms) is put in a cot to rest and some time later is found dead. Although an infant may be face down in the cot with the bedclothes over him or her, suffocation should not be assumed. Sometimes vomit, which may be blood tinged, is found around the mouth or on the bedding, but regurgitation usually occurs after death and is not the cause of death.

In some cases necropsy discloses an unsuspected congenital abnormality or rapidly fatal infection. But usually there is no evidence of severe disease, though there might be slight reddening of the tracheal mucosa, which in other babies normally resolves spontaneously.

Parents often blame themselves and may worry that the infant suffocated as a result of neglect. They should be told that their feelings of guilt are a natural reaction and the doctor should explain to them, and to anyone who was looking after the baby when death occurred, that cot death is a well recognised but ill understood condition and that no one is to blame for the infant's death.

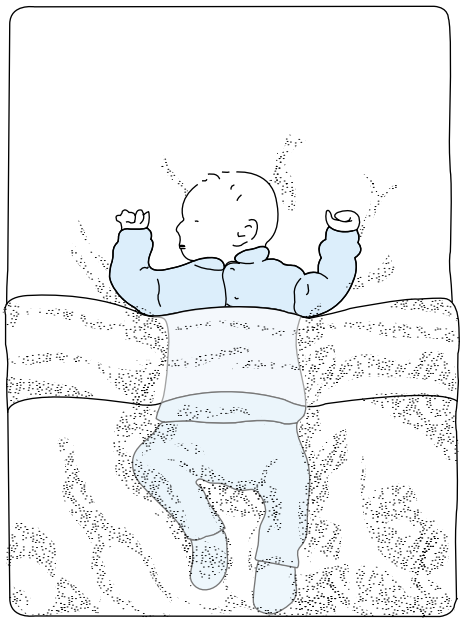


Figure 7.21 Correct sleeping position, with feet at the bottom of the cot.

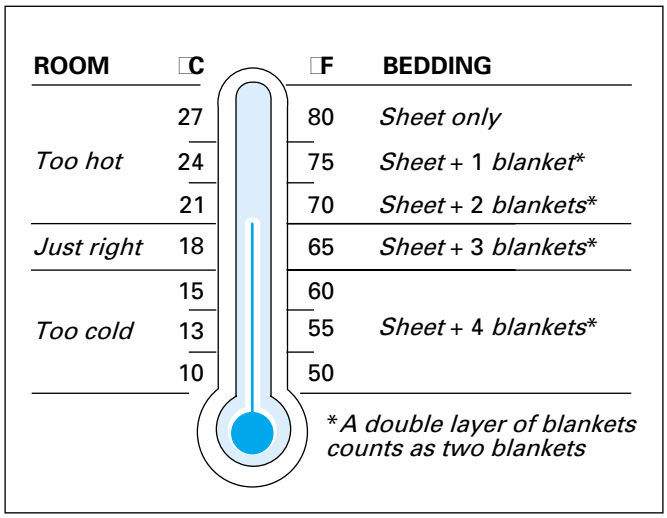


Figure 7.22 Amount of bedding required for a range of room temperatures to keep the infant's temperature at the optimum level.

Box 7.1 Foundation for the Study of Infant Deaths (FSID)

Artillery House, 11–19 Artillery Row, London SW1 1RT
General enquiries telephone: 020 7222 8003
24-hour helpline: 020 7233 2090

Box 7.2 Prevention of sudden infant death

Although sudden infant death is very rare, it is important to follow the latest guidelines for prevention. These are:

- Put your baby to sleep on his or her back
- Do not smoke; keep your baby out of smoky atmospheres
- Do not let your baby get too warm
- Put your baby at the foot of the cot so that he or she cannot wriggle under the covers
- Contact your doctor if you think your baby is unwell

Birth details

Please use ball point pen and press firmly

Date of birth_____ / _____ / _____
Time of birth_____
Weeks of pregnancy_____
Method of delivery_____
Place of birth _____
Problems during pregnancy or birth:
(1) _____
(2) _____
(3) _____
(4) _____

Meconium passed within 24 hr	Yes/No
In special care baby unit	Yes/No
Any neonatal contraindication to immunisation	Yes/No
Risk factors for congenital dislocation of hip	Yes/No
Risk factors for hearing loss	Yes/No/Not Known
Feeding at discharge	Breast/bottle/both
Blood tests done	
Phenylketonuria	Yes/No Norm/Abn
Thyroid test	Yes/No Norm/Abn
Haemoglobinopathies	Yes/No Norm/Abn
BCG	Yes/No

Infant Hospital NO._____
Birth Wt. _____ kg Head circ. _____ cm
Apgar at 1 & 5 mins _____
(Tick box if examination done and write in 'comments' if problem)

		Comments
Skull sutures	☐	_____
Skin	☐	_____
Eyes – red reflex	☐	_____
Palate	☐	_____
Heart	☐	_____
Abdomen	☐	_____
Genitalia	☐	_____
Hips	☐	_____
Limbs/spine	☐	_____

Significant abnormality or condition _____
_____
_____
Follow-up hospital appointment Yes/No
Reason/details _____
_____
_____

Birth details

Immunisations

ALL CHILDREN SHOULD RECEIVE IMMUNISATIONS except a very few children who:
1. are suffering from a feverish illness – when the immunisation should be postponed until full recovery.
2. have had a severe reaction to a previous immunisation.
3. have an illness or are taking medicines that interfere with their ability to fight infections.

CHILDREN TAKING ANTIBIOTICS CAN BE IMMUNISED.
Before each immunisation the doctor or nurse will make sure that it is all right to give your child the vaccine.

Further comments on consultation sheets: YES/NO

Please use ball point pen and press firmly.

Age due	Site of injection	Vaccine	Date given	Batch number	Signature in full	Treatment centre code
Dose 1 2 months	RL/LL.... RL/LL....	*D.T.P or *D.T. Polio Hib Meningo C				
Dose 2 3 months	RL/LL.... RL/LL....	*D.T.P or *D.T. Polio Hib Meningo C				
Dose 3 4 months	RL/LL.... RL/LL....	*D.T.P or *D.T. Polio Hib Meningo C				
12 months		Measles, Mumps, Rubella (MMR)				

*Delete one

Immunisations

Figure 7.23

Parents must also be told that because the death is of unknown cause the coroner will have to be told as a matter of routine and that there will be a necropsy. If the parents want to see the infant's body (and they should be asked), the infant is clothed and a doctor or nurse should be present to answer questions and provide support. Parents should be encouraged to see the infant's body as this will help the grieving process. Discussion with the parents will determine who is best to help with their grieving.

The family doctor and health visitor should be informed of the death immediately. The family doctor should explain the results of the necropsy to the parents or arrange for a paediatrician to do this.

Brothers and sisters of the infant who has died will need reassurance that they will not die and are in a safe and loving environment. The concept of the finality of death may not be attained until the age of eight years. Children may be disturbed by the emotional distress of the parents and this may cause behavioural or sleep problems.

Foreign travel with infants

Families are more mobile and are increasingly seeking advice about air travel with young children. Infants are at increased risk of experiencing earache during take-off and landing because the drainage of their Eustachian tubes may be partially blocked with fluid and mucus. Feeding the infant may help by producing a partial Valsalva manoeuvre, but may not be possible because of safety restrictions in the aircraft.

Travel to foreign countries is best avoided until the completion of the routine infant immunisations at four months of age. BCG should also be given if travelling to a country with a high prevalence of tuberculosis.

Antimalarial prophylaxis is essential for young infants and can be safely given to all infants, provided that they do not have neonatal jaundice, if travelling to a country where malaria is endemic. The same drugs as are prescribed for adults can be used in appropriate dosage for infants:

- first month of life – one eighth adult dose
- one month to one year of age – one quarter adult dose.

Boiled or bottled water should be used for drinking water and preparation of infant formula feeds abroad. Parents should be advised not to use commercially produced mineral water with a sodium content more than 20 mg/l, as a high salt intake can be harmful in young infants. Changes in drinking water and diet may cause an alteration in intestinal flora and diarrhoea in young infants. Meticulous attention to hygiene will help to prevent gastroenteritis. Families of infants should be advised to carry sachets of powdered oral infant rehydration mixture when they travel abroad, as it may not be readily available elsewhere.

The skin of young infants is very susceptible to sunburn and must always be protected with liberal applications of high factor sunblock, hats, and sun shields or umbrellas in hot sunny climates.

8 Dislocated and dislocatable hip in the newborn

Congenital dislocated hip and dislocatable hip are probably the most important asymptomatic congenital abnormalities to detect, as early treatment is simple and usually effective. In the first 12 hours of life ten in 1000 infants in Britain have a hip abnormality. When the examination is carried out 24–36 hours after birth the incidence falls to about seven per 1000 births. Formerly, when they were all left untreated in the newborn, the incidence of established dislocated hip was one in 800 children. There may be a family history of the condition; the anomaly is more common in girls and after the extended breech position *in utero*. There is a higher incidence in some countries, such as Italy and the former Yugoslavia.

The best time to examine the infant's hips is between the ages of 12 and 36 hours, as the tendency to provoke regurgitation is less and there is less ligamentous laxity by that time. The examiner's hands should be warm and the infant should be placed on her back with the sheet spread completely flat. She should be fully relaxed and this may be encouraged by putting an empty sterile feeding teat in her mouth if necessary. The gentle abduction test, followed by Barlow's test, should be carried out in all cases. Unnecessary trauma to the delicate hip joint and its capsule must be avoided.

Gentle abduction test

The gentle abduction test will detect hips that are in the dislocated position at rest. Each hip should be examined separately, while the opposite thigh is gently fixed by the examiner's other hand. Both the knee and the hip should be flexed to a right angle and the knee held so that the examiner's thumb is parallel to the medical aspect of the lower thigh, while the middle two fingers lie along the whole length of the lateral aspect of the femur. The tips of the examiner's fingers thus lie over the greater trochanter.

The thigh should be held *lightly* and neither pushed down towards the cot surface nor pulled up towards the examiner's face. It should then be allowed to abduct very gently and slowly by the weight of the infant's leg until abduction is complete. The thigh should never be forcibly abducted and it is unnecessary to obtain abduction beyond 10° above the flat. While abducting the thigh, the examiner may feel or see the head of the femur slip, jerk, or jolt forward into the acetabulum. A temporary interruption in the flow of abduction at a point about midway through abduction precedes the sensation of this abnormal movement of the head of the femur.

If the joint capsule is very lax, the reduction jolt may be missed unless great care is taken.

Barlow's test

The object of Barlow's test is to identify dislocatable hips in which the head of the femur can be gently jolted posteriorly over the posterior lip of the acetabular labrum to lie temporarily out of the acetabulum and those dislocated hips in which the head of the femur can be jolted forwards to lie temporarily in the acetabulum.

Each hip should again be examined separately while the opposite thigh is gently fixed by the examiner's other hand. The infant's hip should be flexed to a right angle and the knee more acutely flexed. The examiner should place a thumb as high as

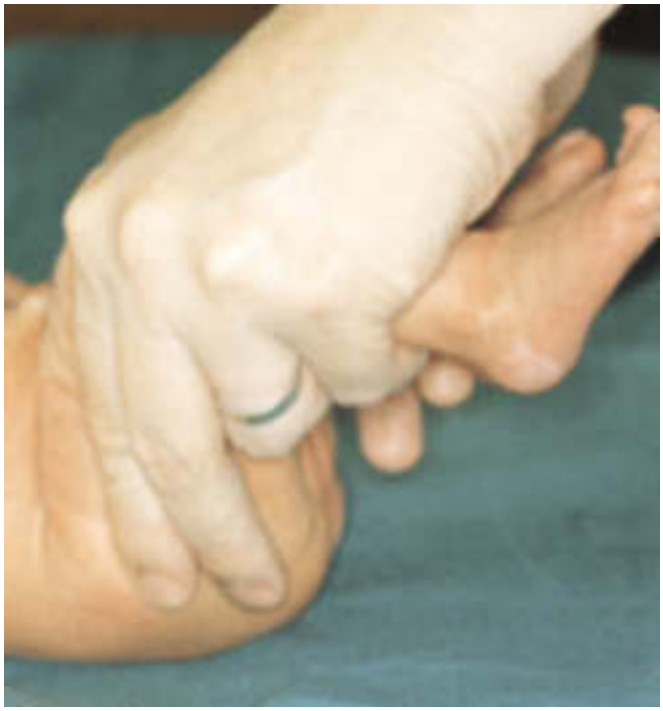


Figure 8.1 Gentle abduction test.

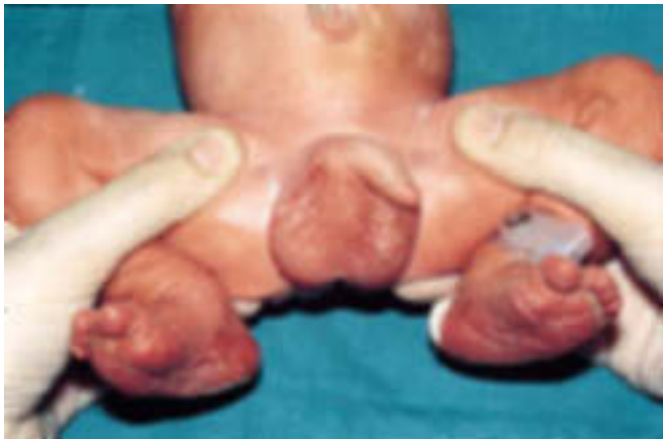


Figure 8.2 Gentle abduction test.

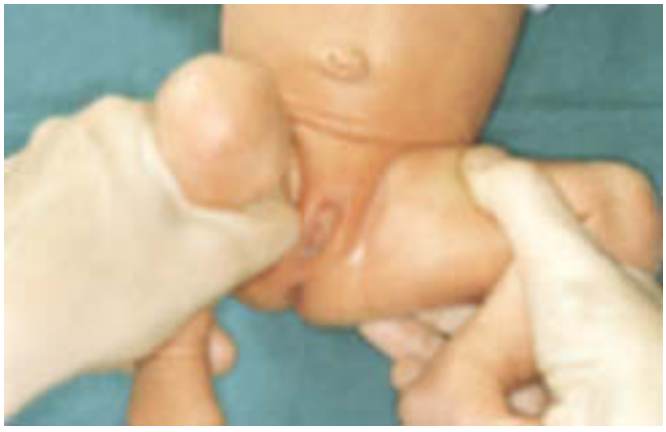


Figure 8.3 Barlow's test.

possible on the medial aspect of the upper femur while the tips of the middle two fingers grasp the greater trochanter laterally.

The thigh is held lightly in a position of only minimal abduction and then an attempt is made to push the femoral head gently posteriorly and slightly superiorly, while at the same time the examiner's hand is internally rotated through not more than 25°. This is followed by reversing the whole movement. No excessive force is used and only a very limited range of movement employed in the test.

The movement of the head of the femur out and in, or in and out, of the acetabulum produces a jolting or jerking movement, *which can be seen and felt by the examiner. It cannot be heard by an individual with average hearing.* The sensation is like that of a gear lever engaging.

Hips that show excessive movement of the head of the femur within the joint, without being actually dislocatable, are classified as normal.

Ligamentous laxity. In at least 10% of infants this examination evokes a noise, click, snap, or grating sensation, but there is no abnormal movement of the femoral head. Such hips should be considered normal and no follow-up is required. Because of confusion about its meaning, the term "clicking hip" should be abandoned.

Examination of the hips should be performed more than 15 hours after birth because few ligamentous clicks will be found.

Ultrasound in screening

If ultrasound screening is undertaken in every infant before leaving hospital, imaging facilities will be required seven days a week. Twenty percent of hips may be considered abnormal and the majority will resolve to normal within four weeks.

Selective screening with ultrasound of infants with a clinical hip abnormality or risk factors for congenital dislocated hip (breech delivery, positive family history, or foot deformity) reduces the screened population and allows treatment options to be delayed and targeted effectively. A recent study showed that this approach has not reduced the prevalence of late cases of congenital dislocated hip and it has been suggested that selective ultrasound screening is dependent upon more vigorous clinical screening and careful selection of risk factors.

Ultrasound in diagnosis and management

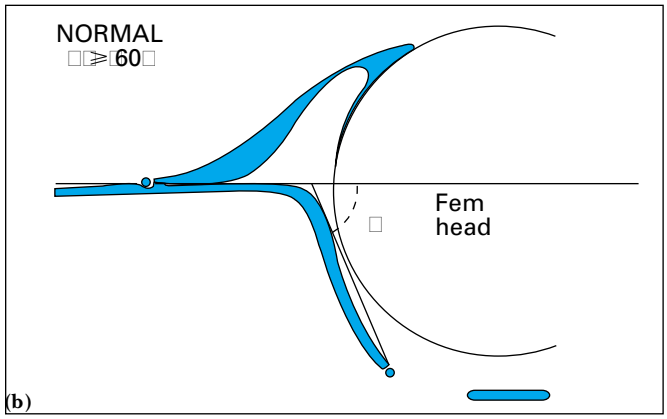
Shallow acetabulae occur with normal hips and delayed ultrasound screening examination to 4–6 weeks in clinically stable hips will allow treatment to be targeted to those hips that require splintage and thus reduce treatment rates without compromising the results of this treatment.

Weekly ultrasound studies can be used to confirm hip relocation and treatment progress while the infant is in a malleable splint or Pavlik harness. Sonographic evidence of continuing femoral head dislocation, despite splintage, allows treatment to be abandoned and thus the risk of avascular necrosis is reduced. The appearance of the ossific nucleus in the cartilaginous femoral head (usually delayed in congenital dislocated hip) is identified sonographically several weeks before it is visualised radiologically.

Recommended screening

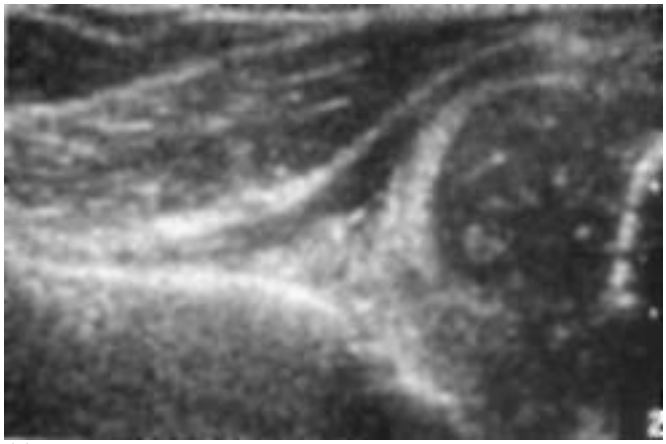
Babies born after a vertex presentation who have normal hips on clinical examination and do not have any risk factors will

(a)

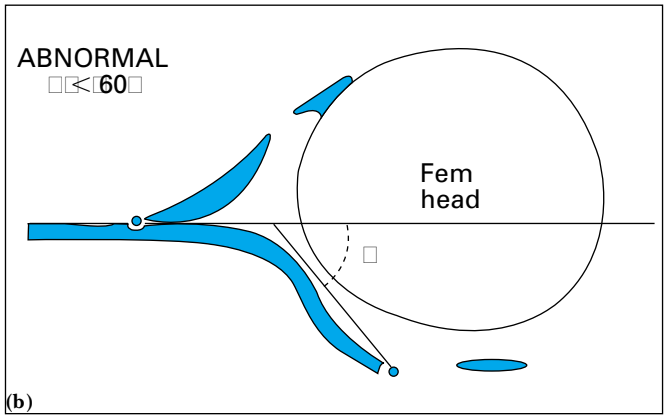


(b)

Figure 8.4 (a) Ultrasound of normal hip **(b)** diagram.



(a)



(b)

Figure 8.5 (a) Ultrasound of abnormal hip **(b)** diagram.

continue to receive clinical screening using the neonatal examination and thereafter by their general practitioner in surveillance clinics.

Babies who are judged to be at risk of congenital dislocated hip because they:

- were born in breech position, or
- have a family history of congenital dislocated hip, or
- have a foot deformity

should have ultrasound at 4–6 weeks of age, in addition to the clinical screening tests. If they are clinically normal and the ultrasound is normal no follow-up is needed. The general practitioner should be sent a standard letter advising that they have been screened by ultrasound but still require follow-up in the surveillance clinic.

If any of the above children have minor abnormalities on the ultrasound scan then they should have a second ultrasound scan six weeks later. Children with subluxed or dislocated hips on ultrasound should be seen again in the paediatric orthopaedic clinic.

Any child who has a clinically dislocated or dislocatable hip at birth should have an immediate ultrasound and be referred to the next paediatric orthopaedic clinic within a week.

Box 8.1 Special risk of congenital dislocated hip

- Born in breech position
- Family history of congenital dislocated hip
- Foot deformity
- Meningomyelocoele

Management of abnormal hips

Apart from cases of irreducible hip dislocation, an infant with a dislocated or dislocatable hip should have a splint applied at about the age of two weeks. Further delay may cause poorer results. The Pavlik harness or variants of the von Rosen splint are used. They are made of malleable metal that has been padded and then covered with waterpoof material. The splint acts equally on both hips, keeping both thighs flexed and at an optimum degree of abduction. Unless substitution by a larger splint is required, the splint must remain in position continuously for at least two months if the hip is dislocatable or for at least three months if it is dislocated at rest. Early consultation with an orthopaedic surgeon, who will continue long term care, is essential.

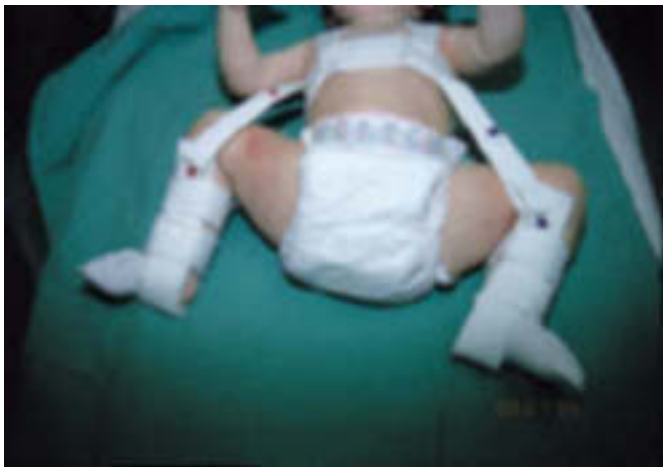


Figure 8.6 Pavlik harness.

Before the splint is first applied, the mother must be told exactly what this treatment entails. She should be encouraged to continue breastfeeding, even though this may prove awkward initially. The infant should be placed naked on the splint, which has been fashioned so that the posterior cross bar is grooved to protect the skin over the spine. If the hip is dislocated, the dislocation must first be gently reduced and the thigh held in the abducted position while the splint is carefully applied. Potential pressure points may be protected by inserting pieces of cotton wool or similar material. The mother should be told to replace these when they become wet or soiled without disturbing the splint. The baby's clothes should be put over the splint and not under it.

Once the splint is in place, the infant should be washed, weighed, and examined without the splint being removed. The splint will need scrupulous adjustment at each visit to allow for growth and to ensure that the hip is not being overabducted. The degree of full abduction in a normal infant diminishes slightly but steadily over the early weeks of life so that the splint should give the comfortable degree of full abduction that is appropriate for the child's age. Failure to adjust progressively for this may cause avascular necrosis of the femoral head. On the other hand, if the degree of thigh abduction is inadequate because the splint is applied too loosely, the hip may remain dislocated. The splint should be checked daily for several days after

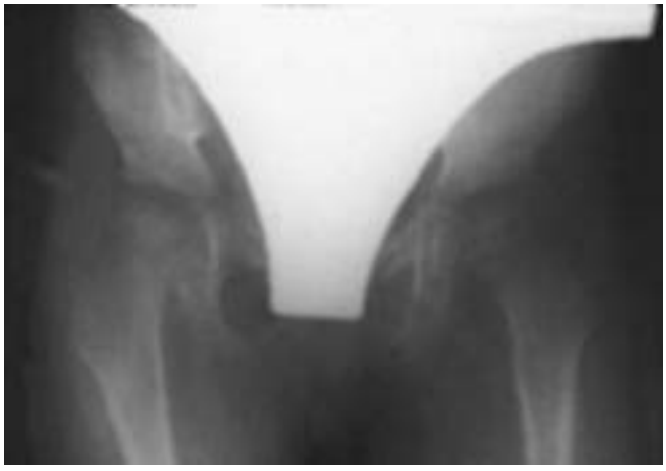


Figure 8.7 Radiograph of normal hips at 18 months.

application and thereafter at intervals of not more than one or two weeks.

Once the splint is finally removed, an anteroposterior radiograph of the hips should be taken with the thighs held parallel, and the infant should be followed up at the ages of six and 12 months, at the least.

Late diagnosis

If a dislocated hip is "missed" in the newborn infant the clinical diagnosis is often difficult until, after some weeks, the classic signs of lack of thigh abduction and, in unilateral cases, asymmetry of the lower buttock creases become apparent. Before the age of four months an ultrasound study is the most useful aid to diagnosis. Radiologically, once ossification occurs in the upper femoral epiphysis at about the age of 3–4 months a dislocation can be more clearly shown.

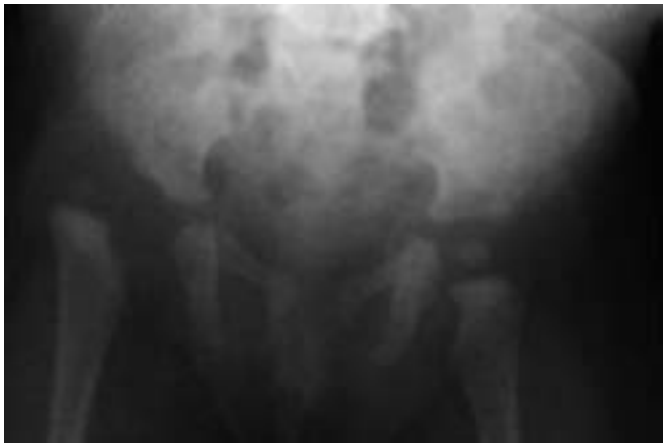


Figure 8.8 Radiograph of dislocated hip at eight months.

Acknowledgement

This chapter was written for the first edition by Dr HVL Finlay and has been revised for each edition.

9 Infection in the newborn

Some infections are fulminating and the infant dies within a few hours. More commonly, however, the onset is insidious and the vague features may include the refusal of feeds by an infant who has previously fed normally, lethargy, hypotonia, apnoeic attacks, or fever. Fever is arbitrarily defined as a temperature over 37.5°C, but newborn infants with infection often have a normal or even a subnormal temperature.

The most common pathogens are group B streptococci, *Escherichia coli, Staphylococcus aureus*, and Pseudomonas. The pathogens causing infection at a particular site can sometimes be predicted (for example, *Staph. aureus* in paronychia) but swabs for bacterial culture should still be taken before giving an antibiotic. Usually an antibiotic needs to be given before the organism has been isolated, but the treatment can be changed later once the sensitivity of the pathogen to antibiotics has been tested *in vitro*. The sensitivity of the pathogens in a particular unit is often known, which may give a guide to effective treatment. If septicaemia is suspected and there is no obvious site of entry of the pathogen, both a penicillin and an aminoglycoside (for example, gentamicin) must be given after a blood culture has been taken.

Intestinal absorption is variable and regurgitation of antibiotics common, so that the intramuscular or intravenous route should always be used initially. Intramuscular injections should be given deeply into the upper lateral aspect of the thigh. Schemes for rotating the sites are essential to prevent local necrosis and to avoid further injections being given into a relatively avascular area. Intravenous antibiotics are given slowly by bolus injection rather than by adding them to the bottle of intravenous fluid.

Box 9.1 Symptoms

- Refusal of feeds
- Lethargy
- Hypotonia
- Apnoea
- Fever

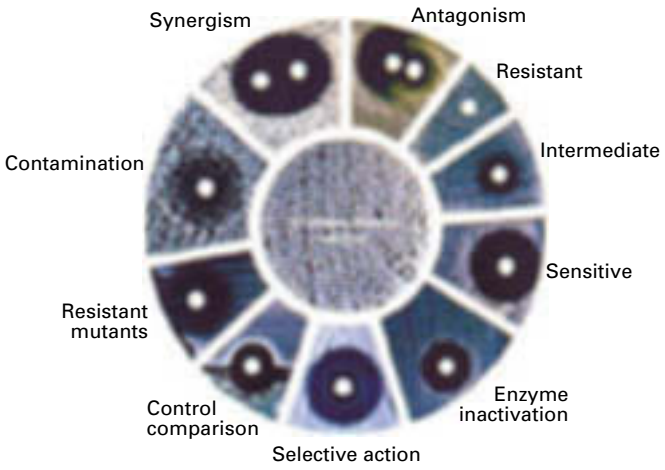


Figure 9.1 Antibiotic disc susceptibility test.

Box 9.2 Group B streptococcal infection

Early onset	• Raised respiratory rate • Peripheral cyanosis
Late onset	• Insidious septicaemia • Meningitis

Group B streptococcal infections

The group B Streptococcus is the commonest pathogen that causes severe infection in the first week of life. It is acquired during birth from the maternal vagina. Although about 10% of mothers are colonised and about 25% of their infants acquire this organism, only one in 1000 infants has symptoms. About half of those with symptoms die. In the early onset type, which occurs in the first few days of life, there may be a persistently raised respiratory rate followed by the vague features of septicaemia and later peripheral cyanosis. The chest radiograph may show extensive areas of consolidation in both lungs or it may be normal. The same organism may cause a more insidious septicaemia and meningitis towards the end of the first week.

Parents should be warned that recurrent group B streptococcal infection may occur for up to three months after birth and intravenous antibiotics will be the only effective treatment.

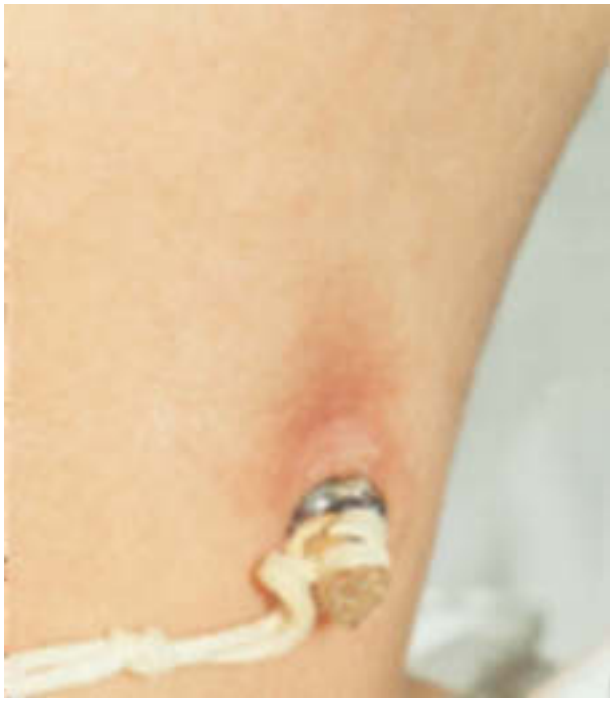


Figure 9.2 Infection of the umbilicus.

Escherichia coli infections

The use of antibiotics with a wide spectrum of action tends to eliminate all bacteria except *E. coli* and Pseudomonas, which flourish in a warm, moist environment. The preterm infant, who has a greater susceptibility to infection, nursed in an incubator with increased humidity provides the ideal setting for these infections, especially if broad spectrum antibiotics are

being given. The umbilicus may become infected and this may be followed by septicaemia and later meningitis.

Early symptoms are ill defined and the nurse may report that the infant appears vaguely unwell. He may have lethargy, anorexia, jaundice, and purpura. The early symptoms of meningitis are similar to those of septicaemia, but late features are vomiting, a high pitched cry, convulsions, raised anterior fontanelle tension, but rarely neck stiffness.

A urinary tract infection may easily be missed if a specimen of urine is not examined in every infant with fever or who is unwell. Rarely, there are physical signs of an associated congenital abnormality of the urinary tract, such as an enlarged kidney due to hydronephrosis or a persistently palpable bladder and poor urinary stream caused by obstruction from urethral valves in a boy.

Outbreaks of acute gastroenteritis due to rotavirus may occur in newborn nurseries.

Figure 9.3 Regular repeated examinations of the urine.

Staphylococcal infections

Conjunctivitis – *Staph. epidermidis* is often cultured from eye swabs of infants with conjunctivitis and it is impossible to determine whether this is the primary cause of the conjunctivitis or whether it is secondary to chemical inflammation. Chemicals, such as chlorhexidine, used in swabbing the mother's perineum may enter the infant's conjunctiva during delivery. Primary infection with *Staph. aureus* may cause severe conjunctivitis. After taking an eye swab from an infant with conjunctivitis, neomycin ophthalmic ointment should be applied to both eyes three times daily for a week.

If mild conjunctivitis persists longer than a week, the possibility of infection with *Chlamydia trachomatis* should be considered. The laboratory may be able to isolate the organism using special techniques. The best treatment consists of a course of tetracycline eye ointment, which should be continued for at least three weeks; a course of oral erythromycin should also be given for two weeks. (See below for gonococcal conjunctivitis.)

Skin sepsis – Clusters of yellow pustules in the axilla or groin that first appear on the fourth day or later are usually due to staphylococci. They may resemble the lesions of neonatal erythema, but those lesions appear earlier and resolve within 48 hours. If there is any doubt, the lesions should be treated as staphylococcal pustules.

Bullous impetigo of the newborn is a rare bullous eruption which can be rapidly fatal and very infectious if adequate antibiotics as intravenous flucloxacillin are not given.

Umbilical infection, shown by a localised red area and a serous exudate, may be due to *Staph. aureus* or a wide variety of other pathogens.

Paronychia is an infection along the side of the nail of the fingers or toes. It often affects several digits together. The lesion appears trivial but it may lead to any of the more serious infections due to the Staphylococcus. Surgery is not required but a long course of flucloxacillin should be given.

A breast abscess is a red, tender swelling, which may not affect the whole breast; usually there are enlarged lymph nodes in the axilla on the same side. Physiological enlargement may be unilateral. Surgical drainage is usually required as well as antibiotics for a breast abscess.

Pneumonia – The signs are the same as those of other types of pneumonia: raised respiratory rate, recession, and sometimes focal or generalised signs in the lungs. The infant appears extremely ill because of the accompanying septicaemia. A chest radiograph may initially show generalised non-specific changes and later lobar consolidation with

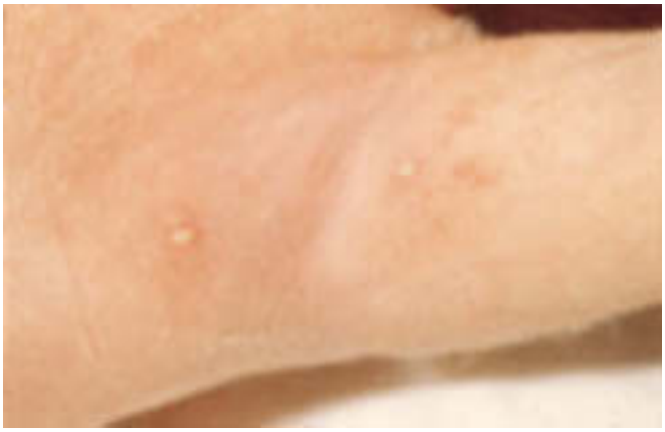


Figure 9.4 Pustules.

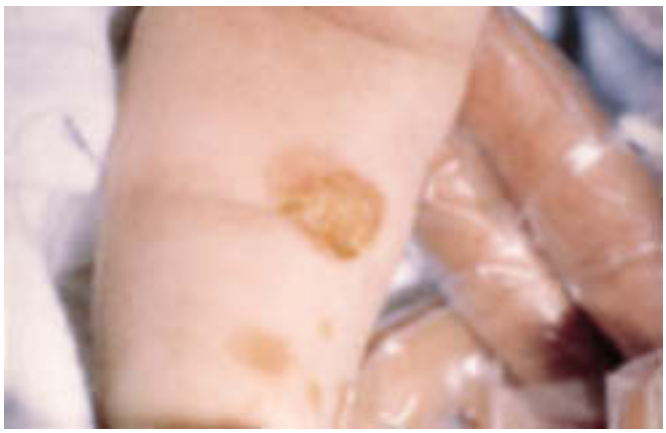


Figure 9.5 Bullous impetigo.

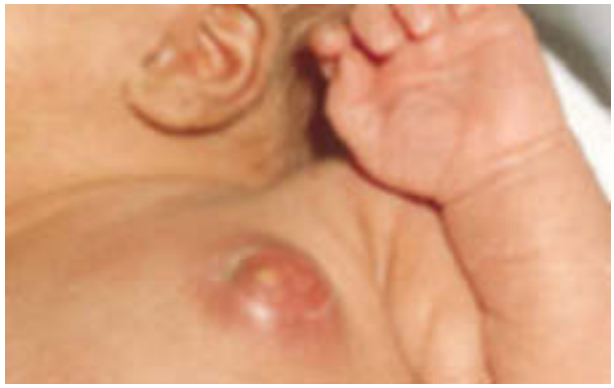


Figure 9.6 Breast abscess.

characteristic pneumatocoeles, which are air filled spaces. Complications are pneumothorax and empyema.

In *osteomyelitis* there are often no specific signs, just an ill infant who will not feed. Reluctance to move a limb and crying when the limb is moved or touched are valuable localising signs, but swelling at the site of the lesion is a late sign. A radioisotope scan may show changes at an early stage of neonatal osteomyelitis and later a radiograph of the limb may show a soft tissue swelling or a raised and thickened periosteum. Rarely, osteomyelitis in the newborn is not due to *Staph. aureus*.

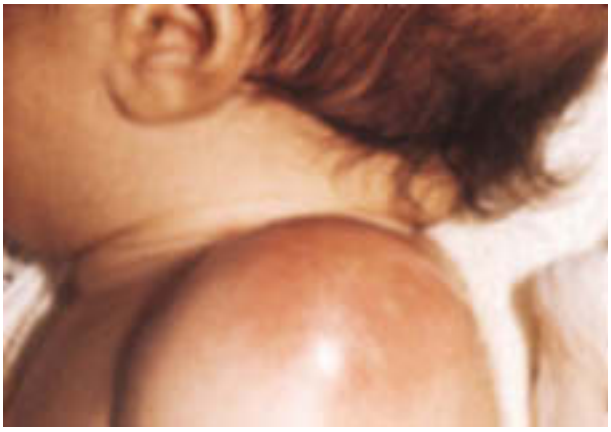


Figure 9.7 Osteomyelitis.

Gonococcal infection

Gonococcal conjunctivitis may be unilateral, usually does not respond completely to local neomycin, and is often present within 48 hours of birth. In cases of severe conjunctivitis, particularly when bilateral and associated with severe oedema, gonococcal infection should be suspected and special arrangements for taking swabs must be made with the laboratory. If the conjunctivitis does not respond rapidly to treatment, swabs should be taken again for gonococcal infection. The paediatrician should be told as soon as the laboratory confirms the diagnosis in view of the medical and legal implications. Urethral, cervical, and rectal swabs should be taken from the mother.

If gonococcal infection is suspected, even if smears do not confirm it, intravenous penicillin is given and the diagnosis is reconsidered when the results of the cultures are available. In addition, chloramphenicol eyedrops are given half hourly for the first six hours and then chloramphenicol ointment is applied two hourly for three days.

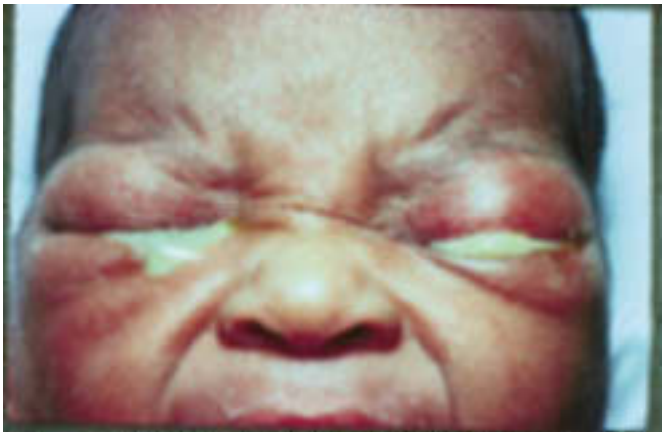


Figure 9.8 Gonococcal ophthalmitis.

Candida infections (fungal)

Candida infections are encouraged by the use of broad spectrum antibiotics. Candida infection of the mouth produces white plaques (thrush) and may make the infant reluctant to feed. It also causes a fiery red scaly eruption of the perineum, usually secondary to ammoniacal dermatitis. A swab made damp with sterile sodium chloride solution should be used to obtain a specimen for laboratory confirmation of the cause of the perineal rash. Oral nystatin suspension should be given after feeds for at least a week for the oral lesions. Nystatin ointment should be applied to the perineal rash each time the napkin is changed until the rash has resolved for a week. Miconazole oral gel and miconazole ointment may be used if nystatin suspension is not effective.

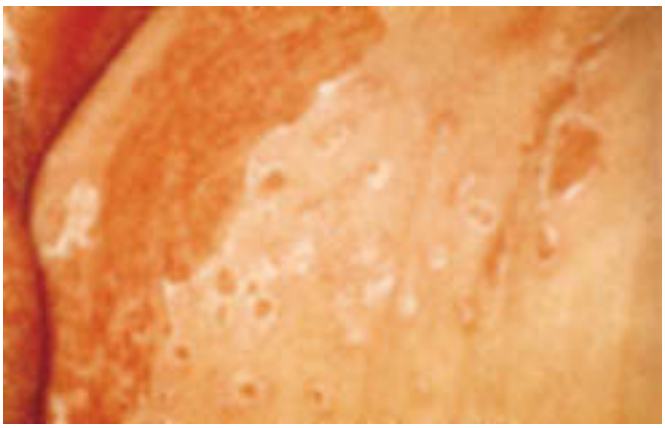


Figure 9.9 Candia infection.

Management and treatment

Careful evaluation of the features noted above may enable a tentative diagnosis to be made and to be confirmed by a single investigation. In many infants, however, there are no specific signs and the following investigations should be carried out.

Urine microscopy and culture – Specimens must be collected carefully and examined promptly to produce reliable results. A fresh midstream clean catch specimen can often be obtained after gentle suprapubic stimulation, especially in boys. If the urine specimen contains more than 10×10^6 pus cells per litre or bacteria are seen in a fresh specimen, a further urine specimen is obtained by suprapubic bladder puncture. Bacteria cultured from this specimen confirm a urinary tract infection.

Figure 9.10 Suprapubic puncture for urine collection.

Ultrasound renal scan should be performed in all infants with a confirmed urinary tract infection before discharge from the unit. Arrangements are made for the infant to be reassessed at regular intervals in the outpatient clinic and further investigations may be needed later. Inadequately treated neonatal urinary infections may cause permanent renal scars and later renal failure.

Other investigations that may need to be performed are (a) culture of gastric aspirate, ear swab, and umbilical swab; (b) chest radiograph; (c) blood culture from a peripheral vein, avoiding the femoral vein; and (d) lumbar puncture.

After the age of two days a blood neutrophil polymorph count higher than $10 \times 10^9/1$ suggests infection. A very low neutrophil polymorph count also suggests overwhelming infection.

As infants may deteriorate rapidly, treatment is started immediately for the most likely organisms involved if the child is ill. If the infant appears ill, antibiotics suitable for septicaemia should be given intravenously or intramuscularly after all the specimens have been taken. If the diagnosis is uncertain after physical examination, investigation of the urine, chest radiograph, and examination of the cerebrospinal fluid (CSF), and the infant appears well, it is better to withhold antibiotics until a paediatrician has been consulted.

Suitable antibiotics are intramuscular or intravenous flucloxacillin for staphylococcal infections; intravenous penicillin for group B streptococcal infections; and intravenous gentamicin for *E. coli* septicaemia. Intravenous cefotaxime or ceftriaxone is suitable for *E. coli* meningitis.

To treat urinary tract infection intravenous or intramuscular gentamicin and ampicillin should be given. Once the sensitivities are established it may be possible to continue treatment with one drug such as trimethoprim. In any case gentamicin should not be given for longer than a week, as eighth nerve damage may occur. Every infant who has had a confirmed urinary tract infection should have a repeat examination of the urine for the presence of infection every three months for two years.

Gentamicin is an effective antibiotic but it can cause deafness and blood levels must be monitored. The serum concentrations of gentamicin should be between 5 and 10 mg/1 one hour after the previous dose. The trough blood level, taken just before the next dose is due, should be less than 2 mg/1.

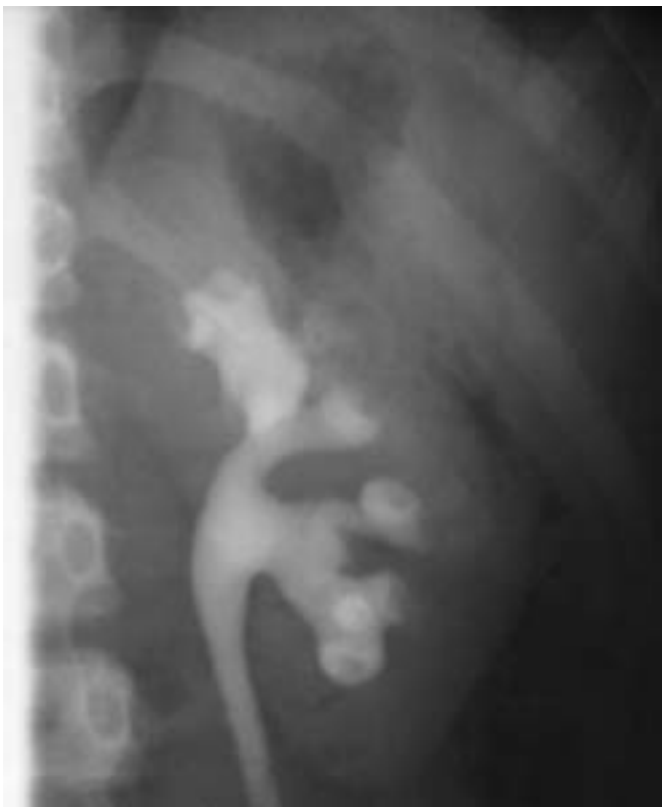


Figure 9.11 Hydronephrosis.

Box 9.3 Antibiotics for specific infections

Staphylococcus	Flucloxacillin
Streptococcus	Benzylpenicillin (IV)
E. coli:	
Septicaemia	Gentamicin
Meningitis	Cefotaxime

Paediatric HIV

Perinatal transmission of HIV from an infected mother to her infant can now be reduced to <5% by preventive measures taken during pregnancy and the perinatal period. It is recommended that all mothers should be encouraged to have an HIV screening test as part of their routine antenatal screening. It is now known that infants are most at risk of becoming infected with HIV during vaginal delivery, when they swallow maternal secretions containing the virus, and later during breastfeeding. The administration of anti-retroviral therapy (ART) to the mother during pregnancy is not harmful to the fetus and, together with elective caesarean section, avoidance of breastfeeding, and the administration of AZT (zidovudine) to the infant, has been shown to reduce the rate of vertical transmission from l5–30% to <5 % in countries where these measures are safe and affordable. The risk of transmission is related to maternal viral load in the third trimester and if this is undetectable in mothers who are on ART, caesarean section may not confer any additional

Box 9.4 Clinical features of HIV

- Recurrent thrush
- Generalised lymphadenopathy
- Parotitis
- Failure to thrive
- Pneumocystis pneumonia
- Developmental delay

advantage. Although breastfeeding doubles the rate of vertical transmission, in developing countries it may still be the preferred method of infant feeding because of greater health risks associated with formula feeding.

All infants born to HIV positive mothers will also test HIV antibody positive for at least the first 12–18 months of life, even those who are uninfected, because of the passage of maternal IgG across the placenta to the baby. Uninfected infants can usually be identified by showing absence of the virus on at least 2–3 tests by PCR (polymerase chain reaction) in the first four months of life. Until the results of the PCR tests are known, the infant will be given oral co-trimoxazole (daily or at least three times a week) from one month of age, in order to prevent infection with PCP (*Pneumocystis carinii*), which can cause a severe life threatening pneumonia in infected infants. Routine immunisation with the triple, Hib, and meningococcal vaccines should be given at the usual age. Some paediatricians recommend IM polio instead of oral polio vaccine for the infant, in order to avoid the small risk of immunocompromised adults in the family developing poliomyelitis. BCG vaccination is usually delayed until the infant is confirmed to be uninfected with HIV.

Infants infected with HIV may remain asymptomatic for 10–12 years after birth but around 20% will develop serious problems in the first year of life. Recurrent thrush, skin rashes, lymphadenopathy, hepatosplenomegaly, failure to thrive, developmental delay, and severe respiratory tract infections are common symptoms in HIV positive infants. ART is only necessary when the infant is shown to have a high viral load and a low number of CD4 lymphocytes. The survival of infected children has been greatly improved by early diagnosis and treatment but there is no definitive cure for HIV infected individuals as yet.

Families with HIV, even those where the children are not infected, benefit from support and help from sympathetic, well informed health care professionals.

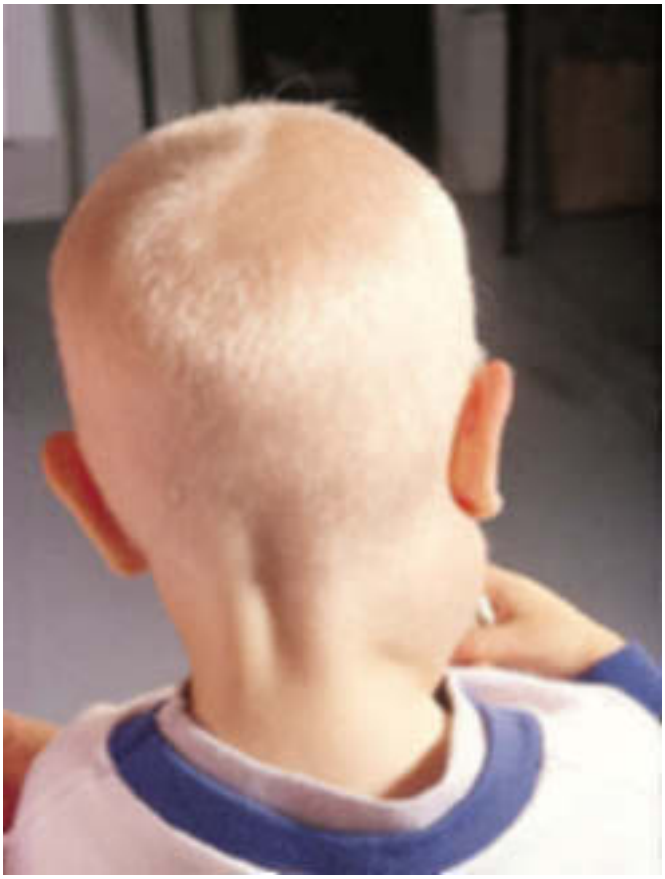


Figure 9.12 Cervical lymphadenopathy due to HIV infection.

Common napkin rashes

Napkin rashes are usually due to ammoniacal dermatitis, seborrhoeic dermatitis, or perianal excoriation.

Ammoniacal dermatitis is caused by ammonia produced by faecal bacterial enzymes acting on urea in the urine. Erythema, papules, scaling, and erosions appear in areas that have been in contact with napkins soaked with urine. The depths of the skinfolds are spared and the prepuce and scrotum are especially vulnerable. The principle of treatment is to keep urine and faeces away from the skin. Initial treatment includes changing napkins frequently or leaving the child without a napkin. A zinc cream is applied to the affected area every time the napkin is changed; in resistant cases a silicone cream is used. Some authorities consider that the formation of ammonia does not play a major part in this condition and they prefer the term "irritant napkin dermatitis".

If the rash does not improve within 10 days Candida infection of the lesions should be considered. This rash has a fiery red appearance with a scaly edge, often with a few early papules separated from the main eruption. Nystatin cream should be applied to the rash each time the napkin is changed for at least a week. The infant's mouth should be examined to determine whether white plaques of Candida are present and an oral suspension of nystatin is needed. Alternatively, miconazole oral gel may be used. In breastfed infants the mother's nipples should be examined, as a scaling dermatitis of the nipples can be due to Candida and may reinfect the infant.

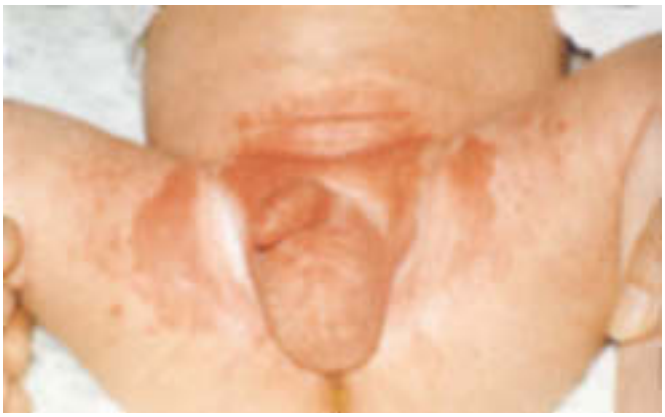


Figure 9.13 Ammoniacal dermatitis.

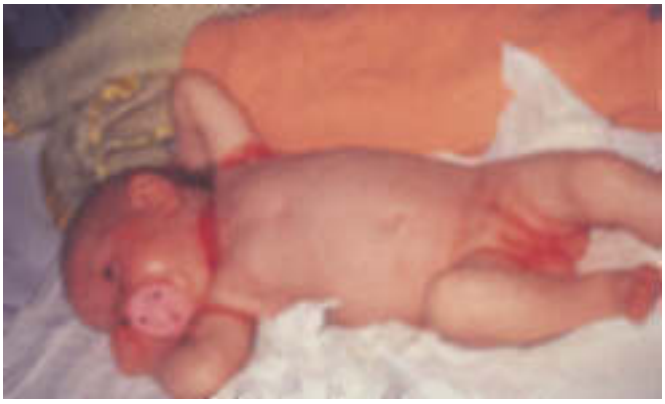


Figure 9.14 Seborrhoeic dermatitis.

Any infant with a rash in the napkin area should be undressed completely to ensure that he or she does not have a rash elsewhere. In *seborrhoeic dermatitis* the erythema and scaling may also affect the axillae, neck, the area behind the ears, scalp, forehead, and eyelids. The scalp may be covered with adherent hard crusted plaques (cradle cap). Secondary infection by staphylococci, streptococci, or Candida is common. The condition always begins before the age of three months and clears completely without treatment by nine months. In contrast to the infant with ammoniacal dermatitis, the infant with seborrhoeic dermatitis is oblivious of the rash. The cause is unknown and there is usually no family history of dermatitis. Treatment with 1% hydrocortisone ointment is rapidly effective and the lesions usually clear completely within a few weeks. In severe cases nystatin or an antibacterial agent may be added to the cream. Cradle cap can be removed by shampoo containing 0·5% salicylic acid.

If the rash affects mainly the perianal area, it is probably due to persistently loose stools. *Perianal excoriation* is commonly found in infants with gastroenteritis and resolves when the diarrhoea stops. Zinc cream or exposure may hasten the resolution of the lesions, but changes in the stool pH and consistency are the most important factors.

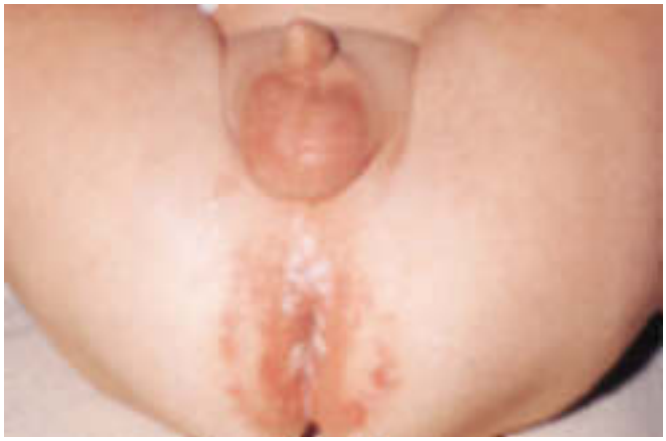


Figure 9.15 Perianal excoriation.

10 Jaundice in the newborn

Jaundice is a yellow colour of the skin caused by a high concentration of bilirubin. Very severe jaundice may damage the cells of the basal ganglia and brainstem. This damage is produced by the fat soluble unconjugated bilirubin. If jaundice is severe, high bilirubin levels may result in deafness, cerebral palsy, or death.

Neonatal jaundice is due to an increased bilirubin load with a transient inefficiency of hepatic excretion resulting from decreased activity of glucuronyl transferase in the liver. There are additional factors. Some of the conjugated bilirubin excreted in the bile is normally deconjugated in the small intestine and reabsorption is enhanced by the slower gut transit in the newborn who takes small volumes of milk. Bilirubin is absorbed from meconium and there is no intestinal flora to degrade bilirubin to urobilinogen.

Jaundice in the skin is visible to the naked eye in white babies at a serum bilirubin level of about 80 μmol/l. In black babies the sclerae should be examined as jaundice is more difficult to recognise. Jaundice first appears in the face and spreads to the periphery of the limbs (Figure 10.3). The level is more than 270 μmol/l if the hands or feet are jaundiced.

Common causes of jaundice in the newborn

Common causes of jaundice include hepatic immaturity, red cell incompatibility, infection, and breastfeeding.

Jaundice due to hepatic immaturity, or "physiological" jaundice, is common both in preterm and in full term babies. A temporary deficiency of glucuronyl transferase enzymes reduces the rate of conjugation of bilirubin, with a consequent retention of unconjugated bilirubin. In full term infants the jaundice always appears after the first 24 hours of life and reaches a peak on the fourth or fifth day. In preterm infants it usually begins 48 hours after birth and may last up to two weeks.

In babies with red cell incompatibility jaundice appears within 24 hours of birth. The main causes are: (*a*) incompatible rhesus grouping, and (*b*) incompatible ABO grouping; the mother's blood is usually group O and the infant's group A or, less commonly, group B.

The common infective causes of jaundice are septicaemia and urinary tract infection. Septicaemia is especially likely to be present if the jaundice appears after the fourth day of life, but it is a possibility in any infant who seems ill. In urinary tract infections the jaundice is of hepatic origin.

In about 2.5% of infants who are breastfed the serum bilirubin rises to levels between 260 and 360 μmol/l in the second or third week of life. These infants have no symptoms. If breastfeeding continues the level remains constant for three or four weeks and falls to normal levels at 4–16 weeks. An abnormal progesterone has been shown in the milk of some of the mothers.

Rare causes of jaundice in the newborn

Rare causes of jaundice include hypothyroidism, galactosaemia, viral hepatitis, and atresia of the bile ducts. These cause prolonged jaundice lasting more than 10 days.

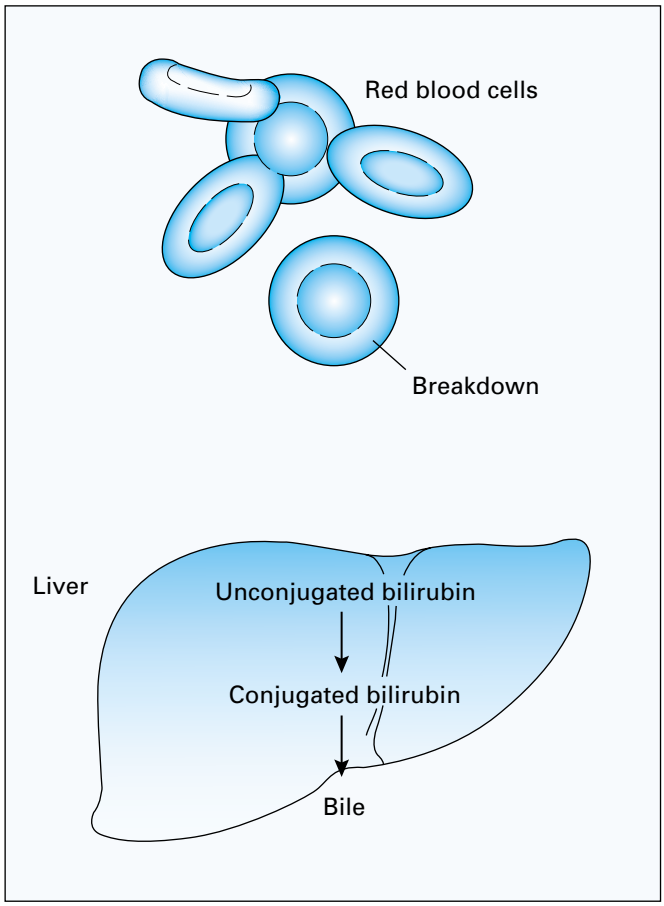


Figure 10.1 Formation and excretion of bilirubin.

Box 10.1 Causes of jaundice

Cause	Onset
Red cell incompatibility	Within 24 hours of birth
Physiological jaundice	After 24 hours
Septicaemia	Usually after fourth day

Figure 10.2 Breast feeding.

Glucose-6-phosphate dehydrogenase deficiency is another cause of prolonged jaundice, but it can also produce a clinical picture similar to blood group incompatibility.

In hypothyroidism physiological jaundice is prolonged, the plasma thyroxine (T4) concentration is reduced, and the thyroid stimulating hormone (TSH) concentration is increased.

In infants with viral hepatitis, which is usually due to intrauterine infection, the stools are pale, the urine dark owing to bile, and there is a high level of conjugated bilirubin in the plasma.

It is difficult to differentiate between hepatitis and atresia of the bile ducts clinically, and they may represent the two ends of a range of disease. If jaundice persists more than 10 days, the advice of a paediatrician should be sought.

In galactosaemia the urine gives a positive result on testing for reducing substances, but the test for glucose may be negative. The infant needs to be referred to a paediatrician immediately for special investigations.

Glucose-6-phosphate dehydrogenase deficiency occurs in infants of Mediterranean, African, or Chinese stock. This hereditary red cell enzyme defect is found in babies with haemolytic episodes that often occur without the usual precipitating factors of drugs or infection. The enzyme is necessary for maintaining the stability of the red cell membrane.

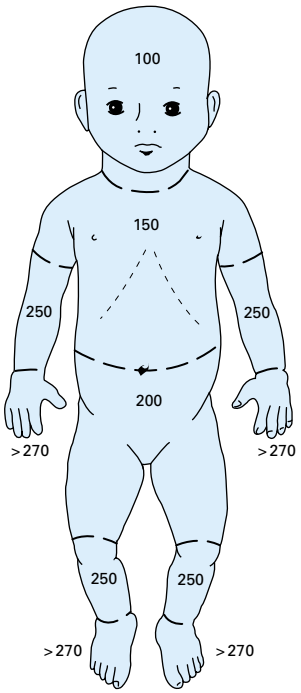


Figure 10.3 Spread of jaundice in the skin.

Box 10.2 Rare causes of jaundice

- Hypothyroidism
- Galactosaemia
- Viral hepatitis
- Atresia of bile ducts
- Glucose-6-phosphate dehydrogenase deficiency

Management of jaundice starting in the first 24 hours

If jaundice appears within 24 hours of birth it must be considered to be due to blood group incompatibility, with a high risk of cerebral palsy, until proved otherwise. Most rhesus problems should be anticipated before the neonatal period. Urgent exchange transfusion may be indicated in infants severely affected by haemolytic disease of the newborn, and it is advisable for the infant to be admitted to hospital immediately for investigation. If the mother is rhesus negative, the infant rhesus positive, and the Coombs test positive, jaundice is due to rhesus incompatibility. The plasma bilirubin concentration should be measured every five to eight hours and the results plotted on a special chart (see p. 44). Once the second estimation has been performed the maximum concentration can be predicted, as the rate of increase is linear. When the serial concentrations fall below the printed line, the infant is unlikely to need any treatment.

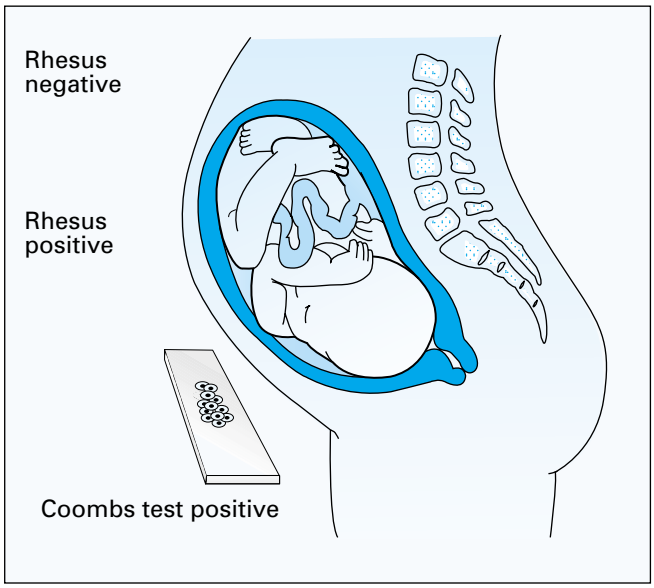


Figure 10.4 Rhesus incompatibility.

Management of jaundice starting after the first 24 hours

The possibility of septicaemia or urinary tract infection should be considered in any ill baby who develops jaundice after the first 24 hours of life. If there is any doubt about when the jaundice first appeared, the possibility of blood group incompatibility should be investigated.

When a doctor visits the infant at home a guide to the plasma bilirubin concentration can be provided by the dermal icterometer. This is a piece of transparent perspex with yellow lines of various shades that correspond to plasma bilirubin concentrations. It should be compressed gently on the infant's nose to indicate the approximate plasma bilirubin concentrations. The dermal icterometer is not accurate in artificial light, when bilirubin values are rising rapidly, or

Box 10.3 Onset after 24 hours but serum bilirubin >300 μmol/l

- Serum bilirubin – conjugated and unconjugated
- Full blood count and white count differential
- Blood group
- Coombs test
- Blood culture
- Urine culture (for asymptomatic infection)
- Glucose-6-phosphate dehydrogenase levels in the appropriate racial groups

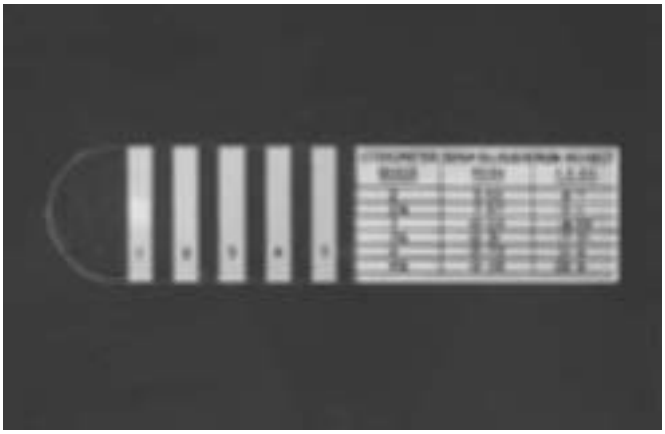
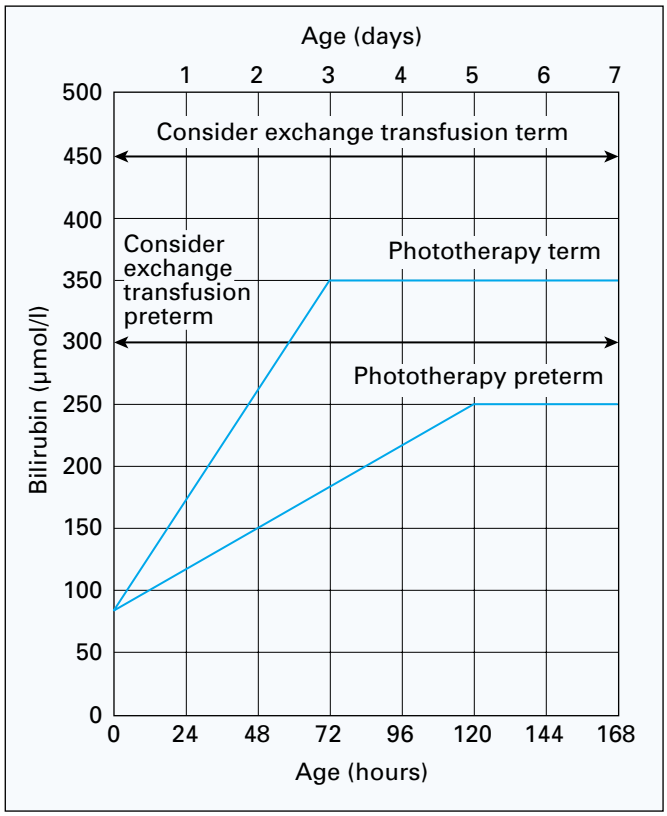


Figure 10.5 Dermal icterometer

when phototherapy has been given. In infants with pigmented skin the dermal icterometer should be compressed against the gums, not the nose.

In full term infants if the dermal icterometer suggests that the plasma bilirubin concentration is about 250 μmol/l or the result lies above the line on the chart the infant should be tranferred to hospital immediately. If the plasma bilirubin level is very high, urgent exchange transfusion or phototherapy may be needed.

In neonatal units there is usually a ward bilirubinometer that measures plasma bilirubin concentrations within a few minutes, using a small specimen of blood obtained by heel prick. If two estimations fall below the line on the chart, treatment will probably not be needed. Those with values above the line may need exposure to light under a phototherapy unit. Phototherapy produces geometric stereoisomers of bilirubin, which have no known long term deleterious effect on the infant. The infant's eyes are shielded with eye pads, but the mother should have the procedure explained first. Although phototherapy units have a shield to reduce transfer of heat from the lamp to the infant, monitoring of the temperature of the infant is essential. Extra fluids may be needed to compensate for the additional evaporative loss and this can be given as plain water to breastfed infants or additional milk to bottlefed babies. Oral fluid decreases the gut transit time and improves the excretion of bilirubin and the associated compounds. The indications for phototherapy are controversial but many units give phototherapy if the plasma bilirubin level is above the line on the chart. Despite phototherapy, an exchange transfusion may still be needed, but the critical level varies with the unit and the gestational age of the infant. Exchange transfusion should be considered if the plasma bilirubin level exceeds 450 μmol/l in a full term infant and 300 μmol/l in a preterm infant. Some units start treatment at lower levels of bilirubin in sick infants.



Bilirubin chart

400 μmol/l = 23.4 mg/100 ml	100 μmol/l = 5.8 mg/100 ml
300 μmol/l = 17.5 mg/100 ml	17.1 μmol/l = 1.0 mg/100 ml
200 μmol/l = 11.7 mg/100 ml	

Figure 10.6 Indications for phototherapy and exchange transfusion.

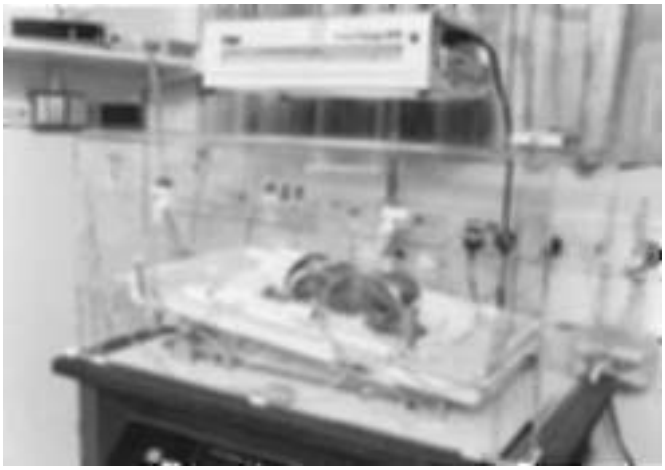


Figure 10.7 Phototherapy.

Prolonged jaundice

If jaundice persists longer than 14 days in a full term infant, blood should be taken for plasma thyroxine and TSH estimations and a specimen of urine collected to measure reducing substances and glucose. The urine should be examined in the laboratory for the presence of infection. If the parents are of Mediterranean, African, or Chinese origin, the screening test for red cell glucose-6-phosphate dehydrogenase should also be performed.

Pale stools and a plasma conjugated bilirubin level greater than 30 μmol/l suggest the possibility of hepatitis or atresia of the bile ducts, and the advice of a paediatrician is needed.

If there is a suspicion that the jaundice is related to breastfeeding, the other conditions causing jaundice should be excluded and the mother advised to continue breastfeeding. If the plasma bilirubin concentration is rising rapidly and breastfeeding is stopped for 48 hours, the infant's plasma bilirubin concentration will fall abruptly and will not usually rise on return to breastfeeding. Although the mother can continue lactation by expressing her milk during this diagnostic test there is a risk that breastfeeding will not be resumed.

Box 10.4 Prolonged jaundice >14 days

- Serum bilirubin – conjugated and unconjugated
- Full blood count
- Thyroid function tests
- Liver enzymes
- Glucose-6-phosphate dehydrogenase if appropriate
- Urine culture
- Urine for reducing substances (for galactosaemia)
- Urine for presence of bilirubin which reflects a high conjugated bilirubin (on a Labstix).

11 Convulsions in the newborn

Although the involuntary movements of a convulsion are usually generalised, they may affect only one limb, the face, or tongue. The distinctive feature is repetitive jerky movements, which may be accompanied by loss of consciousness, apnoea, or rigidity. Often the convulsion has stopped by the time the baby is seen by a doctor. A brisk Moro reflex or jerky normal movements in a baby may be misinterpreted as a convulsion by an inexperienced observer. If the convulsion is not seen, apnoea or cyanosis may indicate that it has occurred, and if there is any doubt it is safer to investigate the baby on the assumption that there has been a convulsion.

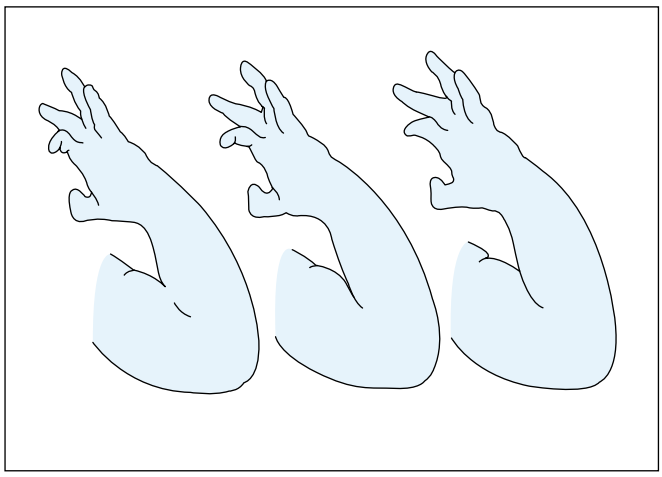


Figure 11.1 Repetitive similar movements indicate a convulsion.

Management

A baby not already in hospital should be admitted. After hypoglycaemia has been excluded by a blood glucose estimation, an anticonvulsant should be given if the convulsion is continuing or recurs while waiting for the results of other investigations. The order of carrying out the various procedures is important. Hypoglycaemia and hypocalcaemia should each be sought for and excluded in that order before the next test. Hypoglycaemia is the more dangerous condition.

The infant is placed on his side and the airway cleared by suction of the pharynx under direct vision. The baby should be nursed in an incubator to improve observation. Oxygen is given in high concentration with a funnel or head box until the convulsion has stopped.

If the plasma glucose and calcium concentrations are normal, a paediatrician should decide whether a lumbar puncture is indicated.

Box 11.1

Exclude:
1 hypoglycaemia
2 hypocalcaemia

Figure 11.2 Glucometer.

Hypoglycaemia

A specimen of blood should be taken from a heel prick immediately for glucose measurement in a glucometer. If the value is less than 3.0 mmol/l, hypoglycaemia *may* be present. A further blood sample should be taken and part of it used to repeat the glucometer test and the remainder taken into a fluoride tube for the laboratory to check the blood glucose concentration. Hypoglycaemia is defined as a laboratory estimated blood glucose level less than 2.6 mmol/l. If the second glucometer value is low, 5 ml of 10% glucose solution per kg body weight should immediately be given intravenously slowly by a scalp or limb vein. A continuous intravenous infusion of 10% glucose solution is then set up at a rate of 60 ml/kg a day.

Hypoglycaemia accompanied by fits is treated with intravenous glucose and when the fits have ceased, continuous intragastric milk should be given.

The amount of intravenous glucose should be reduced gradually over at least 24 hours. During this period three hourly glucometer tests are carried out, supplemented by regular laboratory blood glucose measurements. If an intravenous infusion of glucose is stopped suddenly by accident, severe reactive hypoglycaemia may follow and cause severe convulsions.

In contrast, asymptomatic hypoglycaemia has a good prognosis. Intragastric milk should be given by continuous infusion but the milk volume should be increased by 25% to raise the blood glucose concentration. Oral or intravenous

Figure 11.3 Syringe pump for milk.

glucose should be prescribed only on the advice of a senior member of the unit. The blood glucose level is repeated an hour after starting the milk.

Hypocalcaemia

If the glucometer test result is normal, blood should be taken for emergency plasma calcium estimation. If convulsions are still occurring or recur after hypoglycaemia has been excluded by the glucometer intramuscular paraldehyde should be given while the results of detailed tests are awaited. If the plasma calcium concentration is lower than 1.8 mmol/l (7 mg/100 ml) treatment depends on whether the convulsions are still occurring or recurring. If the convulsions have stopped, 1–2 ml of 10% calcium gluconate is added to each feed. The calcium gluconate should be added to the feed and not given directly to the infant. The total dose of calcium gluconate in 24 hours should not exceed 12 ml of the 10% solution in the full term infant.

If the convulsions continue and the plasma calcium concentration is low, 10% calcium gluconate should be diluted to 2.5% with 5% glucose solution in a syringe and given slowly intravenously into a scalp or limb vein until the convulsions cease or until a maximum of 4 ml/kg body weight of the 2.5% diluted solution has been given. The heart rate is monitored with a cardiac monitor or stethoscope during the procedure and the injection stopped if bradycardia occurs. Calcium gluconate should be added to the feeds until the plasma calcium concentration rises to normal. Calcium gluconate must never be given intramuscularly or allowed to escape out of a vein as severe tissue necrosis may occur.

Intravenous calcium may not control hypocalcaemic convulsions immediately. Some authorities have found that hypocalcaemic convulsions are treated effectively by intramuscular magnesium sulphate (0.2 ml/kg of a 10% solution/dose).

Hypomagnesaemia

In some units there is a high incidence of fits associated with low plasma magnesium concentrations. If there are recurrent fits or the plasma calcium concentration does not rise despite supplementary calcium gluconate, the plasma magnesium concentration should be estimated. If this is less than 0.6 mmol/l (1.5 mg/100 ml) intramuscular magnesium sulphate is given. The dose is 0.2 ml/kg of a 10% solution given intramuscularly every six hours; if the plasma level is normal no further treatment is needed. The plasma estimation must be repeated after two days. The main toxic effect is hypotonia, which can be reversed by intravenous calcium gluconate if the features are severe.

Prognosis

Hypoglycaemia and hypocalcaemia may be found together, especially in infants of diabetic mothers, but hypoglycaemia is the more dangerous. Symptomatic hypoglycaemia may be followed by mental impairment but symptomatic hypocalcaemia or hypomagnesaemia has an excellent prognosis, although enamel hypoplasia, which predisposes to dental caries, is a late complication of hypocalcaemia. Intravenous glucose is not hazardous provided that the correct dose is given, but intravenous calcium salts may lead to cardiac arrest.

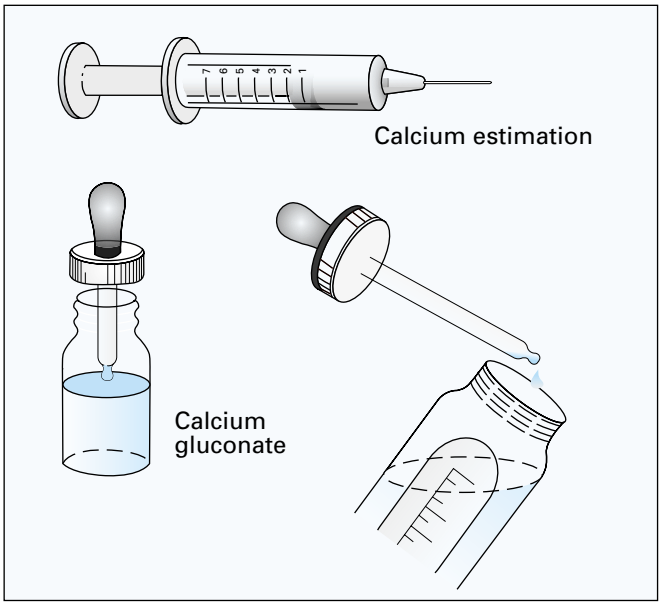


Figure 11.4 Calcium supplements.

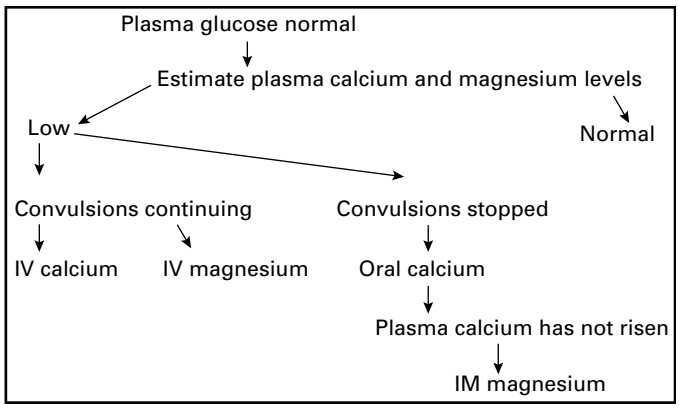


Figure 11.5 Biochemical investigation of a convulsion.

Box 11.2 Prognosis

- Symptomatic hypoglycaemia: cautious prognosis
- Symptomatic hypocalcaemia: good prognosis

Cerebral lesions

After hypoglycaemia and hypocalcaemia have been excluded, meningitis, intracranial haemorrhage, and cerebral oedema due to perinatal asphyxia should be considered. Lethargy, hypotonia, and raised tension of the anterior fontanelle are suggestive of a cerebral lesion but meningitis may be present with no specific symptoms or signs. Lumbar puncture is indicated if there is a possibility of meningitis. Ultrasound examination is reliable in excluding an intraventricular or intracerebral haemorrhage but may not detect a small subarachnoid or subdural haemorrhage.

If the first convulsion is prolonged or the convulsion recurs, intravenous phenytoin or phenobarbitone should be given. The initial dose of phenytoin or phenobarbitone is 15 mg/kg given intravenously over half an hour. The maintenance dose is started 24 hours later and is usually required for only two or three days.

Box 11.3 Cerebral causes of a convulsion

- Meningitis
- Cerebral oedema
- Intracranial haemorrhage

12 Vomiting

Vomiting in the newborn: types of vomit

Vomiting is the forceful expulsion of gastric contents through the mouth. Mothers often confuse vomiting with mild regurgitation, which is the effortless bringing up of small amounts of milk during and between feeds, usually accompanied by air. If the milk dribbles down the chest it is likely to be regurgitation. Babies often bring up small amounts of milk with air and this is of no importance. Recurrent vomiting may be a sign of lethal disease, but a careful history and examination enable a diagnosis to be made with the minimum of special investigations.

Frothy mucoid vomit

Oesophageal atresia with tracheo-oesophageal fistula may present with vomiting, coughing, and cyanosis when the infant begins the first feed. Many of these infants drool frothy material continuously rather than vomit. Vomiting of *frothy* mucoid material may be the only definite observation, but the condition should be suspected in any baby who has *any* symptoms during the first feed. As the fluid expelled is not gastric contents, vomiting is not an accurate description but this is the term often used.

Bilestained vomit

The vomit in infants with intestinal obstruction is usually yellow, due to bile staining, but occasionally it consists only of milk. The cause may be atresia, stenosis, or volvulus of the small gut, necrotising enterocolitis, or congenital intestinal aganglionosis (Hirschsprung's disease) of the large gut. Abdominal distension is usually present and there may be visible peristalsis. A plain radiograph should be taken in the erect position immediately. An alternative is a radiograph in the supine position and, in addition a lateral view with the baby lying on his or her back and a horizontal X-ray beam. They often show fluid levels, dilated loops of gut proximal to the obstruction, and the absence of gas shadows distally. Ideally, every infant who vomits bile should be seen by a surgeon within an hour.

Bloodstained vomit

Bloodstained vomit may be caused by trauma from a feeding tube, swallowed maternal blood, or, most seriously, haemorrhagic disease of the newborn. Trauma caused by a feeding tube may produce a few specks of blood in the vomit. Maternal blood may be swallowed before delivery after premature separation of the placenta or after delivery as the result of bleeding from a cracked nipple. Maternal haemoglobin in the vomit can be recognised in the laboratory.

Haemorrhagic disease of the newborn begins between the second and fourth days of life in the early type or in the third or fourth week in the late type. The first symptom may be haematemesis or melaena and the bleeding can be profuse. An immediate dose of 1 mg vitamin K_1 should be given intramuscularly and a transfusion of fresh blood given urgently if bleeding has been severe or persists after vitamin K treatment.

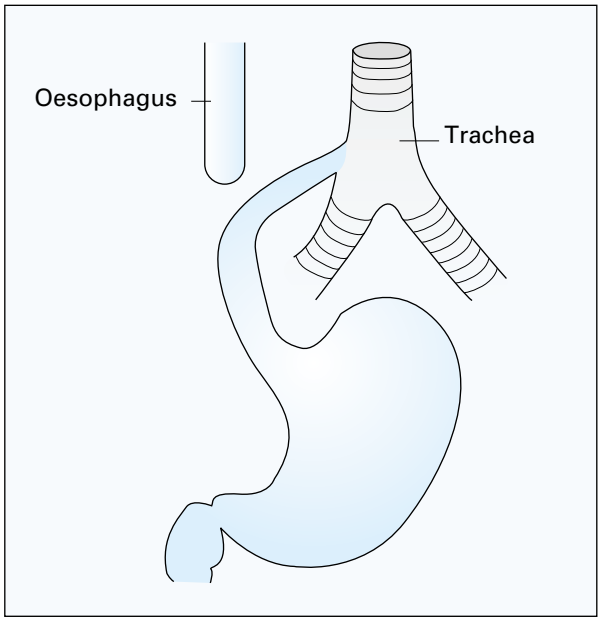


Figure 12.1 Oesophageal atresia.

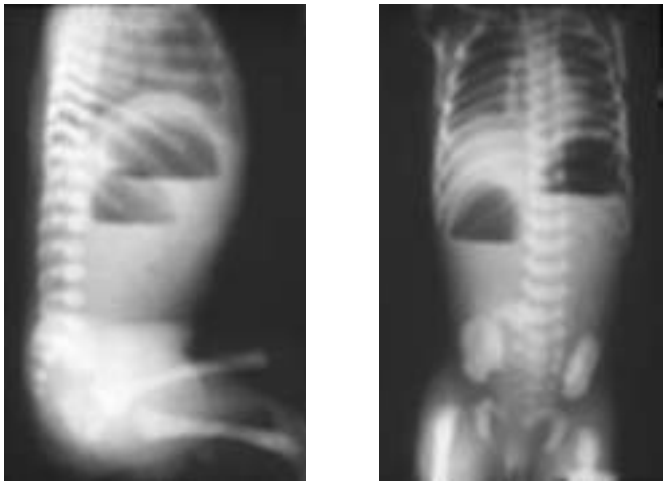


Figure 12.2 Erect lateral (left) and anteroposterior (right) radiographs to show fluid levels.

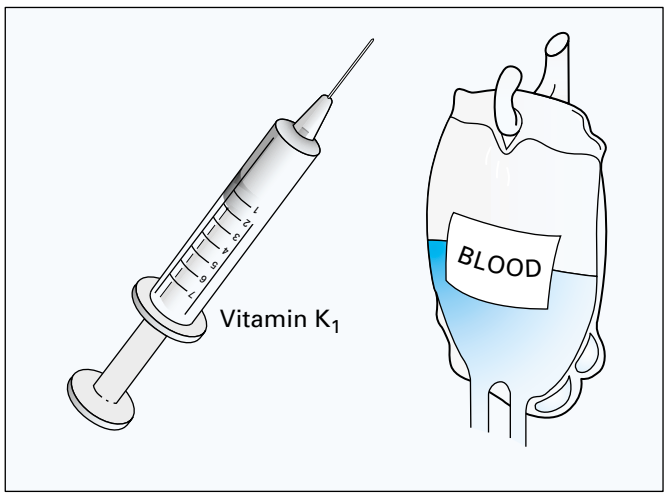


Figure 12.3 Treatment of haemorrhagic disease of the newborn.

Milk

Vomiting of milk may be caused by infections, feeding problems, necrotising enterocolitis, intracranial haemorrhage, or drugs.

Gastroenteritis, urinary tract infection, septicaemia, and meningitis may all be associated with vomiting. A ravenous infant may swallow excessive air at the beginning of the feed and, if not properly "winded", may later regurgitate milk with air. Larger feeds, more frequent feeds, or a larger hole in the teat is needed.

Necrotising enterocolitis occurs in epidemics in neonatal baby units. Lethargy and refusal of feeds are followed by vomiting and abdominal distension. In the majority of the infants there is blood in the stool. Predisposing factors are prematurity, perinatal hypoxia, hypotension, umbilical vessel catheterisation, and prolonged rupture of the membranes. Embolism or thrombosis of mesenteric vessels is followed by ischaemic changes, which vary from mucosal ulceration to complete necrosis of the gut wall. Bacteria invade the necrotic tissue and healing is followed by scarring and sometimes a stricture. Despite optimal treatment there is a 25% mortality rate and early advice from a paediatric surgeon is advisable.

Raised intracranial pressure due to intracranial haemorrhage may cause vomiting of milk, as may several drugs, especially digoxin.

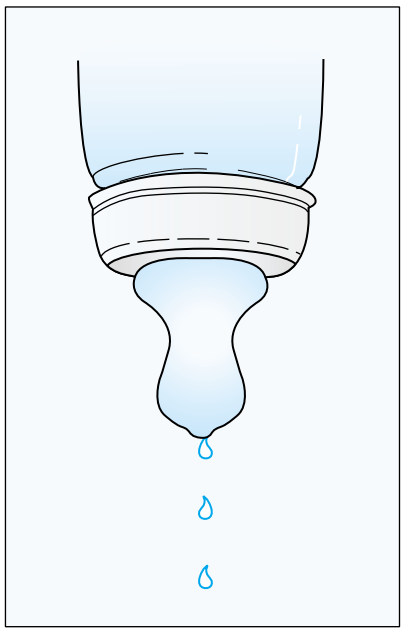


Figure 12.4 Drops of milk should follow quickly and there should not be a continuous stream.

Vomiting from the first week to the first year: causes

As in the newborn, vomiting in infants older than one week may be a symptom of a feeding problem or an infection such as urinary tract infection, otitis media, gastroenteritis, septicaemia, or meningitis (see p. 90).

If no cause of the vomiting is found and the symptoms are mild, urine should be collected for microscopy and culture and should be examined for protein, bile, and reducing substances.

Box 12.1 Vomiting after the first week

- Feeding problems
- Infections:
  - Urinary tract infection
  - Septicaemia
  - Meningitis

Pyloric stenosis

Pyloric stenosis must be considered in every infant less than three months of age who vomits. Rarely the vomiting may occur in the first week of life, but it usually begins in the second or third week, though there may be a delay before the infant is seen by a doctor. Usually the vomit is produced forcefully and reaches some distance from the infant. The infant often accepts another feed immediately after vomiting. Stools are infrequent. If the symptoms have been present for more than a few days, there will be a loss of weight due to dehydration and loss of subcutaneous fat. Scanty urine is associated with dehydration.

The essential diagnostic sign is the presence of a pyloric mass palpated during a test feed. Constant practice is needed to appreciate a pyloric mass. Even if no pyloric mass is felt during the first test feed, if the diagnosis of pyloric stenosis is probable the infant should be admitted for rehydration and the examination repeated. Preliminary aspiration and measurement of gastric contents is helpful, particularly if no feed has been given during the preceding four hours. Metabolic alkalosis strongly suggests the diagnosis.

In a small proportion of infants the diagnosis of pyloric stenosis is suspected clinically but no pyloric mass is palpable during a test feed. The diagnosis may be confirmed, preferably by an ultrasound study. A barium study is needed rarely.

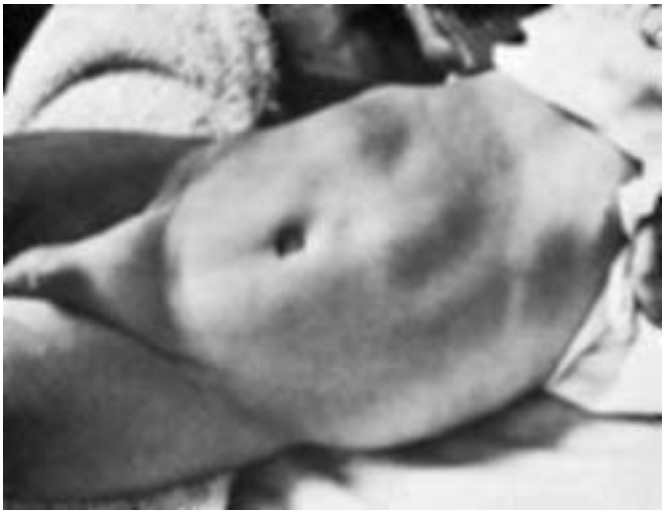


Figure 12.5 Pyloric stenosis.

Intestinal obstruction

Infants who vomit greenish yellow bile are likely to have intestinal obstruction. They should be admitted immediately and seen by a surgeon within an hour. Abdominal distension is often present and peristalsis may be visible. Duodenal stenosis usually presents during the first few days of life but malrotation of the gut with associated volvulus may produce symptoms at any time during childhood.

An inguinal hernia is more likely to incarcerate in the early months of life than later. Incarceration should be suspected if the hernia is tender or is not reduced easily; immediate surgery is required. The risk of obstruction is always present and early surgical treatment is advisable in every baby with an inguinal hernia. The baby must remain in the ward until the operation is performed.

An intussusception is a partial or complete intestinal obstruction due to invagination of a portion of the gut into a more distal portion. It may occur at any age, although the maximum incidence is at 3–11 months. An intussusception may be easily diagnosed in a child who has all the typical features, but these children are not common. The distinctive feature is the periodicity of the attacks, which may consist of severe screaming, drawing up of the legs, and severe pallor. Some episodes consist of pallor alone. The attack lasts a few minutes and may recur about 20 minutes later, though attacks may be more frequent. There may be vomiting and one or two loose stools may be passed initially, suggesting acute gastroenteritis. Bloodstained mucus may be passed rectally or shown by rectal examination. But some patients pass no blood rectally. Between attacks the infant appears normal and may have no abnormal signs apart from a palpable mass.

It is difficult to examine the abdomen during an attack because the child cries continuously, but between attacks a mass, most commonly over the right upper quadrant, can be felt in 70% of children.

If surgical shock is present, then rapid resuscitation should be carried out and intravenous fluids, including blood, given. Plain radiograph of the abdomen may show evidence of intestinal obstruction or a density in the area of the lesion. Ultrasound may show a doughnut configuration with hypoechogenic rims and a dense central echogenic core. An urgent surgical opinion should be obtained. If the symptoms have been present for less than 48 hours and there are no signs of intestinal perforation, an air or barium enema should be given urgently while the surgeon remains nearby. In over 75% of cases it is possible to reduce the intussusception by the enema. If the intussusception is not reduced, then immediate laparotomy is needed to reduce the lesion manually or to perform an intestinal resection. In about 6% of cases there is a persisting mechanical cause of the intussusception and this will not be detected by the enema.

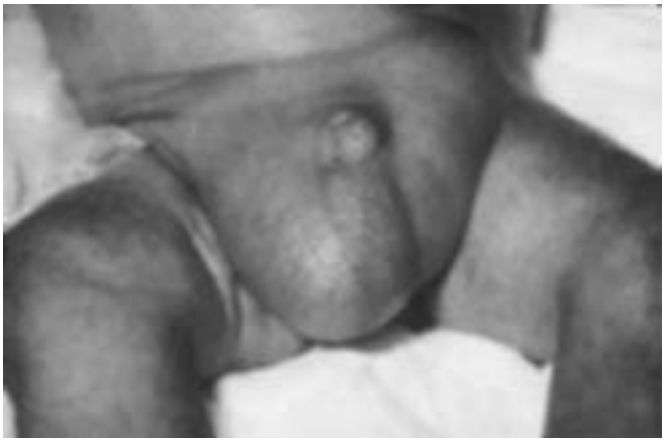


Figure 12.6 Inguinal hernia.

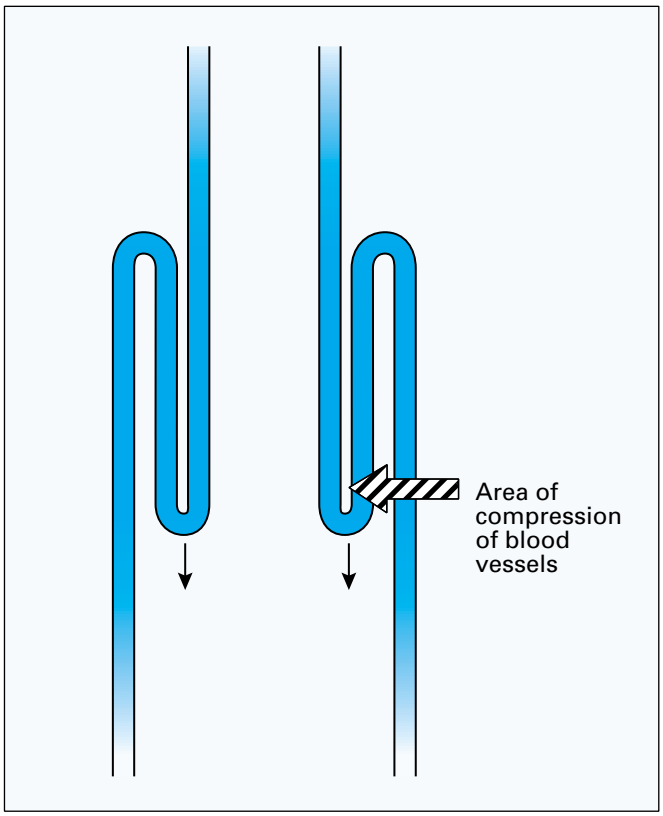


Figure 12.7 Intussusception.

Gastro-oesophageal reflux

Vomiting due to gastro-oesophageal reflux starts during the first week of life and the vomitus may be blood stained. Aspiration into the lungs may cause recurrent bronchospasm and severe vomiting may cause failure to thrive, dysphagia, or stricture formation. During the first year the lower oesophageal sphincter pressure increases and oesophageal mobility becomes more organised. These factors reduce the regurgitation of gastric contents into the oesophagus when the intra-abdominal pressure rises, for example during crying.

The diagnosis can be confirmed by 24 hour pH monitoring of the lower oesophagus. A probe the size of a

Box 12.2 Features of gastrooesophageal reflux

- Vomiting
- Aspiration
- Failure to thrive
- Dysphagia
- Stenosis (rare)

nasogastric tube is placed just above the gastro-oesophageal sphincter. The pH recording is analysed by computer. Barium swallow examination is often negative despite typical symptoms.

Vomiting usually resolves by the age of one year without specific treatment. If symptoms are severe, the feeds can be thickened with carob seed flour or ground rice and the infant may be nursed with his head higher than his feet on his side. If the vomiting is persistent and severe, a paediatrician should be consulted. Most paediatricians recommend a trial period with thickened feeds before a pH recording.

Whooping cough

Vomiting may be so severe in infants with whooping cough that the mother is more worried by the vomiting than the cough. During the first five days of the illness (catarrhal phase) there is a short, dry nocturnal cough. Later, bouts of 10–20 short coughs occur day and night. The cough is dry and each cough is on the same high note or goes up in a musical scale. The long attack of coughing is followed by a sharp indrawing of breath, which causes the whoop. Some children with proved pertussis infection never develop the whoop. Feeding often provokes a spasm of coughing and this may culminate in vomiting. Afterwards there is a short refractory period during which the baby can be fed again without provoking more coughing. In uncomplicated cases there are no abnormal signs in the respiratory system.

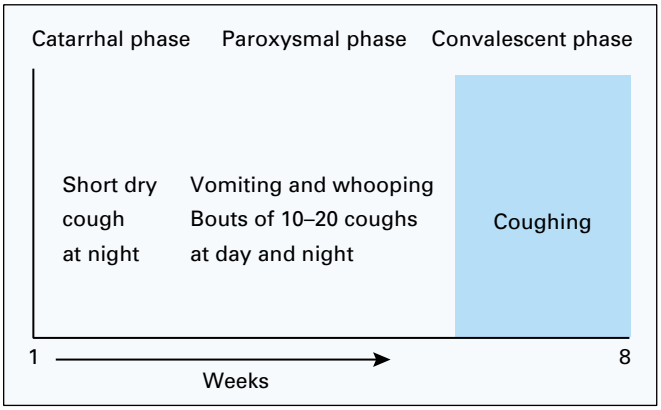


Figure 12.8 Phases of whooping cough.

Adrenogenital syndrome

The adrenogenital syndrome (salt losing type) commonly presents with vomiting as the only symptom in boys. The diagnosis is easier in girls, as virilisation of the external genitalia will have been noticed at birth. Symptoms usually begin between the seventh and tenth days and may be fatal within a few days if extra salt and salt retaining adrenocorticosteroids are not given. Intravenous fluids are essential. The diagnosis is confirmed by raised plasma 17 OH-progesterone concentrations, high plasma potassium, and low plasma sodium concentrations. The plasma electrolyte concentrations are normal at birth and pronounced changes may occur suddenly.

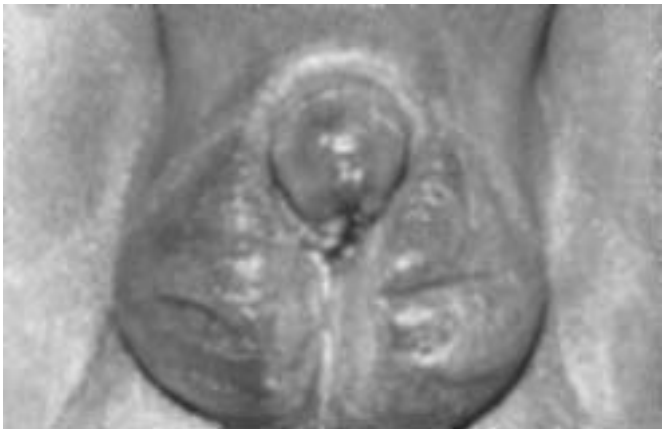


Figure 12.9 Adrenogenital syndrome.

13 Diarrhoea

Diarrhoea is the passage of loose stools more often than would be expected from the diet and age of the infant. It indicates a change in bowel habit. A stool should be examined personally. A rectal examination is often followed by a fresh stool, which can be examined. When diarrhoea is severe, the stools may be mistaken for urine. When this is a possibility, a urine bag should be placed in position and the infant nursed on a sheet of polyethylene.

The stools of newborn infants vary with their diet. The normal stools of breastfed infants are never formed, may be passed at hourly intervals, may contain mucus, and may be green. When lactation becomes established between the third and fifth days intestinal hurry is common, resulting in frequent stools. Later the stools tend to become less frequent and more pasty and by the age of three weeks they may be passed once every two or three days. No treatment is needed. In contrast, the normal stools of bottlefed infants are formed and do not contain fluid or mucus. With certain cow's milk preparations the stools may be dark green, but this has no sinister meaning.

Figure 13.1 Stool of breastfed infant.

Figure 13.2 Changing stool (mixture of meconium and stool).

Acute gastroenteritis

Acute gastroenteritis is an acute infection mainly affecting the small intestine that causes diarrhoea with or without vomiting. The main danger is dehydration and electrolyte imbalance, which may develop rapidly, but the infant may also be very infectious for other infants in a ward or nursery. Gastroenteritis is particularly dangerous to infants aged under two years.

The early signs of dehydration are often difficult to detect, but recent weight loss is often a valuable indicator. Sunken eyes, inelastic skin, and a dry tongue are late signs, but if the infant has not passed urine for several hours severe dehydration is probable. Clinical signs of dehydration are particularly difficult to detect in fat toddlers.

The infant must be examined in detail to exclude any other acute infection.

The rotavirus is the most common cause of gastroenteritis in infants and children throughout the world. It affects every age group and infection easily spreads throughout a family, although the infected adults may have few or no symptoms. Several distinct episodes of diarrhoea can be due to the rotavirus, as there are several serotypes. The incubation period is 24–48 hours and a respiratory illness, including otitis media, precedes the gastrointestinal symptoms in about half the patients. Vomiting which lasts for one to three days is followed by abnormal stools for about five days. Treatment is aimed at keeping infants well hydrated until they recover spontaneously. The frequency of the stools is reduced by dietary treatment, but the abnormal consistency of the stools persists for up to a week.

If infants are given an antibiotic early in the illness (for example, when acute otitis media is suspected as the primary diagnosis) the subsequent diarrhoea may be attributed to the antibiotic rather than to the rotavirus infection. Other drugs (for example, iron) may be associated with diarrhoea.

Box 13.1 Dangers of acute gastroenteritis

- Dehydration
- Electrolyte abnormalities
- Cross infection

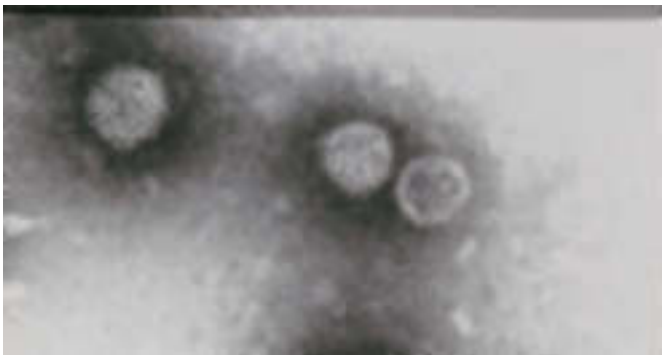


Figure 13.3 Rotavirus.

Management

Clinical signs of severe dehydration or the loss of 5% or more of body weight are definite indications for admission. If infants relapse after treatment or social problems prevent them being

treated at home they may need to be admitted. Infants who vomit persistently usually need to be admitted, though mild symptoms may be managed at home by giving small volumes of liquid by mouth every hour.

In mild cases the main principle of management is to stop cow's milk and solids and give a glucose or sucrose solution orally. After 24 hours fruit or vegetable purées may be introduced and then other items from the child's normal diet. Cow's milk and cow's milk products are re-introduced gradually after the first 24 hours of treatment. Vomiting may be reduced by giving small volumes of fluid frequently. The child should be allowed to drink as much as he wants but needs at least the volume shown in Table 13.1.

Rarely, a breastfed infant has gastroenteritis, but the symptoms are usually mild. The appropriate volume of rehydrating fluid is given by bottle or spoon before each breastfeed.

Kaolin should not be prescribed as it deflects the mother's attention from the main treatment. No antibiotics should be given to children with gastroenteritis treated at home.

The ideal oral rehydrating fluid is a glucose–electrolyte mixture, but a 4% sucrose solution is easily available and safe. Single dose sachets of glucose–electrolyte powder (Dioralyte) or glucose–sucrose–electrolyte powder (Rehydrat) are available, which enable mothers to make up the mixture accurately at home. A safe alternative is 4% sucrose solution, which can be made up by the mother using two level teaspoonfuls of granulated sucrose in 200 ml water. *It is dangerous for mothers to add salt to this mixture.*

In severe cases of dehydration or persistent vomiting oral fluids must be replaced with intravenous fluids in infants admitted to hospital. During the next day the fluid requirement should be given as an oral glucose–electrolyte mixture, alternating with full strength cow's milk formula, and fruit and vegetable purées are introduced. Most children are discharged from hospital on normal diets within a few days of admission.

Infants in hospital with diarrhoea must be barrier nursed in a cubicle, which should ideally be in an annexe to the children's ward.

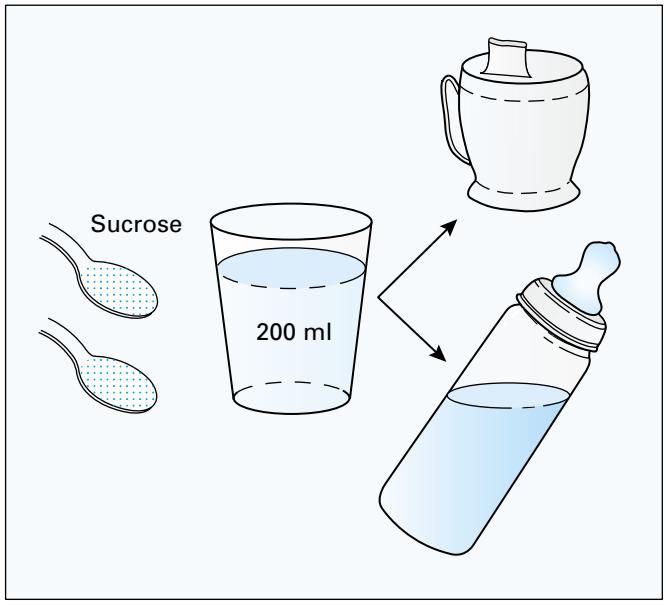


Figure 13.4 Daily fluid intake requirement.

Table 13.1 Fluid intake in relation to weight

Baby's weight		Daily fluid intake	
kg	lb	ml	fl.oz
Under 4	Under 9	500	18
4	9	600	21
5	11	750	26
6	13	900	32
7	15	1050	37
8	18	1200	42
9	20	1350	48
Over 10	Over 22	1500	53

Investigations

Ideally a stool should be sent to the laboratory for detection of pathogens, but this is not necessary for mild cases treated at home. Only a small proportion of children have bacteria such as Campylobacter, Salmonella, Shigella, or pathogenic *Escherichia coli* isolated from their stools. Cryptosporidium, a protozoon which can be seen by light microscopy, is a common pathogen. Most cases of gastroenteritis in children are caused by viruses, usually rotavirus, and can be identified by an ELISA slide test.

Children needing intravenous fluids should have their plasma electrolyte, bicarbonate, and urea concentrations measured urgently.

If two or more infants in a ward or nursery have diarrhoea at the same time cross infection should be presumed, even if their stool cultures show no pathogens. Stools from all the infants on the ward should be sent for culture and tests for rotavirus. Admissions to the ward may have to be stopped.

Box 13.2 Investigations

- Stool culture
- ELISA rotavirus stool test
- Urine microscopy and culture
- Plasma sodium, potassium, and urea estimations

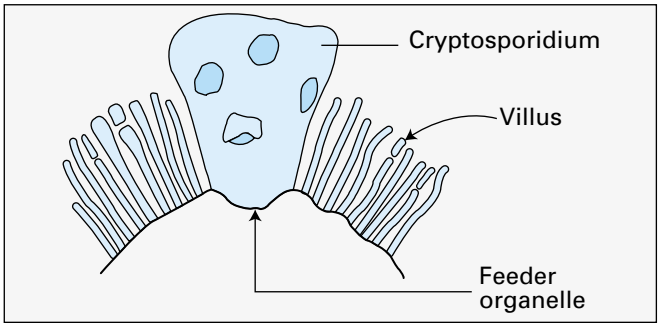


Figure 13.5 Cryptosporidium is intracellular, separated from rest of cytoplasm by a feeder organelle.

Progress

The infant should be reviewed again by the doctor within 24 hours of starting treatment to ensure that the illness is resolving, the infant is not losing too much weight, and the

mother understands the management. Severe dehydration can occur within a few hours and it is helpful to have a specific policy to ensure adequate follow-up visits.

The main cause of relapse or persistent symptoms is failure to follow a plan of treatment. Social problems may indicate the need for admission to hospital. A few infants aged under two years have temporary mucosal damage and are intolerant to cow's milk or other foods. This causes the diarrhoea to persist for longer than two weeks and is considered on p. 69.

Box 13.3 Causes of relapse

- Failure to follow plan
- Temporary intolerance to cow's milk protein

Gastroenteritis in developing countries

In developing countries the continuation of breastfeeding during attacks of gastroenteritis may be essential for survival. Although infants who are completely breastfed rarely have severe gastroenteritis, weaning foods made up with water may infect a breastfed infant. These infants can be managed by continuing the breastfeeding and supplementing the fluid intake to prevent dehydration until the infant spontaneously recovers. Supplements may be given by mouth in mild cases and intravenously in severe cases. An easier method is to give them by continuous intragastric infusion, for which the fluid does not have to be sterile.

Oral rehydrating fluids can be made up using specially designed spoons to measure the sugar and salt. Mothers and older siblings can be taught to use this mixture at the beginning of an episode of diarrhoea rather than wait until the child is dehydrated. Simple slogans such as "a cup of fluid for every stool" are effective.

Figure 13.6 Spoons to measure sugar and salt for rehydration fluid.

14 Mother–infant attachment

There is no specific sensitive period immediately after delivery that is the only time that emotional attachment can form. Any important effects of separation do not persist beyond days or weeks after reunion. Parents who have read obsolete publications may need reassurance that early separation is not followed by permanent damage to mother–infant relationships. A comparison of preterm and term babies indicated that parents find the behaviour of preterm babies less predictable and more frustrating. Though frequent and sustained contact with a sick or preterm baby may increase the degree of parental anxiety, this anxiety induces a feeling of involvement in the care of their baby.

A biological need

Mothers and infants need a close attachment to give the infant the degree of security necessary for optimal emotional and physical development.

Most mothers have strong maternal feelings, which enable them to achieve a firm bond of affection with their babies without difficulty, even after an initial period of separation. The strength of their maternal feelings probably depends on the quality of mothering they received in infancy.

In contrast, a mother may be unable to achieve this attachment without close contact with her baby and even then, it may take a few days before the baby appears to her to be an individual and her own. Failure to form a normal attachment probably accounts for the higher incidence of "battered babies" among preterm infants and among the infants of mothers who were themselves deprived of maternal care.

Forming an attachment

Attachment occurs in five main ways. Firstly, infants can follow the mother's eyes immediately after birth and this eye to eye contact is an important factor.

Secondly, a mother left with her naked infant touches each part of the baby's body with her fingertips.

Thirdly, during the first few days after delivery mothers often appear to overprotect their infants and become overanxious about crying and minor difficulties, such as those of feeding.

Fourthly, even in the first days of life babies mimic the facial expressions of others and can, for example, put out their tongues at them, providing "feedback".

Lastly, physical contact during breastfeeding and the presence of the baby next to the mother throughout the entire 24 hours also promote attachment.

Separation

Separation is sometimes unavoidable if, for example, the infant has to be transferred to another unit for surgery or the mother is receiving heavy sedation for hypertension. The mother is equally separated from her child when she is severely depressed, but this may not appear so obvious.

In the past many babies were separated from their mothers for reasons that would not now be acceptable. Owing to doctors' anxiety, infants with jaundice were admitted to special care units even though they needed no special care.

Figure 14.1 A newborn infant looking at mother.

Figure 14.2 Maternal feelings are related to the quality of the mother's own mothering.

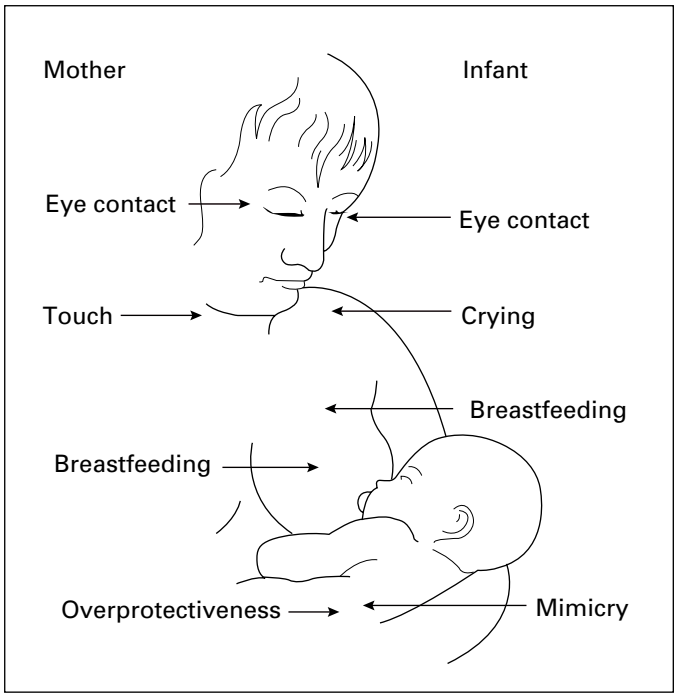


Figure 14.3 Factors in attachment.

The number of nursing staff available on individual maternity wards will determine whether it is safe for babies to remain with their mothers, but no infant should be separated from his or her mother without good reason.

A postnatal ward should be staffed with midwives who have special experience in nursing babies of low birthweight. This allows babies needing tube feeding and other forms of special care to remain with their mothers and avoids the separation resulting from admission to the neonatal unit. The care of healthy but low birthweight infants in the same environment as their mothers is called transitional care to indicate that they need more support than normal term infants.

Difficulty in forming attachments

Some aspects of the history may suggest that a mother may have difficulty in forming an attachment to her infant and will need special help from nursing and medical staff. Particularly vulnerable are mothers who had poor maternal care in their own childhood or who have had a request for an abortion rejected or unmarried mothers who have not decided whether they want the infant to be adopted. If a previous infant was stillborn or died in the neonatal period or a close relative has recently died, help may also be needed. Special attention should be given to a mother under 17 or over 35 years of age and having her first baby.

When the baby is born, the mother may refuse to handle or feed him or her and be more concerned with her own minor symptoms than the infant's care. She may feel detached and the infant's problems may appear to her to be more serious than they are. Similar symptoms may be the first indication of a severe depression in the mother and she may need psychiatric help.

Encouraging attachment

To encourage attachment the principle is to avoid unnecessary separation. Unless infants require special nursing they should be given to their mothers in the labour room, even if an abnormality such as Down's syndrome is present. Newborn infants are able to feed moments after birth even if the mother has received sedation and the mother should be encouraged to put the baby to her breast if she intends to breastfeed. Infants should remain with the mother throughout the 24 hours of the day and be taken out at night only if they continually disturb the other mothers in the room. Breastfeeding should be actively encouraged, though this may take time and perseverance by both mothers and nursing staff. Mothers of preterm infants should be encouraged to express their milk and thus feel that they are actively contributing to the infant's welfare.

In neonatal units the atmosphere should make mothers feel that they are welcome at any time and they should be encouraged to look at, touch, change, feed, and later breastfeed their infants. When the mother is about to visit a very sick infant the reasons for the use of special apparatus should be explained beforehand. A mother can usually visit her infant in the neonatal unit the day after a caesarean section by being wheeled in a chair.

If a mother fails to visit her infant for long periods after she has been discharged, an inquiry should be made whether any remediable reason, such as lack of transport, is responsible.

Before the infant is discharged from the unit the mother should remain with her infant in a room on the unit for at

Figure 14.4 Separation may be necessary for medical reasons.

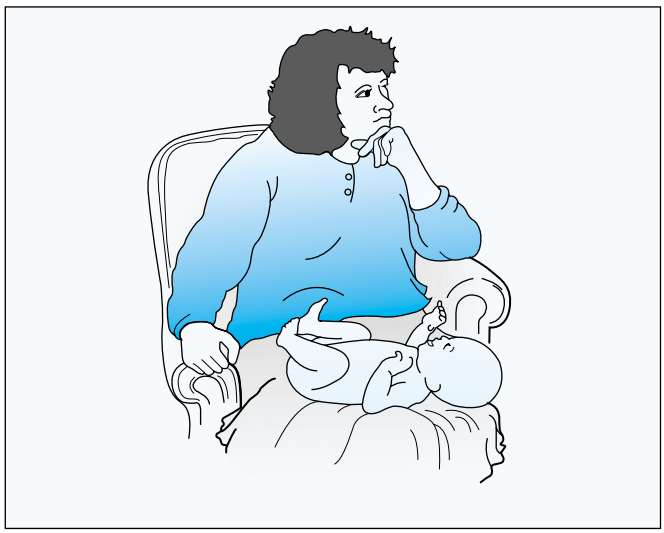


Figure 14.5 Detachment due to depression.

Figure 14.6 Parents should be encouraged to care for their infant in the neonatal unit.

least 24 hours, but longer if possible. This gives her an opportunity to gain confidence in her ability to cope with her baby, who recently appeared to be so fragile and needing expert nursing care to survive.

What the newborn baby can do

Many mothers, especially those having their first baby, believe that newborn babies are blind until they are six weeks old. It is not surprising that these mothers treat their babies like inanimate dolls and feel ashamed when an unexpected visitor catches them talking to the baby. The normal newborn infant can see, hear, and appreciate pain immediately after birth, even if the mother has received heavy sedation. During the hour after birth the infant is often wide awake, looking round for a feed before going to sleep for a few hours.

The distance between the eyes of the mother and the infant when the mother is breastfeeding is the distance at which infants can best focus on an object. This eye to eye contact provides the first means of communication between mother and infant and is probably the reason why mothers find that covering the infant's eyes during phototherapy disturbs them. Mothers of blind infants have difficulty in feeling close to their infants. Newborn infants will become alert, frown, and gradually try to focus on a red object dangled about 30 cm before them. They stare intently at the object and will follow it with short jerking movements of the eyes if it is moved slowly from side to side. Infants are also sensitive to the intensity of light and will shut their eyes tightly and keep them shut if bright light is turned on. They can discriminate shapes and patterns and the arrangements of lines from birth. They prefer patterns to dull or bright solid colours and look longer at stripes and angles than at circular patterns.

Newborn infants can hear. They respond to sound by blinking, jerking their limbs, or drawing in breath. They may stop feeding. Mothers often speak to their infants in a high pitched voice and an infant responds more consistently to his mother's than his father's voice. An infant of three days of age shows preference for sweet and dislike of bitter flavoured fluid. At about the same age he can differentiate smells and distinguish between his own mother's and other mothers' breast pads.

Analysis of sound films shows that both listener and speaker are moving in time to the words of the speaker, creating a type of dance. For example, as the speaker pauses for breath or accentuates a syllable, the infant may raise an eyebrow or lower a foot. The mother notices these changes and this may encourage her to continue speaking. At a few weeks or even a few days after birth the infant may mimic gestures such as tongue protrusion, lip protrusion, or opening the mouth.

If the doctor talks to babies while examining them, mothers will not feel foolish when they do it themselves.

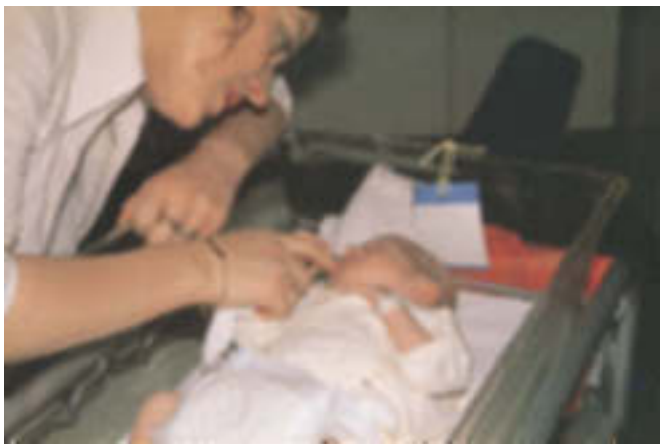


Figure 14.7 Talking to and touching the baby.

Figure 14.8 Newborn infant looking at the camera.

Figure 14.9 Infant looking at mother's face.

15 Growth and growth charts

Weight and head circumference

The size of the normal infant at birth is determined mainly by the mother's size. The greatest changes to adjust for other hereditary factors take place during the first three months of life. Although children tend to attain a stature between those of their parents, some children take after one parent, grandparent, or an even more remote member of the family. All these children are perfectly healthy. It may be difficult to distinguish these normal adjustments of growth from disease, especially during the first few months of life. Growth charts are therefore essential for diagnosis during this period and helpful in discussions with parents.

Length (or height) is difficult to measure accurately in a very young baby and for some years we have used the head circumference as a reference measurement for comparison with the weight. The use of the head circumference in this way is valuable only in the first two years of life. The head circumference is measured round the occipitofrontal circumference (the largest circumference), which should be determined with a disposable paper tape measure. Linen tape measures stretch and produce inaccurate results.

Figure 15.1 Measuring the head circumference.

Normal growth and small normal infants

An infant usually has a similar centile at birth for both head circumference and weight. Children of large parents tend to be towards the 98th centile and those of small parents near the 2nd centile. Most of these children remain on the same centiles for the rest of their lives. Plotting these measurements on a chart in routine clinics helps to confirm that the child is receiving adequate food and is growing normally.

One of the most common problems in paediatric outpatient clinics is short parents who think that their infant does not eat enough. Plotting growth measurements on a chart from measurements already recorded at the local clinic confirms to the doctor and the parents that the child is normal and that the child's final size will be similar to that of the parents (chart A). Charts are better than cards for recording growth.

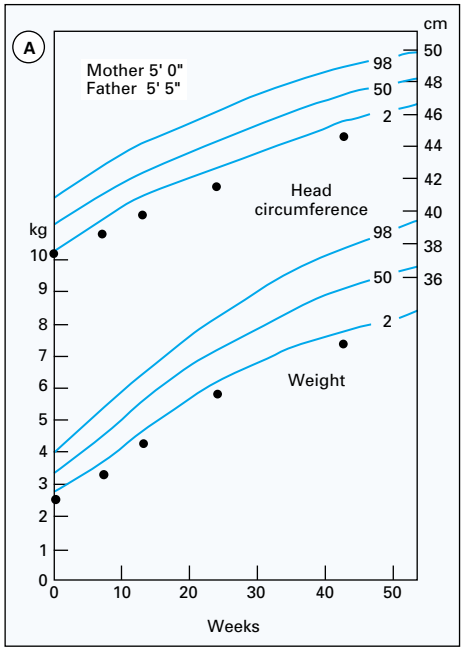


Chart A Small normal infant.

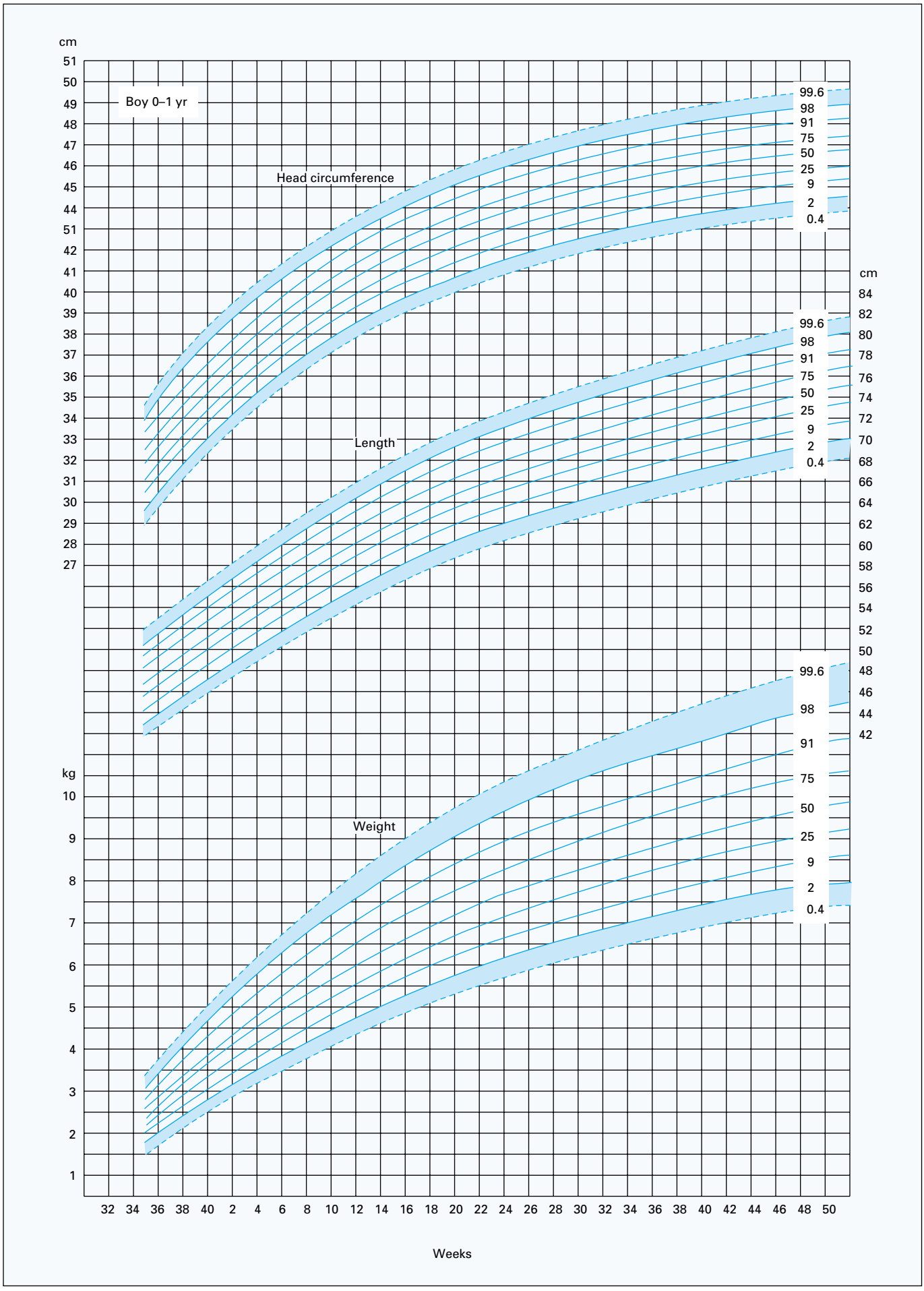


Figure 15.2 Growth chart.

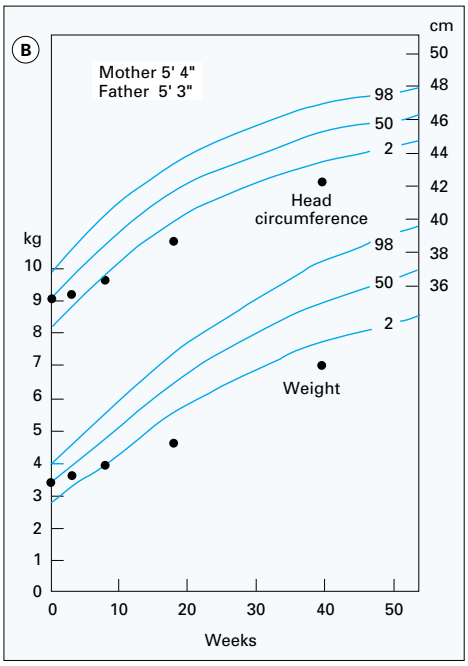


Chart B Taking after father.

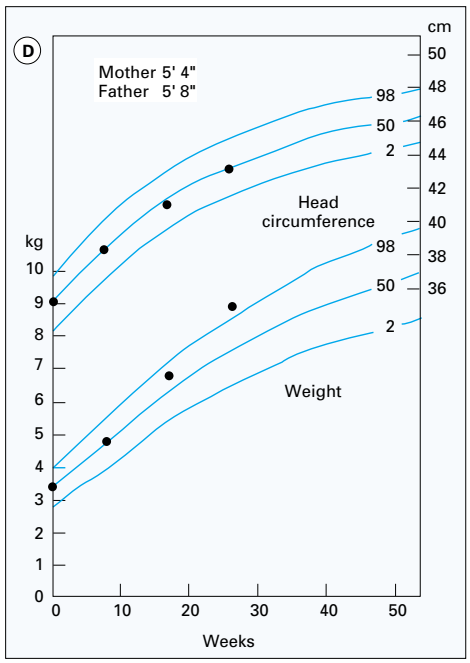


Chart D Obesity.

Taking after father or mother

Some children are more similar to one parent than the other in final size. If only their weights are recorded, they may appear to be either failing to thrive if the father is small (chart B) or gaining weight excessively if the father is tall (chart C). If both weight and head circumference are plotted on a chart, the weight centile line and head circumference centile line can be seen to be running in parallel. In other words, the whole of the child's size is approaching that of a particular parent.

Getting fat

Growth charts give an early warning of obesity. Chart D shows that at the age of 20 weeks this child's weight centile started deviating upwards although the head circumference centile remained the same. If the mother had been shown the growth chart at the age of 20 weeks she might have been able to prevent the phenomenon seen at 30 weeks.

The standard growth charts are based on data collected when the majority of babies were fed with cow's milk preparations. Some normal breastfed babies gain weight rapidly in the first 10 weeks and then more slowly during the subsequent 20 weeks (chart E). This is a normal growth pattern and is not an indication of obesity.

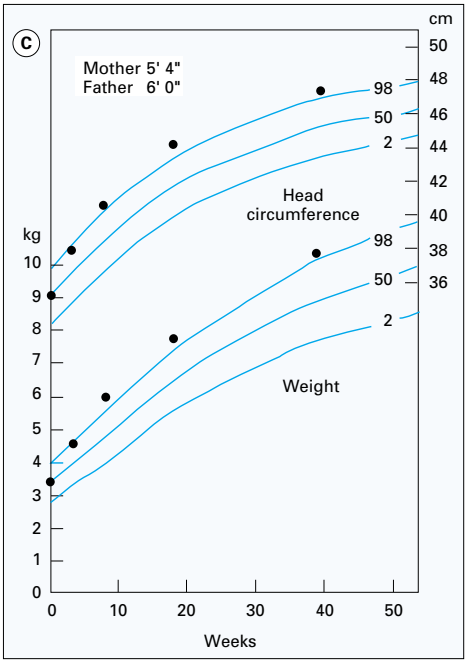


Chart C Taking after father.

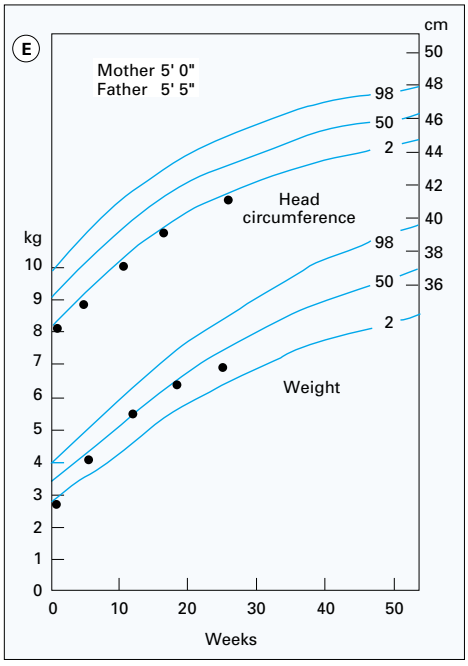


Chart E Normal breastfed infant.

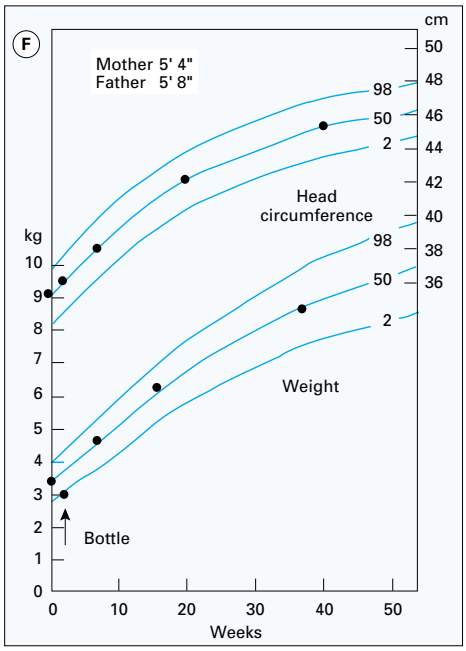


Chart F Poor lactation.

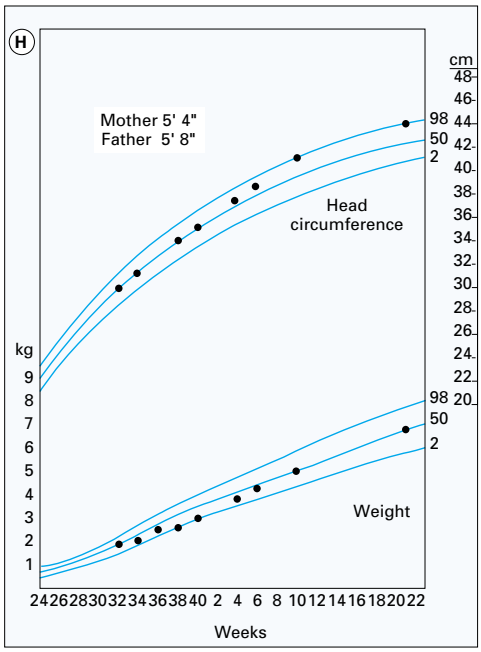


Chart H Catch-up growth in preterm infant.

Poor lactation and failure to thrive

The increased incidence of breastfeeding has resulted in a few infants receiving insufficient milk as a result of their mothers' not recognising poor lactation. This can be detected at an early stage by weighing infants regularly, although normal breastfed infants may not regain their birthweights before the age of two weeks. The onset of full lactation is extremely variable and this must be taken into account when considering the adequacy of weight gain (chart F). Plotting the infant's weight on a growth chart is the first investigation required to diagnose whether failure to thrive is present and also to give a presumptive diagnosis of the cause (chart G).

Preterm

When preterm infants reach home, they may gain weight rapidly, when they will cross the centile lines for both weight and head circumference in parallel (chart H). Preterm infants may have a relatively large head measurement because of the head's discoid shape, but continuing growth on the same centile line shows that the head is normal. Babies of extremely low gestational age, especially below 26 weeks, with chronic lung disease may continue to grow poorly despite feeding well.

Infants whose growth *in utero* has been restricted due to poor nutrition related to placental function may have catch-up growth or may continue to grow poorly postnatally (chart I).

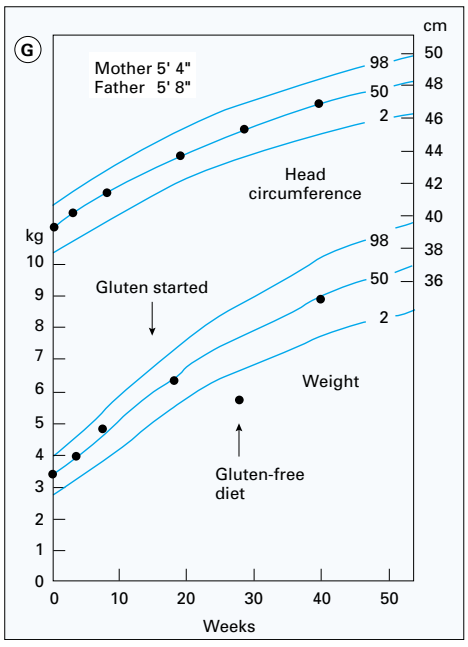


Chart G Coeliac disease.

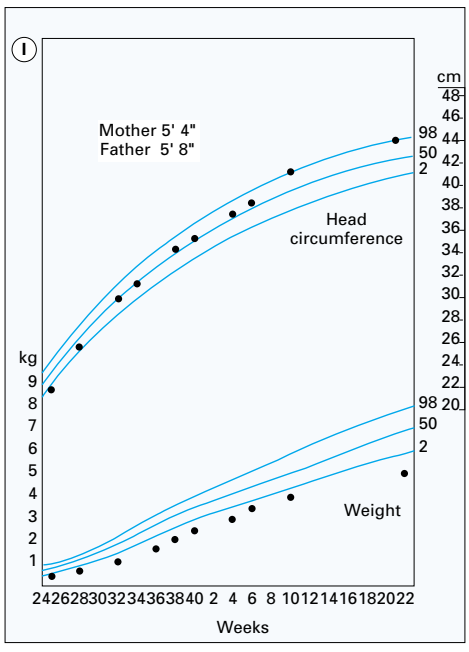


Chart I Poor weight gain and discoid head in preterm infant.

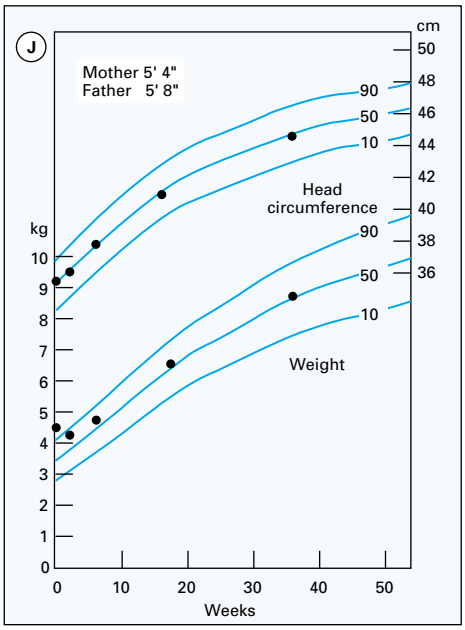


Chart J Infant of diabetic mother.

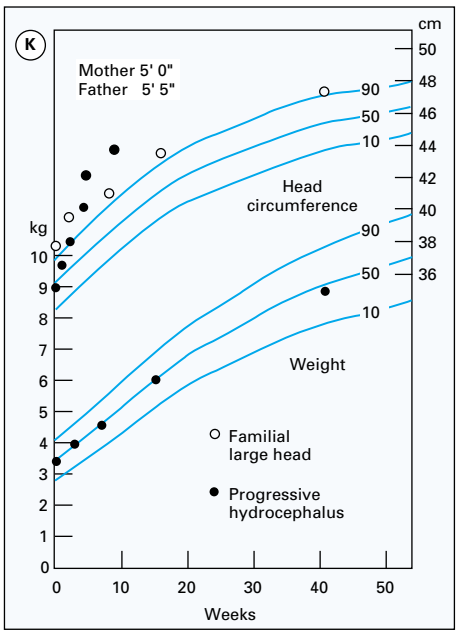


Chart K Progressive hydrocephalus.

Infant of diabetic mother

The infant of the diabetic mother may be grossly obese at birth but slims down within a few months of birth (chart J). These infants may not gain any weight at all for the first two months of life and the mother may be accused of starving the infant.

Progressive hydrocephalus

Progressive hydrocephalus can be confirmed by showing a head circumference centile that progressively increases while the weight centile shows no change (chart K). It may occur in association with intraventricular haemorrhage in preterm infants or myelomeningocoele. Hydrocephalus can be confirmed by ultrasound examination of the brain if the anterior fontanelle is still open. This condition must be differentiated from familial large head, where the head circumference centile remains constant and is parallel to the weight centile.

16 Feeding and feeding problems

There has been no significant change in the incidence of breastfeeding in Great Britain since 1980. In 1990, 63% of mothers breastfed compared with 64% in 1984 and 65% in 1980. There is a steep gradient in the incidence of breastfeeding among mothers of first babies from 50% in social class V to 89% in social class I. The incidence of breastfeeding among mothers having their first child was 57% for those who left full time education at the age of 16 years or under compared with 93% among mothers who left school at the age of 19. In 1990 one third had stopped breastfeeding by six weeks. The proportion of breastfed babies receiving additional bottles at six weeks has been rising since 1980, when it was 28%, to 34% in 1985 and 39% in 1990.

Full term infants usually regain their birth weight between the seventh and tenth day, and thereafter the infant should gain about 20–40 g/day for the next 100 days. Unmodified milk ("doorstep" milk) should not be given until after the age of one year and should be accompanied by vitamin supplements, particularly vitamin D, until the age of two years. Progress on a growth chart is the best guide to ensuring that an infant is receiving the correct amount of milk.

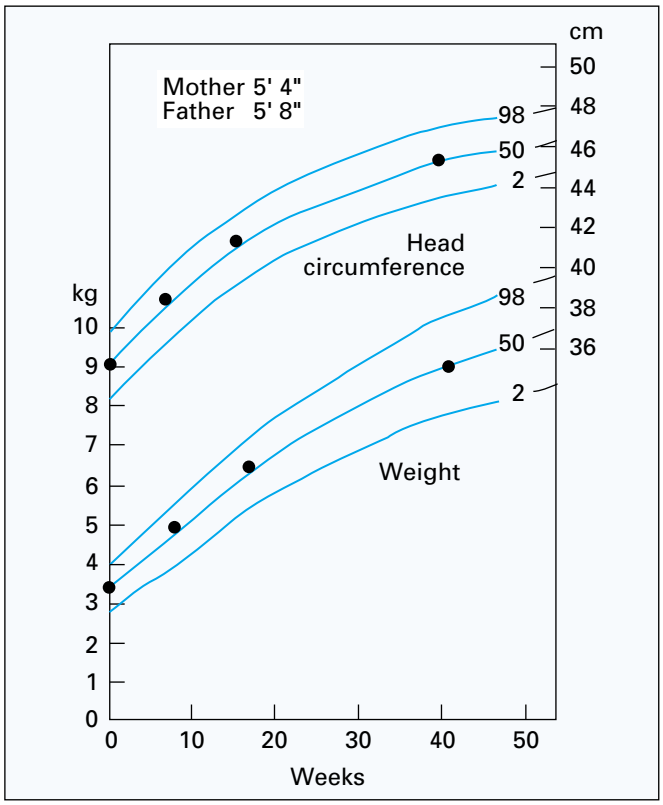


Figure 16.1 Normal growth chart.

Breastfeeding

Breastfeeding should be encouraged. Firstly, the fat and protein of human milk are more completely absorbed than those of cow's milk. The composition of human milk varies during a feed and these subtle changes cannot be mimicked by cow's milk preparations. The significance of these changes is unknown but may be related to the control of intake by appetite.

Secondly, although the fat composition and therefore the fatty acid composition of breast milk vary during a feed, these changes cannot be replaced exactly by cow's milk preparations and the differences in body composition resulting from these different milks may have a long term effect.

Thirdly, human milk contains antibodies and iron binding protein (lactoferrin) which may protect the infant against infections. Gastroenteritis is rare in breastfed infants.

Breastfeeding also plays an important part in mother–infant attachment. If the mother is encouraged during the antenatal period to expect to be able to breastfeed her baby and eventually to enjoy it, she is likely to accept early difficulties with patience and understanding. The close contact and intimacy, and often supreme enjoyment, of breastfeeding provide the best "feedback" between the baby and the mother. The mother's intimate personal relationship with her baby is something that she has to work out for herself. Many feel insecure and inadequate at first and are only too glad to change to bottle feeding whenever the slightest difficulty arises. The attendant should resist such requests and instead use sympathy, understanding, and skill to encourage the mother to gain confidence in handling her own baby.

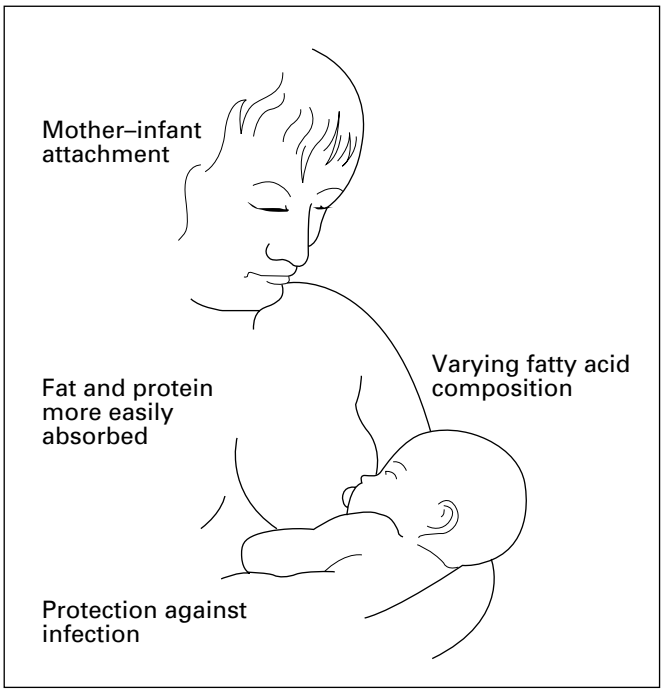


Figure 16.2 Advantages of breastfeeding.

Latching to the breast

The sensation of the nipple against the palate stimulates the baby to suck. Placing the baby in the correct position encourages successful feeding and avoids damage to the nipple. The baby's chin must drive into the breast to enable

the nipple to reach the palate, so the baby needs to put his head back and up. If the baby's head becomes too flexed, the nipple touches the lower jaw and the tongue and the nose is too close to the breast. Helping the chin to thrust forwards and the head to tilt back is hindered by pressure on the back of the head but improved by supporting the baby's back at shoulder level, with the baby facing the mother, chest to chest.

Rooming in

A normal infant is put to the breast for a few minutes at each side either immediately or a few hours after birth. Only a small amount of colostrum is obtained but sucking by the infant stimulates the production of more milk. Some infants are reluctant to take the nipple initially and the mother needs strong reassurance that this is common.

During the first week, and probably later, most of the feed is obtained within four minutes in some babies. Thus the length of time that the baby is on the breast bears little relation to the amount of milk received by the infant. Some infants take a shorter time and others a longer time to take a full feed.

The most satisfactory method of breastfeeding is "on demand". Babies commonly feed every two or three hours during the first few weeks and these frequent feeds are a powerful stimulus to lactation. This is the main advantage of rooming in, where the infant's cot is by the side of mother's bed and she can pick him up and feed him when he cries. Mothers should be encouraged to do this. Infants initially fed on demand usually settle down to a regular schedule after a few weeks. Most breastfed infants feed three hourly rather than four hourly. After the seventh day, if the infant appears to be hungry after a feed or is progressively losing weight test feeding may be considered.

Test feeding

Test feeding should be avoided whenever possible, as it can create anxiety in some mothers. The infant is weighed, without changing the napkin or clothes, before and after each feed during a 24 hour period. If the feed is deficient, putting the baby to the breast more frequently may stimulate increased lactation in the early days or the deficit may be made up with a bottle feed given after feeding from both breasts (complementary feeding). Fortunately, a well nourished full term infant can tolerate a degree of underfeeding without harm for a few days.

Complementary feeds are rarely necessary in the first five days, and after this they should be used only when absolutely necessary. The feel of the bottle teat is quite different from that of the nipple, and when the infant is accustomed to one it may be difficult to persuade him to take from the other.

Contraindications to breastfeeding

There are few contraindications to breastfeeding. Some women have a revulsion to the idea and it would be a mistake to try to persuade them, but psychiatric illness in the mother may be aggravated if the baby is taken off the breast. Severely cracked nipple is a temporary contraindication to feeding from the affected breast, but the milk should be expressed with a pump. Feeding should continue from a breast with acute mastitis while the mother is receiving an antibiotic; it should also continue if the nipple is only mildly cracked.

No drugs should be taken by a lactating mother unless there are strong clinical indications. Most drugs that are essential for the mother are secreted in the milk in

Figure 16.3 Breastfeeding.

Figure 16.4 Rooming in.

Figure 16.5 Test feeding.

Box 16.1 Contraindications to breastfeeding

- Maternal dislike
- Severely cracked nipple
- *Rarely,* drugs

insignificant amounts, so breastfeeding should not be stopped unless there is a special reason. Antibiotics are excreted in minute amounts in the milk but there is the theoretical possibility of sensitising the infant. Warfarin, senna, barbiturates, phenytoin, digoxin, steroids, antacids, and occasional doses of paracetamol pass into the milk in unimportant amounts. Kaolin is not absorbed by the mother. Oestrogens in oral contraceptives may reduce lactation, but the progesterone-only pill is an effective contraceptive and has no effect on lactation. A mother receiving carbimazole may continue to breastfeed provided that the infant's plasma thyroxine concentration is monitored.

Mothers receiving radioactive antithyroid treatment or cytotoxic drugs should not breastfeed. Lithium given to the mother may cause hypotonia, hypothermia, and episodes of cyanosis in a breastfed infant.

Box 16.2 Problems with breastfeeding

- Intestinal hurry: *no treatment*
- Cracked nipple: *breastfeed or rest*
- Breast engorgement: *expression*
- Gulping: *expression*

Problems with breastfeeding

On the fourth or fifth day, when there is a plentiful supply of breast milk, the infant may take up to eight feeds or more a day. Intestinal hurry and frequent loose green stools are common at this stage. Conversely, some normal breastfed infants pass stools only once a week when they reach the age of a few weeks, and this also requires no treatment.

Cracked nipple is usually due to malplacement of the infant on the nipple so that the whole of the pigmented area is not in the infant's mouth. Pulling the infant off the breast abruptly is another cause. If the nipple is slightly cracked, breastfeeding should continue with advice on latching to the breast. If the crack is severe, the infant should be taken off that breast for a day or so and a bland ointment, such as lanolin, placed on the nipple every few hours. A fissure of the nipple occurring after the first few weeks is often aggravated by thrush infection of the baby's mouth and the mother's nipple. It is best treated with miconazole or nystatin gel to both until four days after it appears to be clear.

In acute mastitis there is fever, pain, flushing, and induration in one breast and enlarged axillary lymph nodes. Flucloxacillin or erythromycin is given to the mother for at least a week and feeding continues from both breasts. Fluctuation in the indurated area indicates that an abscess is present and the need for surgical incision and suppression of lactation with bromocryptine.

Feeding on demand usually prevents maternal breast engorgement, which can occur towards the end of the first week of the baby's life. This is easily alleviated by expression after feeding, preferably with a pump. If engorgement has already occurred the help of an experienced midwife is necessary.

During the first feed of the morning milk may spurt quickly from the breast and a ravenous infant may swallow excessive air, which may be regurgitated later with milk. This is often accompanied by severe crying. It can be alleviated by manual expression of the first 30 ml of milk, which can be given to the infant later if necessary.

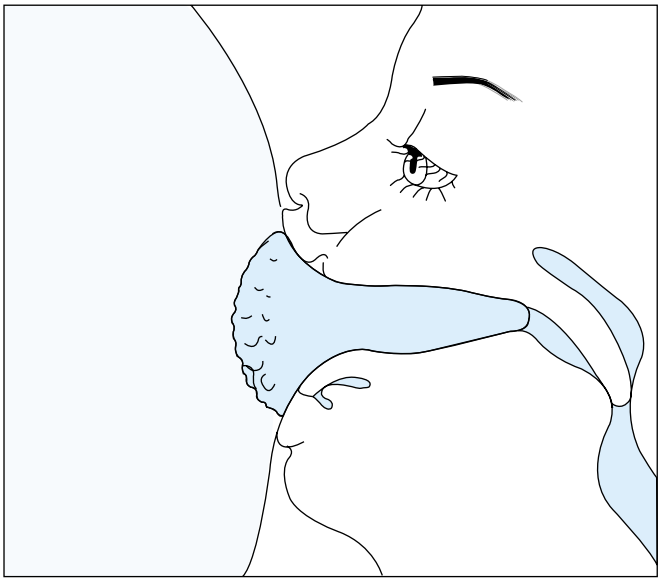


Figure 16.6 Good placement of the nipple in the infant's mouth.

Figure 16.7 Breast pump.

Bottle feeding

All the cow's milk preparations available in Britain have sodium and protein concentrations similar to those of human milk. The powder should be measured accurately, avoiding heaped on packed scoops. The instructions on each packet must be followed. Ready to feed bottles are used in most obstetric units.

If feeds are made up for a 24 hour period they should be stored in the refrigerator. Bottles can be sterilised in dilute hypochlorite solution, but processing in an autoclave is the best method in hospitals. Other methods include steam or a special microwave oven kit. Teats should be well washed before sterilisation.

The size of the hole in the teat should allow individual drops of milk to follow each other quickly when the bottle is inverted, and this should be checked each week.

"On demand" feeding

Feeds are usually given "on demand" or three or four hourly. Most infants need to be fed every three hours. The milk must not be made up to a stronger concentration than that recommended on the packet. Few babies can manage without a night feed for the first few months.

A normal full term infant receives 30 ml of milk per kg body weight during the first day of feeding by bottle. Feeds should be increased by 20 ml/kg each day until a maximum of 150 ml/kg is reached on the seventh day of feeding. Underfeeding causes small amounts of green mucus to be passed frequently, but this is more likely to be found with breastfed than with bottlefed infants.

Problems with bottle feeding

If the hole in the teat is too small the infant may swallow excessive air during the feed and regurgitate it later with milk, accompanied by bouts of crying. It is valuable to observe the rate at which the drops of milk are formed when the infant's bottle is inverted. The drops should follow each other quickly but there should not be a continuous stream. If the hole is too small it may be made larger with a hot needle. If the hole is too large, infants may swallow excessive air as they gulp to avoid choking.

By taking a careful history it is usually possible to determine the likely cause of any symptoms. If growth is poor, infants need more frequent or larger feeds. If the weather is hot and infants are not receiving extra water, they may be thirsty and should have additional water. Mothers tend to use gripe water as a panacea, not realising that it contains bicarbonate, which produces carbon dioxide in the stomach.

Reluctance to feed

In an infant who has fed normally before, reluctance to feed may be a dangerous symptom. It may be due to any severe disease, such as congenital heart disease or a lower respiratory tract infection. On the other hand, when an infant has a mild upper respiratory tract infection the nose may become blocked with mucus making it difficult for the baby to feed. Thrush produces white plaques on the buccal mucosa and tongue, which become sore. It can be treated by a five day course of oral nystatin drops or miconazole oral gel, 1 ml twice a day.

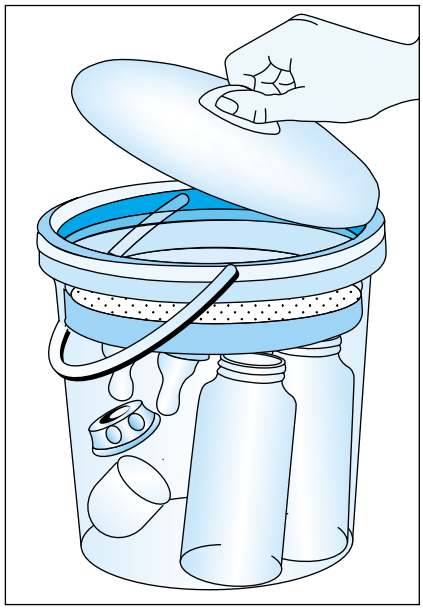


Figure 16.8 Container for sterilising bottles.

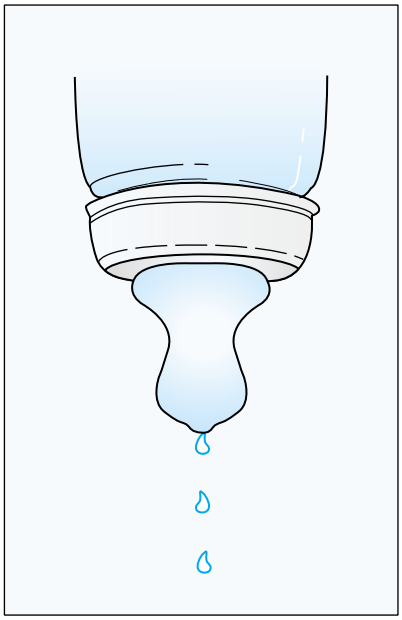


Figure 16.9 Drops should follow each other quickly.

Figure 16.10 Infant refusing a feed.

17 Failure to thrive

Mothers become anxious if their infants do not gain weight at the rate that is expected. The majority of these babies are perfectly healthy but the expectations are too high. Initially, it is essential to determine whether the baby is gaining weight normally. If weight gain is abnormal, failure to thrive is present and a cause should be sought. Some of the material in this chapter is also found in Chapter 15.

Box 17.1

- Is the weight gain normal?
- If not, what is the cause?

Normal weight gain

Small parents

An infant usually has a similar centile at birth for both head circumference and weight. Children of large parents tend to be towards the 98th centile and those of small parents near the 2nd centile. Most of these children remain on the same centiles for the rest of their lives. Plotting these measurements on a chart helps to confirm that children are receiving adequate food and are growing normally. A common problem is the infant of short parents who consider that their child does not eat enough. Plotting growth measurements on a chart from measurements already recorded at the local clinic confirms to the doctor and the parents that the child is normal and that his or her final size will be similar to that of the parents. Charts are better than lists of measurements for seeing the trends in growth.

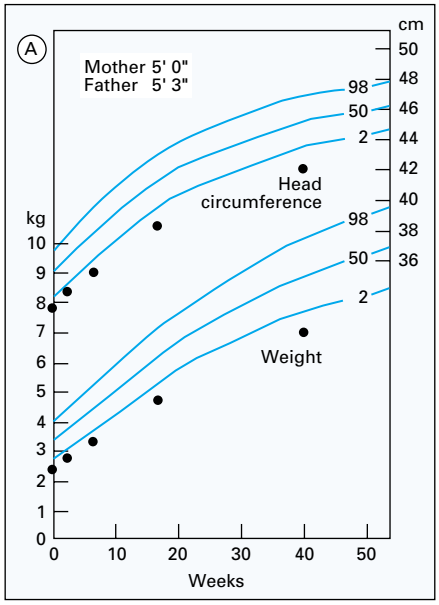


Chart A Normal infant of small parents.

Taking after father or mother

Some children are more similar to one parent than the other in final size. If only their weight is recorded they may appear to be either failing to thrive if the father is small (chart B) or gain in weight excessively if the father is tall (chart C). If both weight and head circumference are plotted on a chart, the weight centile line and head circumference centile line can be seen to be running in parallel. In other words, the whole of the child's size is approaching that of a particular parent.

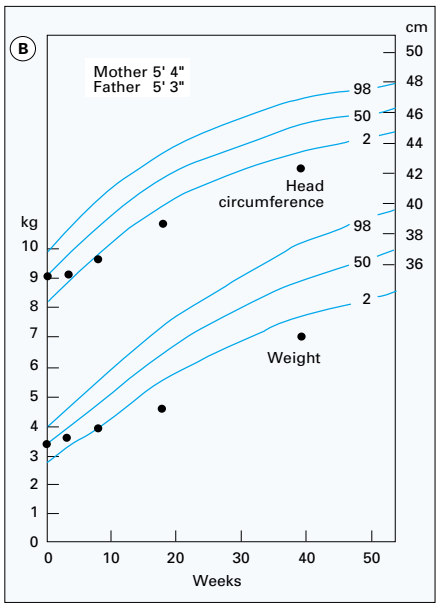


Chart B Taking after father.

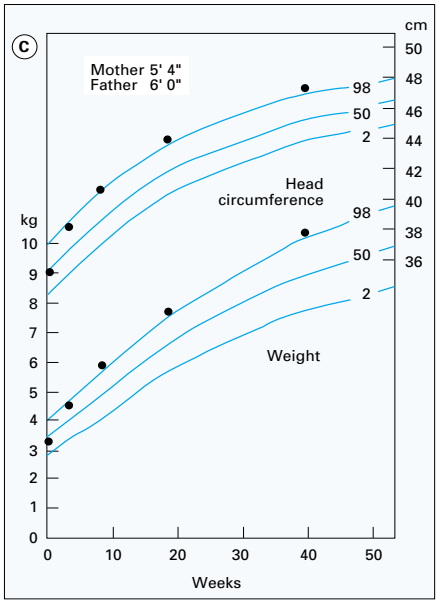


Chart C Taking after father.

Low birthweight

Infants who are of low birthweight may have been born early (preterm) or have experienced intrauterine malnutrition or both. Although some of them have a period of "catch-up" growth during the first few months of life, others remain below the 2nd centile for the whole of their lives. This is considered to be related to the intrauterine malnutrition (see chart D).

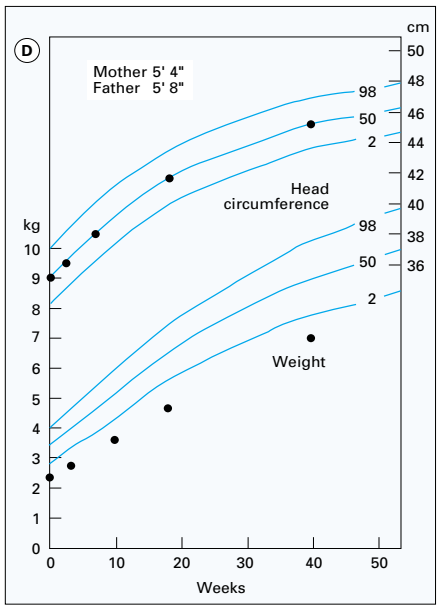


Chart D Low birthweight.

Physical causes of failure to thrive

A progressive fall in the weight centile with a constant head circumference centile is the best confirmation of failure to thrive. Measurements on a single occasion may show a weight centile that is below the head circumference and length centile, but these findings may be present in a normal baby whose family has a body shape that is slightly different from the majority. The plotting of serial measurements on a growth chart before starting to take a history often shortens that process.

Deficient intake of food or the excessive loss from malabsorption or metabolic disease cause failure to thrive. Psychiatric and social problems, which are the predominant factor in the majority of the infants with failure to thrive in this country, probably reduce the food intake.

A detailed history is taken, paying particular attention to the family history, birth, feeding, and maternal anxieties. Details of any vomiting, diarrhoea, or abdominal distension are noted. The complete examination includes an assessment of developmental age and evidence of wasting, particularly of the inner aspects of the thighs and buttocks.

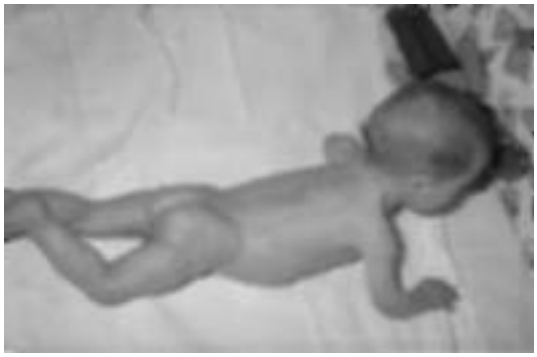


Figure 17.1 Wasting of buttocks.

Figure 17.2 Wasting of thighs.

Deficient intake

Poor lactation – The high incidence of breastfeeding in some areas has resulted in a few infants receiving insufficient milk as a result of their mothers' not recognising poor lactation. This can be detected at an early stage by weighing infants regularly (see Chart E), although normal breastfed infants may not regain their birth weights before the age of two weeks. The onset of full lactation is extremely variable and this must be taken into account when considering the adequacy of weight gain (see Chapter 16)

Problems with bottle feeding – Poor tuition may result in deficient milk intake (see Chapter 16)

Cerebral palsy – Infants with cerebral palsy may present with slow feeding and resulting poor milk intake before changes in the motor function in the limbs can be detected. These infants will usually have some evidence of developmental delay.

Congenital heart disease – There may be deficient milk intake due to slow feeding and there is also a high metabolic rate. Central cyanosis, a murmur, or both are usually present.

Urinary infection – A clean catch urine specimen examined promptly in the laboratory is essential for this diagnosis.

Excessive loss

Gastro-oesophageal reflux – Persistent vomiting from birth may result in failure to thrive if the infant does not take in additional milk to compensate for the loss (see p. 51)

Congenital pyloric stenosis – This condition may present at any time until the age of three months. It should be considered in any baby less than that age who vomits persistently (see p. 50)

Malabsorption – In this country the main causes of malabsorption are cow's milk protein intolerance, cystic

fibrosis, and coeliac disease (see Chart F). These conditions all cause loose stools that are more frequent than normal but may not be recognised by a mother with a first child. A detailed history of bowel function should be taken and personal examination of the stool may be helpful.

Diabetes mellitus – This can be excluded by a negative multistix test for glucose in the urine.

Psychiatric and social factors of failure to thrive

A physical cause for failure to thrive may be present in a family with psychiatric and social problems and therefore a physical cause must be excluded for all infants. The exclusion of a physical cause is often helpful in persuading the parents to accept that psychiatric or social factors are the main reason for the problem. Features in the history that may suggest this possibility include maternal depression, marital discord, or a disorganised household. Maternal food preferences may result in the exclusion of certain foods, such as cow's milk, from the diet without reason and without adequate supervision, and may result in a deficient energy intake.

There are more subtle ways in which maternal emotional factors may affect the infant. The mother may be depressed and tense, and this anxiety is transmitted to the child, who does not feed. The mother then reacts by removing the food. In other families the child may be intrinsically less responsive than the average child to food and this may affect the mother's response to the child at mealtimes. In both these examples the amount of food taken by the child falls to a plateau where the infant appears to be satisfied with the amount given. Another possibility is that the mother keeps to a diet herself as she perceives that she is overweight and also gives a diet to the infant in a similar way.

Some infants are deliberately underfed as part of child abuse. Normal weight gain occurs when fostering has been carried out.

Effective management of infants with this diagnosis requires a team approach involving a health visitor, the family doctor, a child psychiatrist, and paediatrician. Regular advice from a health visitor or dietitian often improves the rate of weight gain.

Investigations

Specific investigations are indicated by the history and examination given above, but if there are no pointers to the diagnosis the following investigations are performed: multistix urine test for glucose, full blood count, sweat test, urine microscopy and culture, and plasma creatinine.

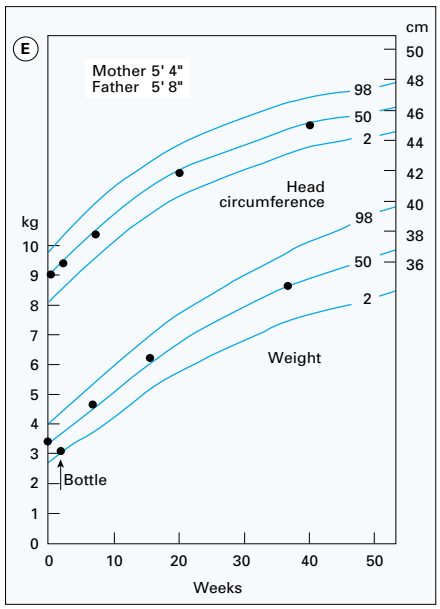


Chart E Poor lactation.

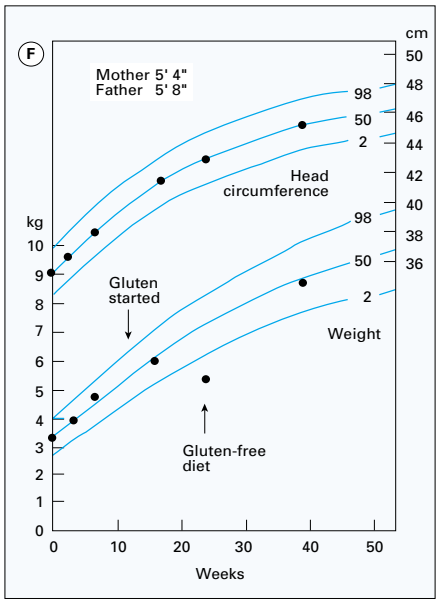


Chart F Coeliac disease.

18 Weaning

Weaning is the process of getting babies used to eating foods other than milk, and using a spoon and cup. For the first four months of life babies will need only milk and additional drinks of boiled water. New foods should start to be introduced at about four months. If solid foods are introduced too early babies may become too fat; if they are introduced too late, after seven months, there may be problems with chewing. Once introduced, solid foods should be given regularly and in gradually increasing quantities. As the amount of solid food increases, the number or size of milk feeds should decrease.

In countries where commercially produced infant cow's milk preparations are safe and affordable, they should be used for the first year of life. Advantages are that the protein and calcium concentrations and vitamin content are more appropriate for babies than unmodified cow's milk.

Mothers need a small, wide plastic teaspoon with no sharp edges, a small plastic cup or bowl, a feeding beaker, and a cotton or plastic backed cloth bib. The feeding equipment should be kept in a sterilising solution.

Firstly, babies have to be taught to take food from the spoon rather than just sucking. As solids are increased, the volume of milk should be reduced. The baby should also be given plenty to drink to replace the milk; at least 100–150 ml of cooled boiled water each day. Diluted fruit juices are not necessary. Most fruit *squashes* are unsuitable for babies.

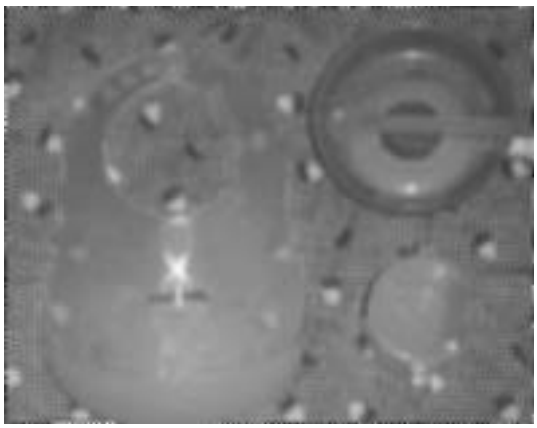


Figure 18.1 Equipment for weaning.

4–6 months

At 4–6 months babies will usually be having breast or bottle feeds at 6–8 am, 10 am, 2 pm, 6 pm, and 10 pm. Solid foods should be introduced to babies before the mid-morning or lunchtime feed – usually a baby cereal or a fruit or vegetable purée.

Mothers should start by using a cereal made from rice and should mix half to one teaspoon of cereal with one to two tablespoonfuls of breast milk, infant formula, or water; it should be given to the baby from the spoon. Sugar, honey or salt should not be added. Wheat based cereals should not be introduced until after the age of six months. At lunchtime babies should be fed with vegetables such as potato or carrot, or fruit such as apple. The vegetables can be fresh or frozen, and the fruit fresh or canned in natural juice. The vegetables should be boiled and the fruit stewed and then liquidised, mashed into purée or sieved. Feeding should start with half or one teaspoonful before or during the lunchtime feed and gradually increased.

At about six months mothers should start to give a wider variety of foods. Banana mashed with a fork is a favourite of many babies. Mashed hard boiled eggs, soups, milk puddings, and strained baby foods should all be tried, with the emphasis on savoury flavours rather than sweet ones. Lightly boiled or scrambled eggs or omelettes should *not* be given, as there is a risk of salmonella infection.

Figure 18.2 Feeding with a spoon while sitting on the mother's lap.

Box 18.1 Menu at four to six months

6–8 am	Milk
10 am	Milk
	Baby cereal
2 pm	Milk
	Fruit or vegetable purée
6 pm	Milk
10 pm	Milk

Preparing food

Home prepared foods are best. The mother knows what is in them and can prepare them easily and cheaply, as she can use

a small portion of food cooked for the rest of the family. Many families have an electric liquidiser or food processor, but a manual blender is equally effective. A sieve is useful for making fruit and vegetable purées. If there is a deep freeze, time can be saved by preparing puréed food in bulk and storing individual portions for up to three months.

Cans, jars, and packets of baby foods are useful but should not replace home made foods. After opening, jars and cans should be kept in a refrigerator and used within 24 hours.

Iron deficiency is common during weaning but can be prevented by an appropriate diet. Iron from vegetables and cereals is not as well absorbed as that from meat. Sources of iron for vegetarians include breakfast cereals and other cereal products, pulses, green vegetables, and nut purée.

6–8 months

When babies are six months old it is no longer necessary to sterilise equipment for weaning. Bottles and teats should be sterilised until a beaker or cup is used, and that should be introduced at about this age.

At 6–7 months a greater variety of tastes and textures should be introduced. A little finely minced chicken, lamb, beef, or liver may be given mixed with potato purée. Vegetables should be cooked until soft, then sieved, blended or liquidised until smooth. White fish can be boiled with milk and, after the bones have been removed, mashed in with potato or vegetable purée. Dahls such as chama dahl and dhudhi, lentils and rice can be given at this age, but should not include salt or hot spices (chilli, ginger, cloves), fat or oil. Soup can be made with lentils or soft chick peas and carrots, cauliflower, or potato, which can then be blended, liquidised, or mashed. Grated cheese, cottage cheese, curd cheese, or paneer may be given, mixed in with other foods, and mashed hard boiled egg can be added to mashed potato. Babies often eat yoghurt plain, but they may need to be tempted with added fruit.

At around seven months parents should start changing the texture of foods to encourage chewing. For example, meat and fish should be less finely minced and potatoes and other vegetables can be mashed rather than puréed. Babies should also be given hard foods such as a crust of bread, a piece of chappatti, pieces of apple, carrot, and banana, which they can hold in their hands, and this will encourage them to chew. These foods should always be given *under supervision* to make sure babies do not choke. Babies should not be given rusks or biscuits every day, as these may encourage the taste for sweet foods.

The timetable of feeds should also start to change at this age and the number of milk feeds should fall.

Babies often reject food. This may be because they are not used to a new taste or texture. However, it may also be because they are thirsty or because the food is too hot, or simply because they want to attract attention. Mothers should be told to keep trying with new foods and to give them in different ways. Babies who become constipated should be given more fruit and vegetables and water.

At about eight months the baby can now sit in her high chair at the table with the rest of the family. Two course meals at lunch and teatime should be introduced – for example, a savoury food followed by a fruit purée. The amount of milk will decrease as more solids are taken. Breastfeeding is continued or an infant formula milk should be given until the end of the first year.

Figure 18.3 Electric liquidiser.

Box 18.2 Menu at six to eight months

6 am	Milk feed
9–10 am	Cereal + milk *or* hard boiled egg Milk feed
1–2 pm	Minced/puréed meat or fish, puréed vegetables and potato, gravy Diluted fruit juice or milk feed
3–4 pm	Diluted fruit juice or water
5–6 pm	Fruit purée/mashed banana Custard or milk pudding Milk feed

Figure 18.4 Sitting in highchair eating with the family.

9–12 months

At this stage babies should be able to eat the same food as the rest of the family, though the food will have to be cut up or mashed. Food should not contain hot spices, particularly chilli, ginger, or cloves, and not too much butter, ghee, or oil. Each day the baby should drink a pint of milk, or less if he or she eats some cheese or yoghurt; two small portions of meat, fish, poultry, or eggs, some fruit, fresh fruit juice, or vegetables; some cereals – for example, a cereal at breakfast and half a slice of bread at tea; and a small amount of butter or margarine. Babies should not be given too many sweet foods, particularly between meals, as they can cause overweight and tooth decay.

Vitamins

Breastfed infants under six months do not need additional vitamins, provided the mother had an adquate vitamin status during pregnancy. From the age of six months infants receiving breast milk as their main drink should be given supplements of vitamins A and D.

Infants receiving 500 ml or more a day of an infant formula milk do not need vitamin supplements. Infants receiving a smaller volume of formula milk or pasteurised whole cow's milk (after the age of one year) should be given vitamin A and D supplements until the age of five years. After the age of one year these supplements may be omitted if the diet has an adequate vitamin content and there is moderate exposure to sunlight.

Box 18.3 Food safety

- Babies should never be left alone when eating, because of the risk of choking.
- Heated foods should be warm but not hot enough to burn a baby's mouth.
- It is not advisable to use a microwave oven to reheat food and drink. This is because "hot spots" can occur in the middle of food if it is not thoroughly stirred after reheating.
- Eggs should be boiled until the white and yolk are solid (about six minutes).
- Peanuts and food containing peanuts or unrefined groundnut oil should not be given to infants from atopic families until they are at least three years of age or to infants who are atopic or allergic to peanuts.
- Whole nuts should not be given to children less than 4–5 years of age because of the risk of choking. However, smooth peanut butter or finely milled nuts are suitable.

19 Review at six weeks

For many years, infants nursed in neonatal units have been routinely assessed at regular intervals as have those who were placed on "at risk" registers for handicaps. Total population screening was introduced when it was realised that many children later found to have disabilities were not on registers or being seen after discharge from neonatal units.

Examination of all infants at specific ages allows abnormalities to be detected at an early stage, provides mothers with a chance to discuss their anxieties, and gives opportunities for education in preventive health.

Developmental examinations should be considered on two levels: the surveillance of an apparently healthy population and the detailed assessment of infants referred from primary examinations because of suspected abnormalities. Primary examinations, which should cover all infants in the district, are best performed by the family doctor. In some districts specially trained health visitors do this work, especially for infants whose parents will not visit the surgery or clinic. Nevertheless, this system is not ideal since health visitors may consider up to 20% of infants abnormal and refer them to an assessment clinic, with all the anxiety that this entails. Some clinics are run jointly by a family doctor and health visitor, who often has wide experience of managing feeding and behaviour problems.

Infants found to be abnormal at the primary examination need to be referred to a doctor with a special interest in developmental assessment. This may be the community or the hospital paediatrician. The more detailed examination performed at this stage will probably need a team of paramedical staff.

A mother who considers that her infant has an abnormality of hearing, sight, or development must be sent straight to the specialist because she cannot be reassured until a detailed assessment has been performed, which usually needs several members of the team. The mother is likely to be right.

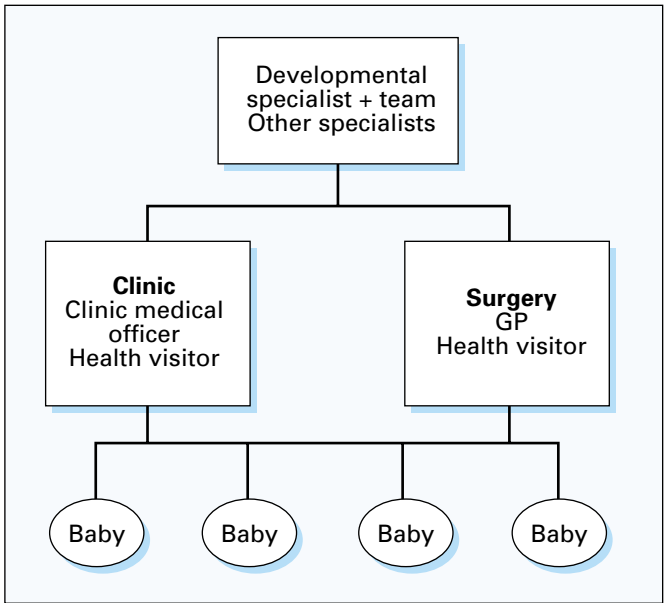


Figure 19.1 Organisation of primary and secondary care.

Primary examination

Infants in the normal population should be assessed within 48 hours of birth, at six weeks, and again at seven or eight months. Preterm infants should be examined about six weeks after the expected date of delivery, when their development should be the same as that of a six week infant born at term.

The examination should take no more than about 15 minutes and should be confined to items that have prognostic significance.

Many districts use a "parent held" manual record that, if completed correctly by professionals and parents, is a complete record of a child's health. All child health surveillance data, the dates of immunisations, and outcomes of visits to general practitioners and hospitals, and contacts with other health professionals can be recorded. Some health advice is also included throughout this record. In addition, the Health Education Authority's book *Birth to Five Years*, which is issued to all first-time mothers, is an excellent manual on child health for all parents (and junior doctors).

The six week examination should include a physical examination and assessment of alertness, vision, and motor function. The checks should be carried out as part of clinical care and not be regarded as pass or fail examinations. A simple form is useful for recording the information quickly. Screening

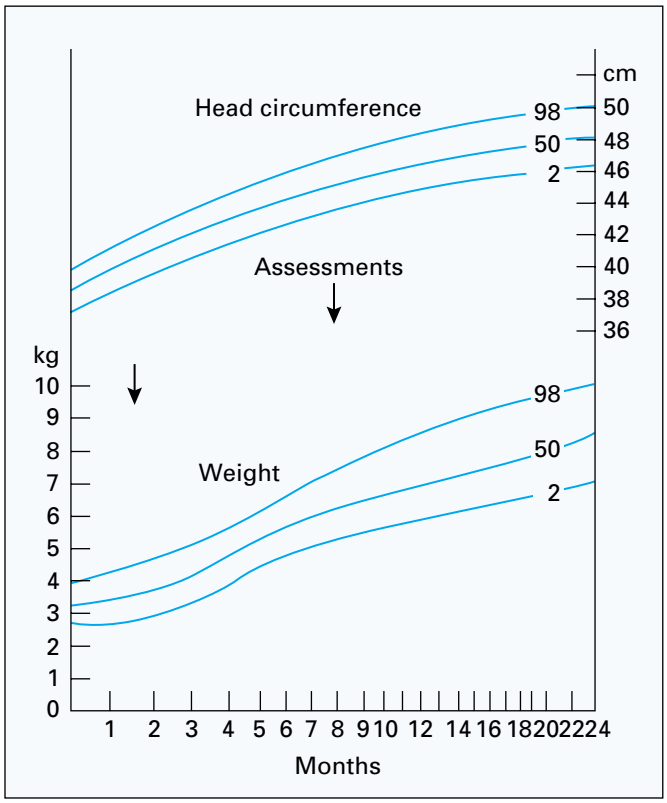


Figure 19.2 Ages for assessments.

6–8 weeks review

Review at 6–8 weeks

For parent to complete:

Health topics for discussion: Tick ☑	
Immunisation	☐
Recognition of illness	☐
Nutrition	☐
Activities to aid development	☐
Dangers: fire, scalds, falls, overheating	☐
Good child rearing practices	☐

This review is done by your health visitor and/or a doctor. Below is a list of things you may want to discuss when you see them. However, if you are worried about your child's health, growth or development you can contact your health visitor or doctor at any time.

Circle 'yes' or 'no' or 'not sure'

Do you feel well yourself?	Yes/no/not sure
Do you have any worries about **feeding** your baby?	Yes/no/not sure
Do you have any concerns about your baby's **weight** gain?	Yes/no/not sure
Does your baby watch your face and follow with his/her eyes?	Yes/no/not sure
Does your baby turn towards the light?	Yes/no/not sure
Does your baby smile at you?	Yes/no/not sure
Do you think your baby can hear you?	Yes/no/not sure
Is your baby startled by loud noises?	Yes/no/not sure
Are there any problems in looking after your baby?	Yes/no/not sure
Do you have any other worries about your baby?	Yes/no/not sure

Comment_____

_____

How are you feeding your baby? Breast/bottle/mixed

Keep hot drinks away from children Use a coiled-flex kettle.

Check the water before you bath your baby. Hot water can scald your baby badly.

Review at 6–8 weeks

* Please place a sticker (if available) otherwise write in space provided.

Surname ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

First names ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

NHS number ☐☐☐☐☐☐☐☐☐☐ Local no ☐☐☐☐☐☐☐

Address_____ Sex M/F

_____Postcode_____ D.O.B.__/__/__

G.P. ☐☐☐☐☐☐☐☐☐☐☐☐ Code ☐☐☐☐☐☐☐☐☐☐

H.V. ☐☐☐☐☐☐☐☐☐☐☐☐ Code ☐☐☐☐☐☐☐☐☐☐

Date of examination _____ /_____ /_____

Age in weeks _____

Name of examiner

☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Examiner code number ☐☐☐☐☐☐☐☐☐☐ Examination centre ☐☐☐☐☐☐☐☐

WEIGHT_____ kg _____centile

HEAD CIRC._____ cm _____centile

FEEDING_____ Breast/Bottle/Mixed

Any previous ongoing medical problems? NO☐ Yes☐

If 'Yes' please specify (1) _____ (2)_____ (3) _____

ICD Code ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐ ☐☐☐☐☐☐☐☐☐☐

ITEM	GUIDE TO CONTENT	CODE (Ring one)*	COMMENT
Physical	Full physical exam. Fontanelles. Palate. Skin	S P O T R N	
Vision	Eyes. Follows. Red reflex.	S P O T R N	
Hearing	Ears. Risk factor. Parents observations	S P O T R N	
Locomotion	Tone, movement, reflexes	S P O T R N	
Manipulation	Moro and grasp	S P O T R N	
Speech/Lang.	Vocalisation	S P O T R N	
Behaviour	Social smile. Cry. Sleep	S P O T R N	
Hips	and skeletal system	S P O T R N	
Genitalia		S P O T R N	
Heart	Femoral pulses	S P O T R N	

*(If one or more codes seem to apply select the last one, e.g. R takes priority over T, T over O etc.)

S = Satisfactory *(normal result)*
P = Problem *(significant condition on record)*
O = Observation *(special recall arranged)*
T = Treatment or investigation underway
R = Referred to any community or hospital service
N = Not examined for this item

Referred to (1)_____ (2)_____
(3)_____

Special recall in_____ wks/mths Signature _____

White copy: Stay in Record Yellow copy: to Child Health Office
Pink copy: to GP/HV

6–8 weeks

Figure 19.3

for hearing defects is not usually done until seven or eight months (see page 80), and other abilities are assessed more sporadically, depending on whether the mother takes the child to a local clinic or to the general practitioner's special baby clinic.

Of greatest importance is the child's alertness, interest in his surroundings, responsiveness, and ability to concentrate.

The physical examination of the infant should always include measurement of weight and head circumference. These values, plotted on a growth chart, will show at a glance whether the infant's growth is normal.

Weighing and measuring

Being undressed and weighed is often the most worrying part of the visit to the child. Weighing the child in vest and napkin reduces upset and time taken. Weights of the vest and napkin should be subtracted from the total weight. Suitable scales that are checked regularly are an essential piece of equipment.

In very young infants it is easier to measure head circumference than length. The measurement should be made, using a paper tape measure, around the occipitofrontal circumference (the largest circumference). Measurements can then be plotted on a growth chart together with weights.

The length of infants aged under two years is measured on a special measuring board and accurate results can be obtained only when there are two dedicated measurers present. One has to hold the infant's head against the top board while the other brings the footboard up to the child's feet while stretching him or her out. When the child is old enough to stand a special stadiometer can be used to measure height, but careful attention to detail is necessary for reproducible results.

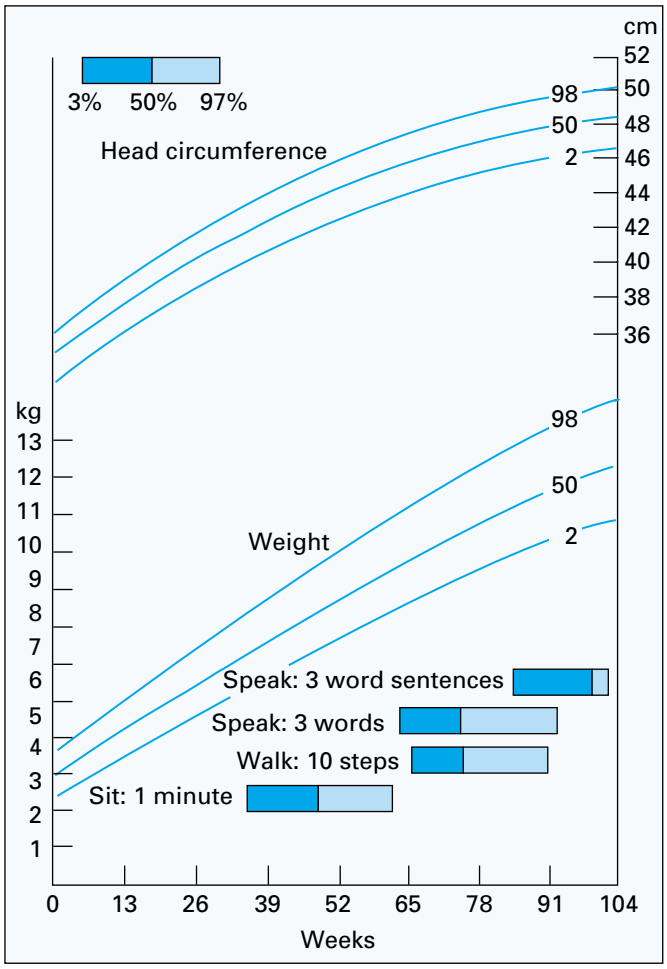


Figure 19.4 Range of ages for attainment of milestones.

Vision and squint

The infant's visual attention is engaged by the examiner crouching about 60 cm in front of the infant so that their eyes are on the same level. When the examiner slowly moves his head a short distance to one side then the other, the normal infant will continue to hold his gaze. Failure to do so is abnormal but may be due to something distracting the infant. Infants who fail to follow must be examined again two weeks later. Possible causes include lack of stimulation at home, delay in development, or blindness.

The second routine test at this age is the detection of the red reflex. If a bright ophthalmoscope is held about 45 cm away from an infant's eye, with the observer looking down the beam of light, a bright red reflex is normally seen in a fair skinned infant and a dark red or grey reflex in an infant with a dark skin. Absence of the reflex indicates an abnormality of the refractive media and requires urgent referral to an ophthalmologist.

An infant who does not fixate on the examiner's face at the second visit should be referred to a developmental specialist. At that stage additional tests will include an attempt to determine whether the infant can see at all by showing him a revolving Catford drum. This induces opticokinetic nystagmus – an involuntary movement of the eyes, which occurs even if the infant does not concentrate well.

In a young infant a squint can be noticed by observing the position of the light reflex on each cornea when a torch is shone into the eyes. Normally the reflex should be in the centre of each pupil or at a corresponding point on each cornea. If there is a possibility of a squint, movements of the eyes should be assessed and the cornea, lens, and fundi should

Box 19.1 Weights of napkin and vest at various ages

Baby	Napkin	Vest
0–3 months	35 g	30 g
3–9 months	55 g	45 g
9–12 months	70 g	60 g

Figure 19.5 Infant following movements of the examiner.

be examined. An infant with a squint needs to be referred to an ophthalmologist.

Motor function

The baby should be placed prone on a couch; normally he will lift his head for a few seconds, though some infants may take a few minutes to do so. Infants who usually sleep prone are more advanced in lifting their heads than others. If the infant does not eventually lift his head he needs to be re-examined two weeks later.

When the infant is held in the prone position, his head rises to the same plane as his trunk and his legs do *not* fall vertically.

Figure 19.6 Lifting the head.

Other assessments

The testes and hips should be examined by the methods described for the newborn infant. It is difficult to detect an abnormal hip at the age of six weeks and ideally the condition should have been detected in the newborn infant.

Cardiac murmurs should be sought for and the femoral pulses checked. Though hearing is not tested at six weeks, the normal infant will respond to sounds by stopping crying, opening his eyes widely or by a startle reflex, depending on the intensity of the sound.

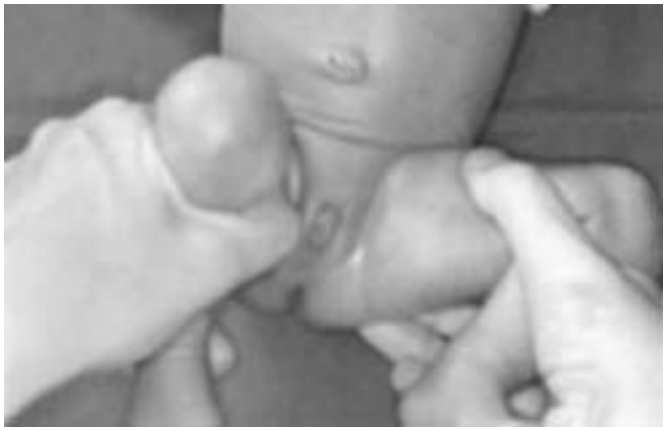


Figure 19.7 Examination of the hips.

Patterns of development in childhood

Progress in a particular field does not occur gradually but in spurts, and when one skill is advancing quickly others tend to go into abeyance.

There is considerable variation in the range of normal development, but among the children who fall outside the normal range there are some who later turn out to be normal. The children who fall into this group need examining more closely and more often to determine which are definitely abnormal. They should be referred to developmental specialists.

Lack of appropriate stimulation may cause delay in physical, intellectual, or social development. This factor must be taken into account if the child has been cared for in an institution or if the mother has a learning disability or has had puerperal depression. These children advance rapidly when the mother is given advice on handling them or they are placed in a better environment. Delayed speech development, probably the commonest disability of childhood, may be prevented by encouraging mothers to talk to their infants from birth.

No child has impaired intelligence if development is delayed in a single field and normal in all others. A child who is slow at learning will be delayed in all fields, except sometimes in walking. Intelligence cannot be assessed accurately until the age of about $4\frac{1}{2}$ years and before that age it is best to use the term developmental delay.

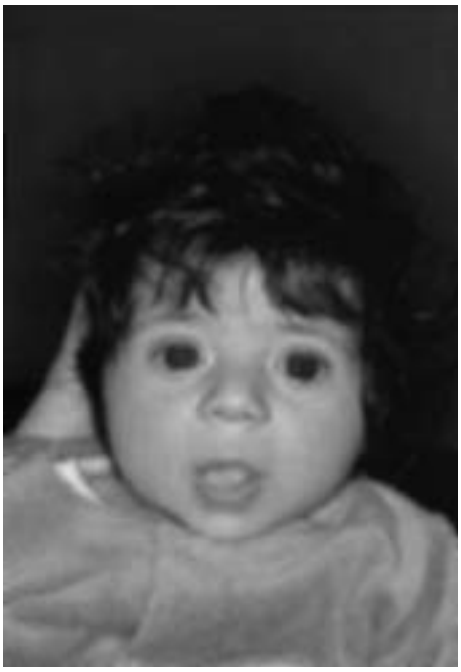
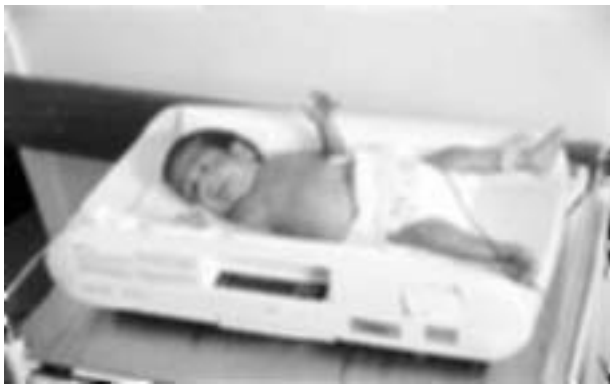
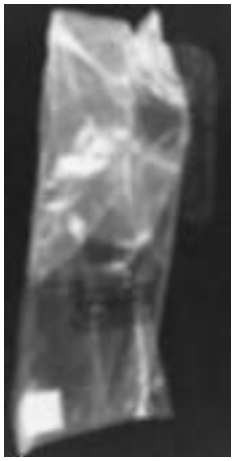


Figure 19.8 Looking at the camera.

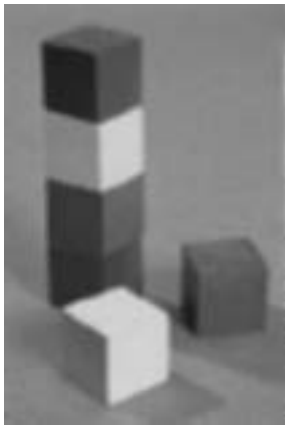


(1)

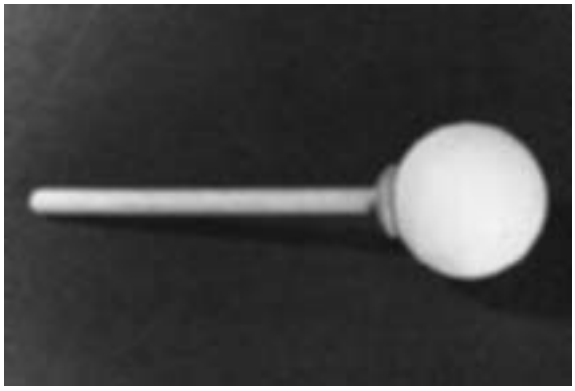
(2)



(3)



(4)



(5)

(6)

(7)

(8)

Fig. 19.9 Essential equipment: (1) scales, (2) disposable paper tape, (3) urine bag, (4) building blocks, (5) Nuffield rattle, (6) ophthalmoscope, (7) picture on ceiling, (8) auriscope.

Equipment for examining children

Most of the equipment needed for examining and assessing children is cheap, but its use would greatly improve the quality of care: scales, disposable paper tape measures, measuring rod, torch, auriscope, spatulas, centile charts, one inch cubes, Nuffield rattle, Ladybird first reading books, miniature cups and saucers, small table and chair, pictures on the ceiling, toys (including wooden puzzles, pencils and paper), disposable napkins, urine bags, heel prick equipment, including lancet and Steriswab, and ophthalmoscope.

20 Review at eight months

Mothers are asked whether they have any problems and whether the infants have had any illnesses since the last examination. They are also asked whether the infants are making sounds and the details of their diet, in particular whether they are chewing solids.

As at six weeks, the infant's alertness and interest are checked. Head circumference and weight should be measured and plotted on the growth chart.

Vision

Eye movements should be noted. If a squint is suspected, the position of the light reflex in each cornea should be observed; if there is a squint, the reflex will be in a different position in each eye. While the infant looks at an object one eye is covered with the doctor's hand. If the eye that is not being covered moves to fixate the object, that eye has an overt squint. The covering hand is slowly removed and if that eye moves to fixate the object, there is a latent squint. If a squint is seen or suspected at any time the infant should be seen by an ophthalmologist.

Near vision may be tested by placing a raisin or pellet of paper about 20 cm in front of an infant and watching whether he reaches out to grab it. Testing distant vision is time consuming at this age and is usually performed only at specialist clinics.

Figure 20.1 Near vision being tested.

Motor development

From the prone position infants should get up on their wrists. When pulled from the supine position they should be able to sit spontaneously for a minute or two. Normal children can sit without help by the age of eight months. Infants should also be able to take their weight on their legs when they are held standing.

If a cube is placed in front of the infant he should grab it with his whole hand. This test should be tried for both the right and the left hand, and the infant should transfer the object from one hand to the other. Infants who perform all these tests normally but are not sitting spontaneously should be seen again two months later.

A dislocated hip will show reduction of abduction and a proximal displacement of the lower buttock crease. Bilateral dislocation is hard to detect clinically but radiographs of the hips will confirm the diagnosis at this age.

Figure 20.2 Resting on the wrists in the prone position.

Hearing

The distraction method is the hearing test used at this age. The distractor holds the infant's attention while the tester produces a sound. Careful attention to detail is essential to produce reliable results. The mother is requested to position the infant at the edge of her lap and not to react to the sounds. The distractor sits at the level of the infant's face and holds the infant's attention with a toy that produces no noise. Eye contact is avoided. The toy is withdrawn from sight and immediately afterwards the tester produces the test sound at a distance of 45 cm from the infant's ear, on a level with the ear,

Figure 20.3 Taking weight on the feet.

8–9 months review

Review at 8–9 months

For parent to complete:

Health topics for discussion: Tick ☑

Accident prevention: choking, scalds, safety in cars and house, sunburn
Dental advice
Developmental needs
Nutrition, etc

This review is done by your health visitor and/or a doctor. Below is a list of things you may want to discuss when you see them. However, if you are worried about your child's health, growth or development you can contact your health visitor or doctor at any time.

Circle 'yes' or 'no' or 'not sure'

Are you feeling well yourself?	Yes/no/not sure
Have you any worries about your child's health?	Yes/no/not sure
Do you have any worries about how your baby is feeding?	Yes/no/not sure
Are you happy your baby is gaining weight?	Yes/no/not sure
Do you have any worries about your baby's development? (see pages 12–13)	Yes/no/not sure
Is your baby sitting alone?	Yes/no/not sure
Is your baby using both hands?	Yes/no/not sure
Does your baby babble (Ba-ba, da-da, etc.)?	Yes/no/not sure
Have you any worries about your baby's eyesight?	Yes/no/not sure
Can s/he recognise carer at a distance?	Yes/no/not sure
Have you noticed a squint (eyes not moving together?)	Yes/no/not sure
Do you think your baby can hear you?	Yes/no/not sure

Comment_____

_____

Are all your child's immunisations up to date? Yes/no

Do you have fire guards to stop your child touching heaters and open fires?

Review at 8–9 months

* Please place a sticker (if available) otherwise write in space provided.

Surname
First names
NHS number Local no
Address Sex M/F
Postcode D.O.B.__/__/__
G.P. Code
H.V. Code

Date of examination _____ /_____ /_____
Age in months _____
Name of examiner
Examiner code number Examination centre
WEIGHT_____ kg _____centile
HEAD CIRC._____ cm _____centile

Any previous ongoing medical problems? NO ☐ Yes ☐
If 'Yes' please specify (1) _____ (2)_____ (3) _____
ICD Code

ITEM	GUIDE TO CONTENT	CODE (Ring one)*	COMMENT
Physical	Parents observations. Diet.	S P O T R N	
Vision	Eye movements. Visual behaviour. Cover test.	S P O T R N	
Hearing	Distraction test.	S P O T R N	
Locomotion	Sitting. Some weight bearing.	S P O T R N	
Manipulation	Transfers. Uses both hands.	S P O T R N	
Speech/Lang.	Babble. Responds to speech.	S P O T R N	
Behaviour	Knows strangers. Enjoys mirror image. Socialisation. Alertness.	S P O T R N	
Hips	Check for CDH	S P O T R N	
Genitalia	Testicular descent (Ring 'N' for girl)	S P O T R N	

*(If one or more codes seem to apply select the last one, e.g. R takes priority over T, T over O etc.)

S = Satisfactory *(normal result)*
P = Problem *(significant condition on record)*
O = Observation *(special recall arranged)*
T = Treatment or investigation underway
R = Referred to any community or hospital service
N = Not examined for this item

Referred to (1)_____ (2)_____
(3)_____
Special recall in_____ wks/mths Signature _____

White copy: Stay in Record Yellow copy: to Child Health Office
Pink copy: to GP/HV

8–9 months

Figure 20.4

Figure 20.5 Testing hips.

but slightly behind it so that the infant does not see the test object or the tester's shadow. The distractor detects whether the infant turns the head or eyes abruptly towards the sound. The speed with which the infant responds to the sound is a useful indication of his alertness. The test sounds should include those with a known pitch and intensity such as a warbler. Voice sounds at minimal intensity should include "psss", "phth" (for high tones), or "oo" (for low tones). If an infant fails to turn to the sound, the ability to turn the head to each side should be assessed. If the infant responds to sounds on two out of three tests performed on one ear, hearing is considered normal in that ear. If the infant fails to respond, the ears should be examined with an auriscope.

The room where the test is performed must be quiet. It should ideally have a carpet and be removed from the waiting area and sounds of traffic. The commonest reasons for a child failing to respond to the test sounds are lack of interest, tiredness, distraction, wax blocking the external meatus, otitis media, and lack of familiarity with the test sounds. It is important to explain to the mother that these are far more likely reasons for lack of response than deafness and that the test will be repeated by the same person one month later. Infants who fail to respond at the second visit should be seen by an audiologist.

High tone deafness cannot be excluded by these tests and if there is a high suspicion of deafness – for example, if the infant had a very high plasma bilirubin concentration in the neonatal period, has congenital rubella syndrome, or has a family history of deafness – the infant should be seen by an audiologist even if the results of these simple tests appear normal. Pure tone audiology may be needed and the audiologist may have to see the child several times before a hearing defect can be completely excluded.

Pattern of development

Most normal children sit unaided by the age of eight months. African children tend to advance in most aspects of motor development faster than other children up to the age of about seven months. In some families one aspect of development – for example, sitting, walking, or speech – may be unusually early or late, the development in all other fields being normal.

Finding that an infant is not sitting by a certain age does not necessarily mean that he is abnormal, as he may have a variation of normal development. These children fall into a group who are late in reaching a developmental milestone and it is reasonable that they should be referred for further examination to exclude more sinister problems.

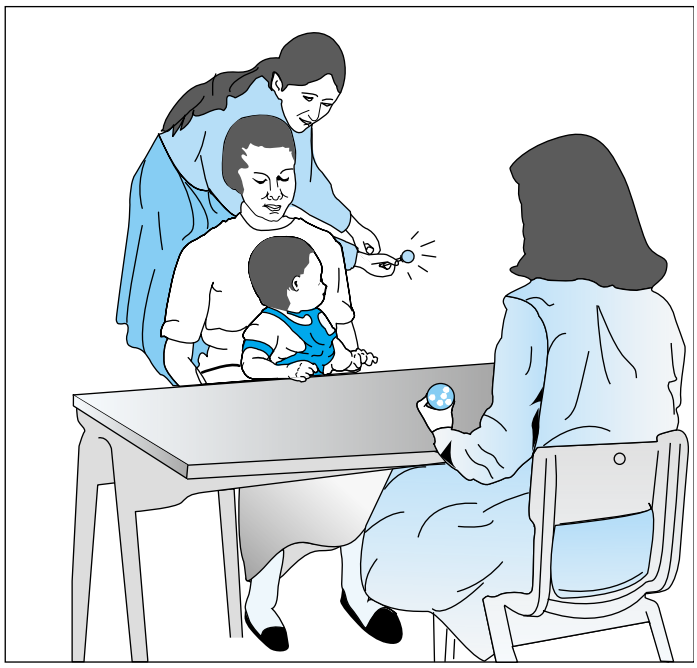


Figure 20.6 Distraction test.

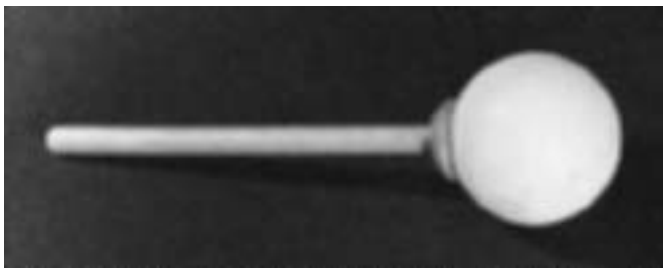


Figure 20.7 High-pitched rattle.

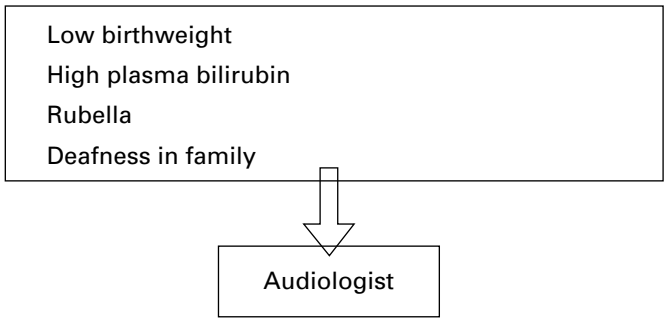


Figure 20.8 Indications for direct referral to the audiologist in the neonatal period.

21 Respiratory infections in the older infant

Infection of the respiratory tract is a common cause of illness of infants. Although pathogens are often not confined to anatomical boundaries, the infections may be classified as: (*a*) upper respiratory tract – common cold, tonsillitis, and otitis media; (*b*) middle respiratory tract – acute laryngitis and epiglottitis; (*c*) lower respiratory tract – bronchitis, bronchiolitis, and pneumonia.

Upper respiratory tract infection is usually the least serious condition but blockage of the nose by mucus may completely obstruct the airway in those infants who cannot breathe through their mouths. Middle respiratory tract infection may totally obstruct airflow at the narrowest part of the airway. Lower respiratory tract infection produces trivial signs initially but may be lethal within a few hours.

Viruses, which cause most respiratory tract infections, and bacterial infections produce similar clinical illness. Different viruses may produce an identical clinical picture or the same virus may cause different clinical syndromes. Clinically it may not be possible to determine whether the infection is due to viruses, bacteria, or both. If the infection is suspected of being bacterial, it is safest to prescribe an antibiotic, as the results of virus studies are often received after the acute symptoms have passed. The commonest bacterial pathogens are pneumococci, *Haemophilus influenzae*, group A β-haemolytic streptococci, and *Staphylococcus aureus*. Group B streptococci, Gram negative bacteria, and anaerobic bacteria are less common.

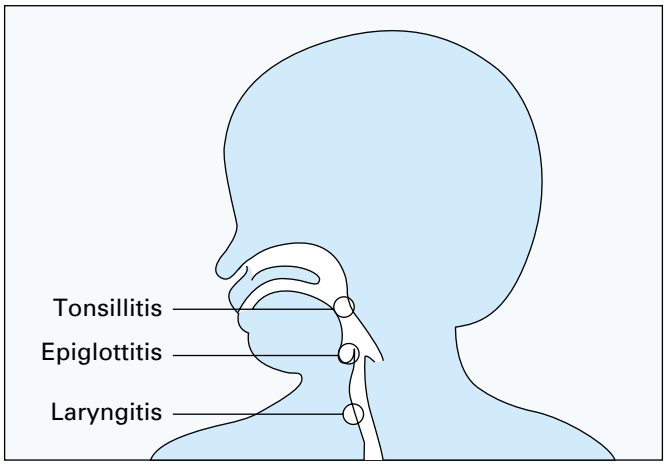


Figure 21.1 Sites of infection in the upper respiratory tract.

Common cold (coryza)

Preschool children have at least 3–6 colds each year. The main symptoms are sneezing, nasal discharge, cough, and, rarely, fever. Nasal obstruction in infants who cannot breathe through their mouths may cause feeding difficulties and, rarely, brief periods of apnoea. Similar symptoms may occur in the early phases of infection with rotavirus and be followed by vomiting and diarrhoea. Postnasal discharge may produce coughing. The commonest complication is acute otitis media, but secondary bacterial infection of the lower respiratory tract sometimes occurs.

There is no specific treatment for the common cold and antibiotics should be avoided. If an infant is not able to feed due to nasal obstruction from mucus, two drops of 0·9% sodium chloride solution can be instilled into each nostril before feeds three times daily. This will wash the mucus into the back of the pharynx and relieve the obstruction. There is a danger with nasal drops that they will run down into the lower respiratory tract and carry the infection there.

Box 21.1 Main symptoms of common cold

- Sneezing
- Nasal discharge
- Cough
- Fever (rare)

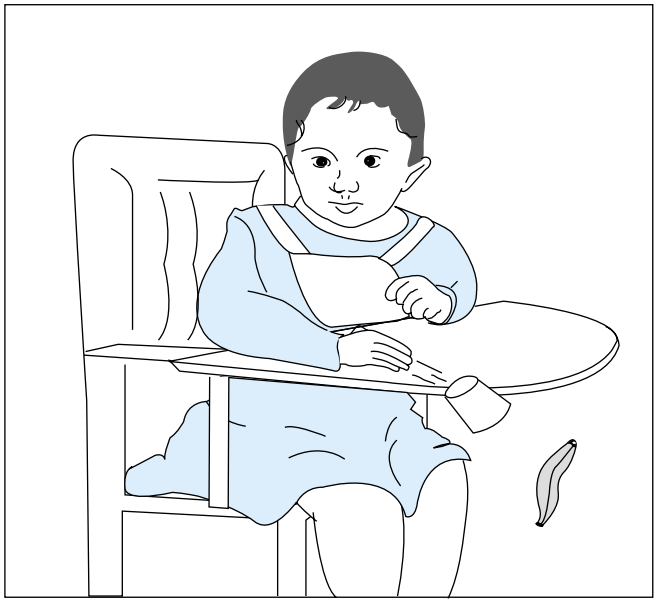


Figure 21.2 Picture of child in highchair pushing food off the table.

Tonsillitis and pharyngitis

In children aged under three years the commonest presenting features of tonsillitis are fever and refusal to eat, but a febrile convulsion may occur at the onset. Older children may complain of a sore throat or enlarged cervical lymph nodes, which may or may not be painful. Viral and bacterial causes cannot be distinguished clinically as a purulent follicular exudate may be present in both. Ideally, a throat swab should be sent to the laboratory before starting treatment, to determine a bacterial cause for the symptoms and to help to

indicate the pathogens currently in the community. There has been a recurrence of group A haemolytic streptococci in outbreaks of sore throat and a more liberal use of penicillin is justified. As this organism is the only important bacterium causing tonsillitis, penicillin is the drug of choice and the only justification for using another antibiotic is a convincing history of hypersensitivity to penicillin. In that case the alternative is erythromycin. In the absence of an outbreak of group A streptococcus infection the indication for oral penicillin is fever or severe systemic symptoms. The drug should be continued for at least 10 days if a streptococcal infection is confirmed. Parents often stop the drug after a few days as the symptoms have often abated and the medicine is unpalatable. The organism is not eradicated unless a full 10-day course is given.

Viral infections often produce two peaks on the temperature chart.

An extensive thick white shaggy exudate on the tonsils (sometimes invading the pharynx) suggests infectious mononucleosis and a full blood count, examination of the blood film, and a Monospot test are indicated. Although rare in Britain, a membranous exudate on the tonsils suggests diphtheria and an urgent expert opinion should be sought.

Fluids can be given while there is dysphagia and regular paracetamol during the first 24–48 hours reduces fever and discomfort.

A peritonsillar abscess (quinsy) is now extremely rare. It displaces the tonsil medially so that the swollen soft palate obscures the tonsil and the uvula is displaced across the midline. The advice of an otolaryngology surgeon is needed urgently.

Otitis media

Pain is the main symptom of acute otitis media and is one of the few causes of a fretful infant who cries all night. The pain is relieved if the media drum ruptures. Viruses probably cause over half the cases of acute otitis, but a viral or bacterial origin cannot be distinguished clinically. The commonest bacteria are pneumococci, group A β-haemolytic streptococci, and *H. influenzae*.

Children are often fascinated by the light of the auriscope and the auriscope speculum can be placed on a doll or the child's forearm for reassurance. Gentleness is essential and the speculum should never be pushed too far into the external meatus because this causes pain. If the pinna is pulled gently outwards and downwards to open the meatal canal, the tympanic membrane is visible with the tip of the speculum in the outer end of the meatus. In early cases of otitis media there are dilated vessels over the upper and posterior part of the drum. Later, the tympanic membrane becomes red, congested, and bulging and the light reflex is lost. Swelling or tenderness behind the pinna should always be sought, as mastoiditis can be easily missed.

The choice of initial treatment lies between amoxycillin and erythromycin. If there is no improvement in the drums after two or three days, another antibiotic should be substituted. Cefaclor or amoxycillin with clavulanic acid may be included in the second line drugs. The duration of the course of antibiotics is controversial. The most common view is that antibiotics should be given for seven days and the ears examined again before the course is stopped. There is some evidence that short courses of 3–5 days of antibiotics given at high doses may be as effective as the longer courses. Ideally, a hearing test should be performed three months after each attack of acute otitis media to detect residual deafness and

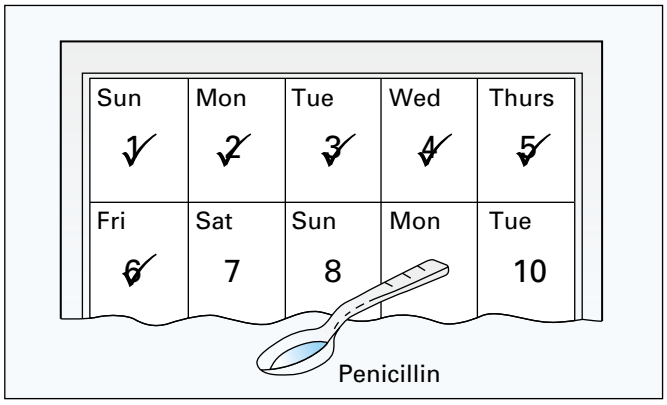


Figure 21.3 Antibiotics should be taken regularly.

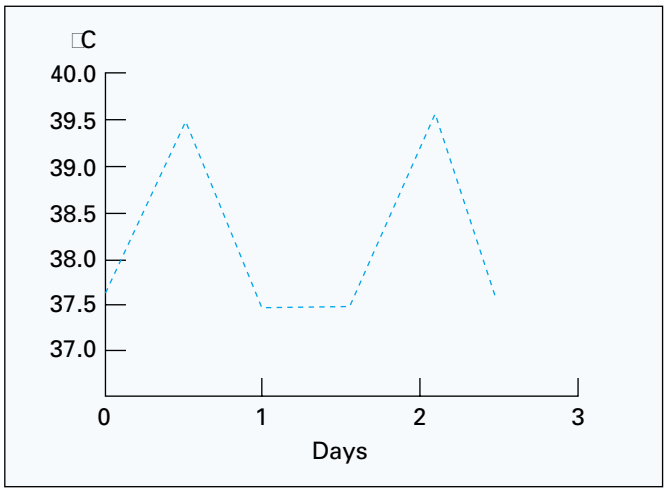


Figure 21.4 Temperature chart in a viral infection.

Figure 21.5 Auriscope.

"glue ear". One study showed that after the first attack of acute otitis media in infants, which was treated with antimicrobial agents, 40% of patients had no middle ear effusion after one month and 90% after three months.

Stridor

Stridor is noisy breathing caused by obstruction in the pharynx, larynx, or trachea. It may be distinguished from partial obstruction of the bronchi by the absence of rhonchi. Although most cases are due to acute laryngitis and may resolve with the minimum of care, similar features may be due to a foreign body and may cause sudden death.

Stridor is recognised as one of the most ominous signs in childhood. Any doctor should be able to recognise the sound over the telephone and arrange to see the child immediately. Examination of the throat may precipitate total obstruction of the airway and should be attempted only in the presence of an anaesthetist and facilities for intubation.

A glance at the child will show whether urgent treatment is needed or whether there is time for a detailed history to be taken. The doctor needs to know when the symptoms started and whether there is nasal discharge or cough. Choking over food, especially peanuts, or the abrupt onset of symptoms after playing alone with small objects suggests that a foreign body is present.

During the taking of the history and the examination mothers should remain near their children and be encouraged to hold them and to talk to them. All unpleasant procedures, such as venepuncture, should be avoided. This reduces the possibility of struggling, which may precipitate complete airway obstruction. Agitation and struggling raise the peak flow rate and move secretions, which results in increased hypoxia and the production of more secretions.

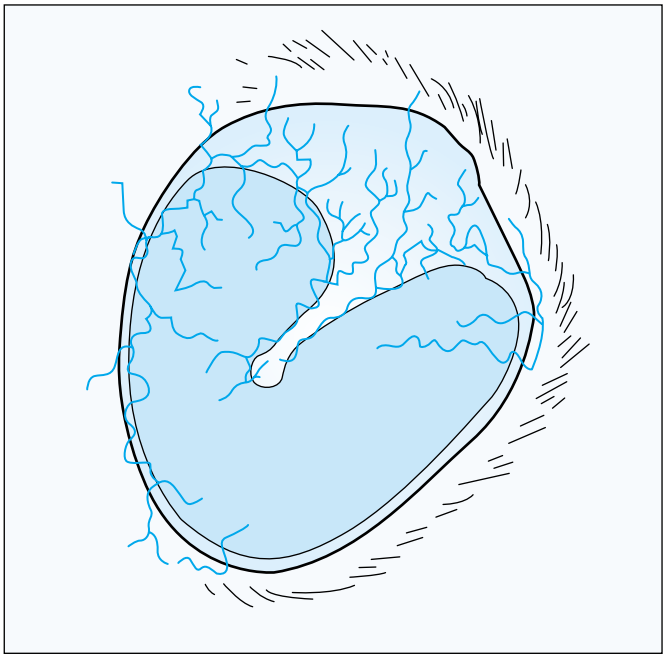


Figure 21.6 Acute otitis media.

Box 21.2 Procedures to be avoided in the presence of stridor

Throat examination
Venepuncture → Cardiac arrest
Struggling

Acute laryngotracheitis

Acute laryngitis causes partial obstruction of the larynx. It is characterised by inspiratory and expiratory stridor (noisy breathing), cough, and hoarseness. The laryngeal obstruction is due to oedema, spasm, and secretions. Affected children are usually aged six months to three years, and the symptoms are most severe in the early hours of the morning. Recession of the intercostal spaces indicates significant obstruction and cyanosis or drowsiness shows that total obstruction of the airway is imminent.

A child often improves considerably after inhaling steam, which is provided easily by turning on the hot taps in the bathroom. Mild cases may be treated successfully at home using this method, but children must be visited every few hours to determine whether they are deteriorating and need to be admitted to hospital. Continuous stridor or recession demands urgent hospital admission. Hypoxaemia or thirst may cause restlessness and should be corrected and sedatives avoided. Nebulised adrenaline or betamethasone may produce considerable improvement in symptoms within an hour. Rarely, the obstruction needs to be relieved by passing an endotracheal tube or performing a tracheostomy.

Acute laryngotracheitis is usually caused by a viral infection and therefore infants with mild symptoms do not need antibiotics. In a few cases *Staph. aureus* or *H. influenzae* is present and the associated septicaemia makes the child appear very ill. Bacterial infection is characterised by plaques of debris and pus on the surface of the trachea, partially obstructing it, just below the vocal cords. Acute epiglottitis and acute

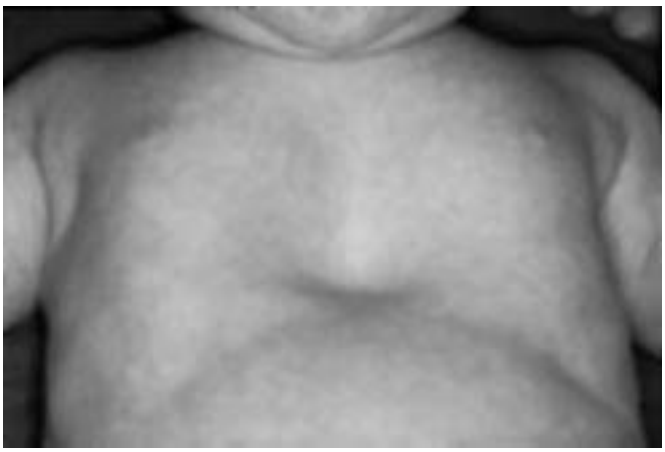


Figure 21.7 Substernal recession.

Box 21.3 Restlessness could indicate:

- hypoxia
- thirst
- hunger
- fear

laryngitis may be indistinguishable clinically since stridor and progressive upper airway obstruction are the main features of both. If bacterial infection is suspected cefotaxime is given intravenously. Some paediatricians give steroids as well. The dose of hydrocortisone is 100 mg intramuscularly or intravenously repeated once after two hours. No effect is seen for at least two hours. Later, betamethasone should be given at a dose of 3 mg intravenously every six hours but only until signs of improvement appear. Children with severe symptoms should be managed in the intensive care unit.

Children with epiglottitis are usually aged over 2 years; drooling and dysphagia are common, and the child usually wants to sit upright. When the obstruction is very severe, the stridor becomes ominously quieter. There is usually an associated septicaemia with *H. influenzae*. Epiglottitis is now rare due to the Hib vaccine.

Other causes of stridor

Even if the symptoms have settled and there are no abnormal signs, a history of the onset of sudden choking or coughing can never be ignored. A radiograph of the neck and chest should be taken and may show a hypertranslucent lung obstructed by a foreign body, a shift of the mediastinum, or, less commonly, collapse of part of the lung or a radio-opaque foreign body. The radiograph may be considered normal. Bronchoscopy may be needed to exclude a foreign body even if the chest radiograph appears to be normal. Stridor in a child who has had scalds or burns or has inhaled steam from a kettle suggests that intubation or tracheostomy may be needed urgently.

If the cause of stridor is likely to be a foreign body below the larynx, the object should be removed immediately by a thoracic surgeon in the main or accident and emergency operating theatre. If the object is above the larynx and if an ENT surgeon or anaesthetist is not immediately available and the child is deteriorating, the safest treatment is to insert a wide needle, such as Medicut size 14, into the trachea in the midline just below the thyroid cartilage. It may be preferable to insert two needles. No attempt should be made to look at the mouth or throat or remove the object, as the struggling that may follow may impact the object and prove fatal. A first aid measure usually performed before arrival is to slap the infant's back between the shoulder blades while holding the infant upside down by the legs. The child should remain in the position he or she finds most comfortable, which is usually upright. Forceful attempts to make the child lie flat – for example, for a radiograph – may result in complete airway obstruction.

Infants with congenital largyngeal stridor, which is due to loose aryepiglottic folds, usually have inspiratory stridor only. The symptoms begin during the first few weeks of life and usually persist for several months, becoming more severe during episodes of upper respiratory tract infections or while crying, but subsiding in sleep.

Acute bronchitis

Acute bronchitis often follows a viral upper respiratory tract infection; there is always a cough, which may be accompanied by wheezing. There is no fever or difficulty with feeding. The respiratory rate is normal and the symptoms resolve within a week. The only signs, which are not always present continuously, are rhonchi. Since bronchitis is usually due to a virus, antibiotics are indicated only if secondary bacterial infection is suspected.

Figure 21.8 Oximeter for measuring blood oxygen saturation.

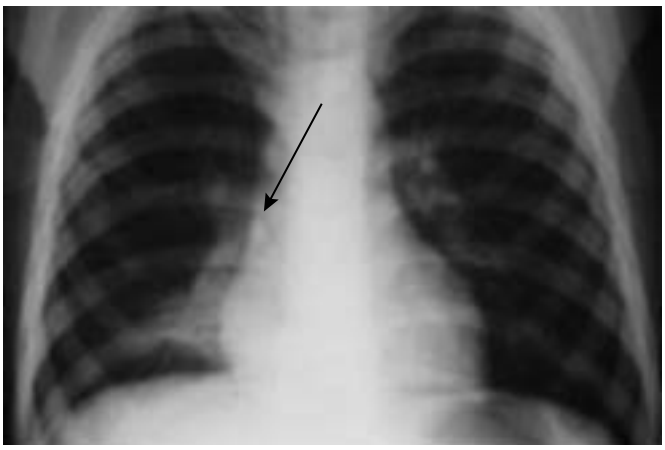


Figure 21.9 Foreign body in right main bronchus (tooth).

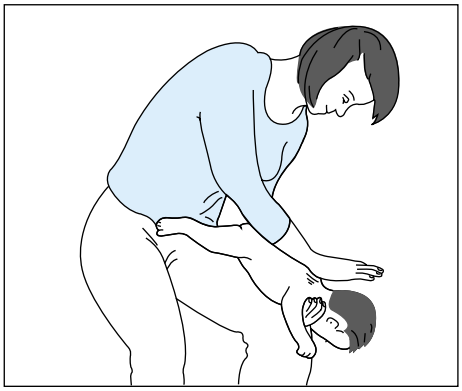


Figure 21.10 Slapping the back to dislodge a foreign body.

Box 21.4 Upper limit for normal respiratory rate related to age

< 2 months	60/min
2–11 months	50/min
12–60 months	40/min

Acute bronchiolitis

Acute bronchiolitis is an acute infection that occurs in winter epidemics in infants aged under one year. For the first few days there may be only a rasping cough but deterioration may occur within a few hours, causing a raised respiratory rate, indrawing of the intercostal spaces, cyanosis, drowsiness, and apparent enlargement of the liver. The respiratory syncytial virus is found in over 70% of cases. During epidemics the disease may be recognised at an early stage, but at the beginning of epidemics the condition may be recognised only when the infant is moribund. Signs in the chest vary at different stages of the illness and there may be no adventitious sounds. The chest radiograph may appear normal even in severely ill infants. In units with ELISA or immunofluorescence techniques for diagnosing respiratory syncytial virus, results are available the same day, treatment can be planned, and cross infection avoided. Oxygen is the most important aspect of treatment and is efficiently given by a nasal cannula with a special low flow regulator. Most infants need intragastric tube feeding or intravenous fluids for a few days. Where a viral cause can be confirmed immediately, antibiotics can be avoided but they should not be withheld from severely ill infants, as there is a possibility of additional bacterial infection. A few infants develop progressive respiratory distress and intermittent positive pressure ventilation may have to be considered.

The infant is discharged from hospital when feeding is normal. The cough may persist for six weeks, but if there is no improvement after three weeks, a sweat test should be performed to exclude cystic fibrosis.

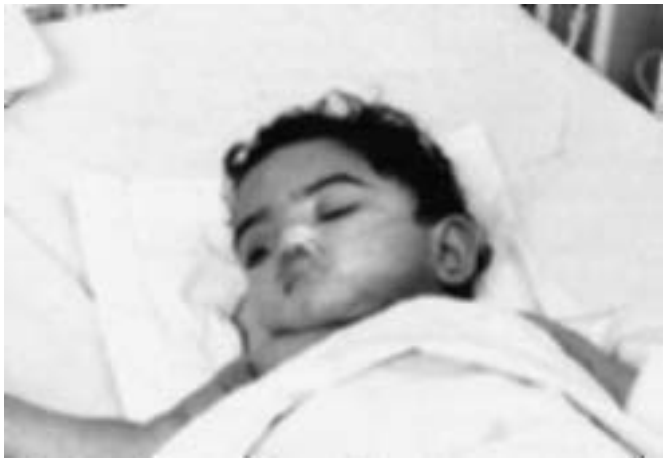


Figure 21.11 Oxygen given by a nasal cannula.

Bronchopneumonia and segmental pneumonia

Pneumonia is acute inflammation of the lung alveoli. In bronchopneumonia the infection is spread throughout the bronchial tree whereas in segmental pneumonia it is confined to the alveoli in one segment or lobe. A raised respiratory rate at rest or indrawing of the intercostal spaces distinguishes pneumonia from bronchitis. The upper limit for a normal respiratory rate is related to age (see Box 21.4). Cough, fever, and flaring of the alae nasi are usually present and there may be reduced breath sounds over the affected area as well as crepitations. A chest radiograph, which is needed for every child with suspected pneumonia, may show extensive changes when there are no localising signs in the chest. The radiograph may show an opacity confined to a single segment or lobe, but there may be bilateral patchy changes. Bacterial cultures of throat swabs and blood should be performed before treatment is started. Ideally, nasopharyngeal secretions should be studied virologically and virus antibody titres of serum collected in the acute and convalescent phases should be measured.

Children with pneumonia are best treated in hospital, as they may need oxygen treatment. Antibiotics should be prescribed for all cases of pneumonia, although a viral cause may be discovered later. If the child is not vomiting and not severely ill, oral erythromycin or amoxycillin is given. Cefotaxime is given intravenously if the symptoms are severe, and erythromycin is added when failure to improve promptly suggests infection with mycoplasma or chlamydia. Antibiotic treatment can be modified when the results of bacterial cultures are available. Intravenous fluids may be needed.

The chest radiograph of a child with segmental or lobar pneumonia should be repeated after one month.

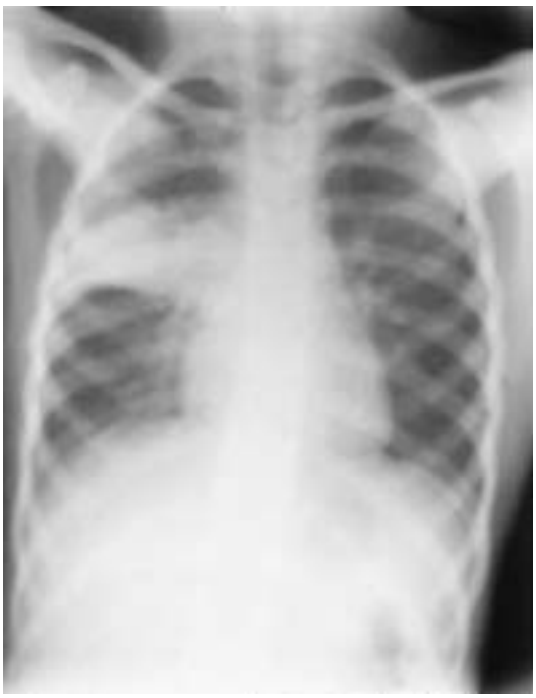


Figure 21.12 Chest radiograph showing pneumonia in the right lung.

Recurrent respiratory infections

Although all doctors concerned with children are familiar with the catarrhal child, the exact pathology of the condition is unknown and it is called by many names – postnasal discharge, perennial rhinitis, or recurrent bronchitis. These children have an increased incidence of colds, tonsillitis, and acute otitis media. Recurrent episodes of symptoms such as fever, nasal discharge, and cough are most common during the second half of the first year of life, the first two years at nursery school, and the first two years at primary school. Recurrent viral or bacterial infections contracted from siblings or fellow pupils may be important, but the considerable differences between the behaviour of children in the same family suggest the possibility of a temporary immunological defect.

Various treatments including nasal drops and oral preparations of antihistamines are given with little effect. A chest radiograph should be performed to exclude persistent segmental or lobar collapse. A sweat test should be carried out to exclude cystic fibrosis and plasma immunoglobulin studies should be conducted to exclude rare syndromes.

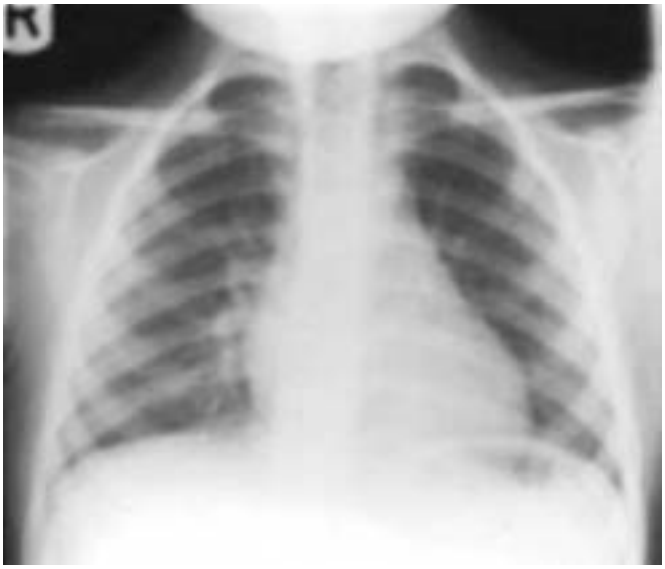


Figure 21.13 Normal chest radiograph.

Recurrent bronchitis

Two separate episodes of acute bronchitis may occur in a normal child in a year. If attacks are more frequent at any age, bronchial asthma should be considered. Viruses cause the majority of attacks of bronchitis and will precipitate most attacks of bronchial asthma. Some paediatricians have reverted recently to the older terms recurrent or wheezy bronchitis, as most children with these features become free of symptoms by the age of five. Although the pathological processes and prognosis may differ between recurrent bronchitis and bronchial asthma, there is no clinical or laboratory method of distinguishing between them and treatment is the same.

After an episode of severe symptoms during an infection with respiratory syncytial virus (bronchiolitis), many children have recurrent episodes of cough and wheezing during the subsequent four years. It is not known whether the respiratory syncytial virus predisposes the child to recurrent respiratory symptoms or whether the child has a predisposition to produce severe symptoms with viral respiratory infections.

If there is a persistent or recurrent cough, a chest radiograph should be performed to exclude persistent segmental or lobar collapse. A Mantoux test for tuberculosis and a sweat test to exclude cystic fibrosis should be performed and plasma concentrations of immunoglobulins and IgG subclasses should be measured to exclude transient or permanent immune deficiencies.

The management of recurrent bronchitis or bronchial asthma is the same. For infants with mild symptoms an oral bronchodilator, for example salbutamol, can be given at the beginning of an episode and continued for a week. If this is not effective, a bronchodilator can be given by a *small* spacer device with a face mask or by air pump and nebuliser. Infants with severe or frequent episodes can be given an inhaled steroid as a prophylactic drug for six weeks and the course can be extended to six months if there is an improvement in symptoms.

Prophylactic drugs can be given with a small spacer device or by an air pump and nebuliser. If infants are receiving both a bronchodilator and a prophylactic drug, the dose of bronchodilator should be given just before the prophylactic drug.

Box 21.5 Causes of recurrent cough and wheezing

Recurrent bronchitis
?
Bronchial asthma

Figure 21.14 Small spacer device with a face mask.

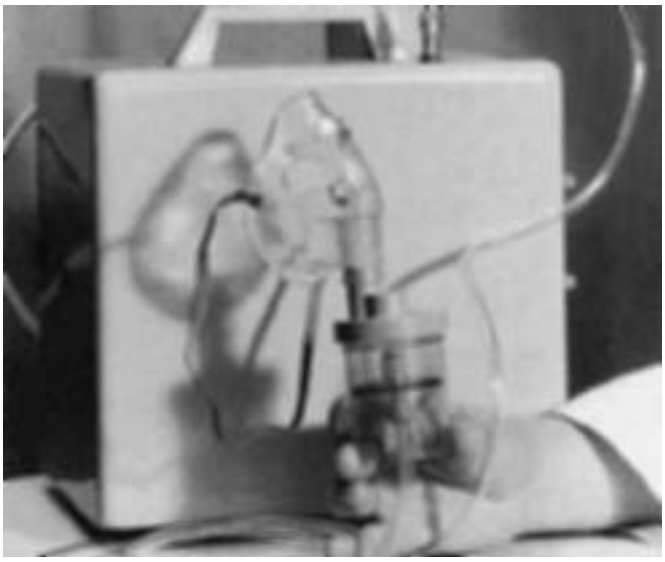


Figure 21.15 Air pump and nebuliser.

22 Whooping cough

Young infants receive no protective immunity to whooping cough from their mothers and have the highest incidence of complications. Immunisation is directed at increasing herd immunity and reducing the exposure of infants to older children who have the disease. Recent changes in advice on contraindications to whooping cough vaccine have been associated with the increased use of this vaccine.

Contraindications to immunisation

Convulsions and encephalopathy have been reported as rare complications, but these conditions may arise from other causes and be falsely attributed to the vaccine. Neurological complications after whooping cough itself are considerably more common than after the vaccine.

As with any other elective immunisation procedure, if a child is suffering from any acute illness then vaccination should be postponed until the child has fully recovered. Minor infections without fever or systemic upset, however, are not reasons to delay immunisation. Vaccination should not be carried out in children who have a history of a severe local or general reaction to a preceding dose; the following reactions should be regarded as severe.

Local – an extensive area of redness and swelling, which becomes indurated and covers most of the anterolateral surface of the thigh or a major part of the circumference of the upper arm.

General – fever (a temperature of 39·5°C or higher) within 48 hours of vaccination, anaphylaxis, bronchospasm, laryngeal oedema, generalised collapse, prolonged unresponsiveness, prolonged inconsolable screaming, and convulsions occurring within 72 hours.

A personal or family history of allergy is *not* a contraindication to immunisation against whooping cough; nor are stable neurological conditions such as cerebral palsy or spina bifida.

Children with problem histories

When there is a personal or family history of *febrile* convulsions then there is an increased risk of these occurring after pertussis immunisation. In such children immunisation is *recommended*, but advice on the *prevention of fever* (see Box 22.2) should be given at the time of immunisation.

In a recent British study, children with a family history of epilepsy were immunised with pertussis vaccine without any appreciable adverse events. These children's developmental progress has been normal. In children who have a close (first degree) relative with *idiopathic epilepsy* there may be a risk of developing a similar condition, irrespective of whether they are vaccinated. Immunisation is *recommended* for these children.

When there is a *still evolving neurological problem* immunisation should be *deferred* until the condition is stable. When there has been a documented history of *cerebral damage in the neonatal period* immunisation *should be carried out unless there is evidence of an evolving neurological abnormality.* If immunisation is to be deferred, this should be stated on the neonatal discharge summary. When there is doubt, appropriate advice should be sought from a consultant paediatrician, district immunisation coordinator, or consultant in public health medicine *rather than withholding the vaccine.*

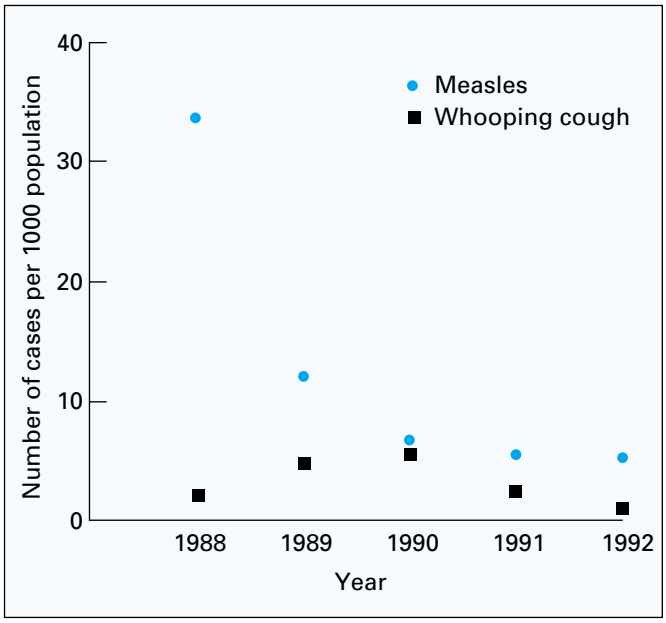


Figure 22.1 Prevalence of measles and whooping cough. (Source: OPCS)

Box 22.1 Mothers can fill out a card in the waiting room at each visit for immunisation

Is the baby unwell in any way? YES/NO
Has the baby had any side effects from previous immunisation? YES/NO
Did the baby behave normally during the first week of life? YES/NO
Has the baby, or anyone in the immediate family, ever had fits or convulsions? YES/NO
Is the baby developing normally? YES/NO

Box 22.2 Post-immunisation fever

The doctor should advise the parent that if fever develops after immunisation the child may be given a dose of paracetamol followed, if necessary, by a second dose 4–6 hours later. The dose of paracetamol for post-immunisation fever for an infant aged 2–3 months is 60 mg; an oral syringe can be obtained from any pharmacy to give the small volume required. The doctor should warn the parent that medical advice should be sought if the fever persists after the second dose.

Diagnosis

Whooping cough is difficult to diagnose during the first 7–14 days of the illness (catarrhal phase), when there is a short dry cough at night. Later, bouts of 10–20 short dry coughs occur day and night; each is on the same high note or rises in pitch. A long attack of coughing is followed by a sharp indrawing of breath, which may produce the crowing sound or whoop. Some children, especially babies, with *Bordetella pertussis* infection never develop the whoop. Feeding with crumbly food often provokes a coughing spasm, which may culminate in vomiting. Afterwards there is a short period when the child can be fed again without provoking coughing. In uncomplicated cases there are no abnormal respiratory signs.

The most important differential diagnosis in infants is bronchiolitis; this is usually due to the respiratory syncytial virus, which produces epidemics of winter cough in infants under one year. For the first few days there may be only bouts of vibratory rasping cough, which never produce a whoop. Later, rhonchi or crepitations are heard in the chest and the infant either deteriorates or improves rapidly within a few days. Older siblings infected with the virus may have a milder illness. Other viruses may cause acute bronchitis with coughing but there are seldom more than two coughs at a time.

A properly taken pernasal swab plated promptly on a specific medium should reveal *B. pertussis* in most patients during the first few weeks of the illness. A blood lymphocyte count of 10×10^9/l or more with normal erythrocyte sedimentation rate suggests whooping cough.

Management

If the diagnosis is suspected in the catarrhal phase (usually because a sibling has had recognisable whooping cough) a 10-day course of erythromycin may be given to the child and to other children in the home. Parents must be warned that an antibiotic may shorten the course of the disease only in the early stages and is unlikely to affect established illness. Vomiting can be treated by giving soft, not crumbly, food or small amounts of fluid hourly.

No medicine reliably reduces the cough. Oral salbutamol has been used in a dosage of 0.3 mg/kg/24 hours divided into three doses, When sleep is disturbed some authorities recommend that the child should be given a bedtime dose of 3–5 mg/kg of phenobarbitone. In severe cases mothers can be taught to give physiotherapy, which may help to clear secretions, especially before the infant goes to sleep. An attack may be stopped by a gentle slap on the back.

The threshold for admission to hospital should be lower for children aged under six months. Convulsions and cyanosis during coughing attacks are absolute indications for admission to an isolation cubicle. Parents often become exhausted by sleep loss and arranging for different members of the family to sleep with the child will give the mother a respite. The cough usually lasts for 8–12 weeks and may recur when the child has any new viral respiratory infection during the subsequent year. If the child is generally ill or the cough has not improved after six weeks, a chest radiograph should be performed to exclude bronchopneumonia or lobar collapse, which need treatment with physiotherapy and antibiotics. Long term effects on the lung, such as bronchiectasis, are rare in developed countries.

The infant will not be infective for other children after about four weeks from the beginning of the illness or about two days after erythromycin is started. The incubation period is about seven days and contacts who have no symptoms two weeks after exposure have usually escaped infection.

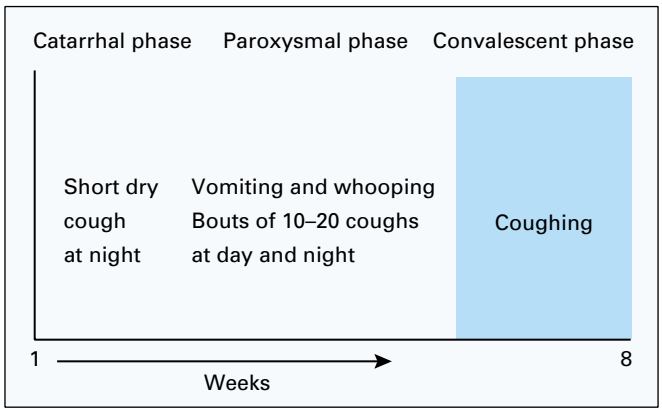


Figure 22.2 Phases of whooping cough.

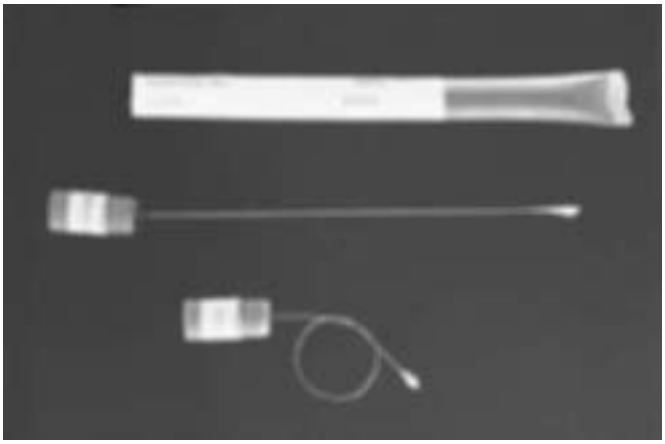


Figure 22.3 Pernasal swab for culture of *B. pertussis*.

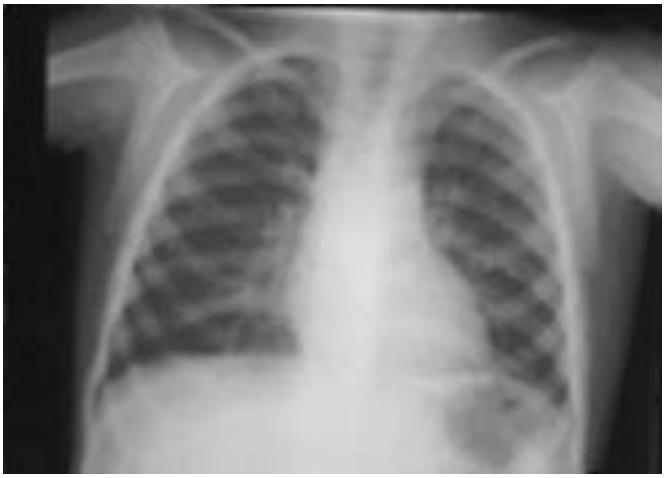


Figure 22.4 Normal radiograph.

23 Fever in the older infant

Examining an ill child

Infants should be allowed to adopt whatever position they like (usually on their mother's knee) where they can be observed. The history taken from the mother must contain enough detail to provide a differential diagnosis before the physical examination begins.

A systematic examination may prove impossible in a protesting infant, so the most relevant system should be examined first. If the infant objects to being fully undressed, only the part being examined need be exposed. The examiner needs warm hands and should talk to the infant continuously to soothe him. Patience and gentleness at this stage are often rewarded by the infant accepting a full examination without protest.

The baby's abdomen is best examined as the infant lies on his mother's knees during a feed. Abdominal breathing is normal and the edge of the liver is easily palpable. Even the newborn will show abdominal tenderness by grimacing or crying. Rectal examination may be performed with the little finger.

Instruments such as auriscopes and stethoscopes should be shown to the infant first and rested on his forearm so that he can see what they are. The auriscope must be used gently and the speculum not pushed too far into the external meatus. When the pinna is pulled gently outwards and downwards, the tympanic membrane can be seen with the tip of the speculum in the outer end of the meatus.

Examination of the throat, which is the most disliked procedure, should be left until last. It should never be performed in an infant with stridor unless the examiner has the facilities to intubate the infant immediately.

The infant should be examined again no longer than 24 hours later and the initiative for this review should not be left to the parents, who may not appreciate the danger of features that have developed.

The normal oral or rectal temperature is about 37.5°C (99.5°F) and the normal axillary (skin) temperature 37.0°C (98.4°F). If the temperature is 0.5°C above these levels, the infant has a fever. In the prodromal period of any infectious disease of childhood, fever may be the only symptom.

Figure 23.1 Throat examination.

Figure 23.2 Ear examination.

Figure 23.3 Auriscope.

Specific causes

Tonsillitis – There may be small areas of pus on the tonsils or the throat may be generally red. The presence of pus does not help to distinguish between bacterial infection due to group A streptococcus (*Streptococcus pyogenes*) and a viral infection. Petechiae on the soft palate usually indicate a viral infection. A throat swab should be taken and a 10-day course of oral penicillin given (see p. 82).

Acute otitis media – An early sign is an increase in the size of the vessels of the upper posterior part of the drum. Later the drum becomes dull pink or red and in severe cases there is bulging. Ear drums can be examined efficiently only if the auriscope has a magnifying lens. It is important to look for swelling or tenderness over the mastoids. The first choice of drugs is oral amoxycillin or erythromycin. If these are not effective, amoxycillin with clavulanic acid or cefalexin is given. The drum should be examined again after two days and if there is no improvement the antibiotic should be changed.

carried out during the acute phase, cystourethrography is performed when the urine is sterile, and isotope (^{90m}Tc dimercaptosuccinic acid (DMSA)) scanning is carried out at least six weeks after the acute phase. DMSA scanning detects scars, ultrasound examination shows obstructive lesions, and cystourethrography detects vesicoureteric reflux and should be performed under the protection of a suitable antibiotic.

Regular urine cultures are advisable at three monthly intervals as well as at times of fever or recurrence of symptoms. The urine should be cultured regularly while vesicoureteric reflux persists and in children with renal damage at least until the age of five years. The child's growth and blood pressure should be measured regularly.

Pneumonia – A raised respiratory rate at rest and fever may be the only signs of pneumonia and a chest radiograph may be necessary to show consolidation. The upper limit for the normal respiratory rate is 60 per minute for infants less than two months of age and 50 per minute between two and 12 months. Movements of the alae nasi and indrawing of the chest wall between the ribs are confirmatory signs.

Osteomyelitis or septic arthritis may present with fever. In the early stages the only helpful sign may be the infant's reluctance to move a limb and radiographs are often normal. Later there is redness, swelling, and tenderness at the site of the osteomyelitis, which usually affects the maxilla, long bones, or vertebrae in infants. Early blood culture and radioisotope bone scans are helpful in the diagnosis.

Malaria – If within the previous two years the child has been in an area where malaria is endemic, blood films should be examined immediately for malarial parasites. The parasites are most numerous during fever but may be found at any time.

Pyrexia of undetermined origin

Infants have no localising symptoms or signs. If they simply have fever, they should be seen again within a few hours or admitted to a cubicle on a children's ward for observation. Usually these children will have a full blood count, urine microscopy, and culture and sometimes blood culture and CSF examination.

Persistent fever lasting more than a few days should prompt the consideration of Kawasaki disease. Features include conjunctivitis, cracked lips, red throat, cervical lymphadenopathy, blotchy rash, oedema of the hands and feet, and reddening of the palms and soles. In the second week there may be peeling of the skin of the finger tips and toes. Prompt admission for intravenous gammaglobulin reduces the risk of coronary artery aneurysms.

Figure 23.7 Regular repeated examinations of the urine.

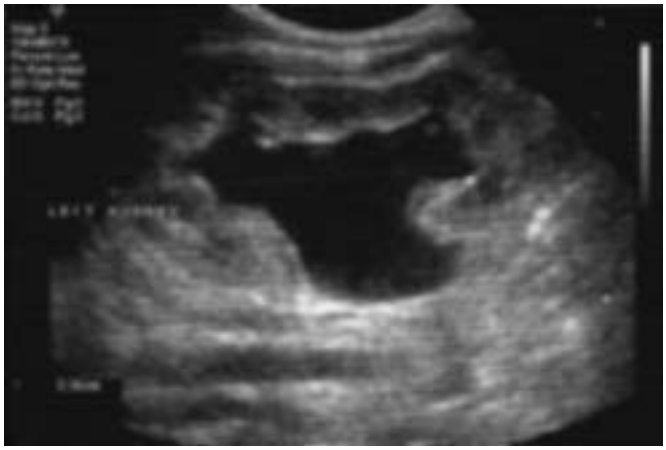


Figure 23.8 Ultrasound scan of kidney: dilated renal pelvis.

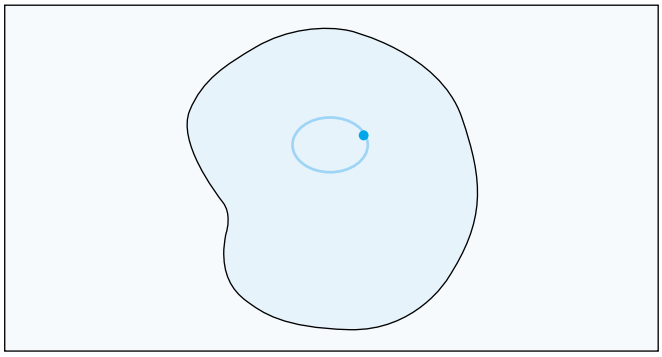


Figure 23.9 Malarial parasite in red cell.

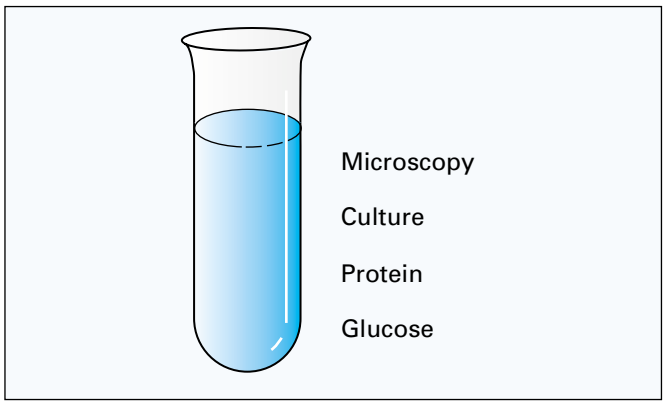


Figure 23.10 Urine examination.

24 Convulsions in the older infant

In infants between the ages of one month and one year convulsions are usually associated with fever. If there is no fever, epilepsy should be considered.

Fits can be divided into generalised or partial seizures. Generalised seizures include tonic-clonic and myoclonic fits. Partial seizures include focal and temporal lobe fits. During some episodes partial seizures may be followed by generalised seizures.

Generalised tonic-clonic fits are the most common type. The child may appear irritable or show other unusual behaviour for a few minutes before an attack. Sudden loss of consciousness occurs during the tonic phase, which lasts 20–30 seconds and is accompanied by temporary cessation of respiratory movements and central cyanosis. The clonic phase follows and there are jerky movements of the limbs and face. The movements gradually stop and the child may sleep for a few minutes before waking confused and irritable.

Although a typical tonic-clonic attack is easily recognised, other forms of fits may be difficult to diagnose from the mother's history. Infantile spasms may begin with momentary episodes of loss of tone, which can occur in bouts and be followed by fits in which the head may suddenly drop forward or the whole infant may move momentarily like a frog. Recurrent episodes with similar features, whether they are changes in the level of consciousness or involuntary movements, should raise the possibility of fits.

Box 24.1 Features of a tonic clonic convulsion

Tonic

- Cry
- Loss of consciousness
- Rigidity
- Apnoea

Clonic

- Repetitive limb movement (rate can be counted)

Sleep

Box 24.2 Dangers

- Inhalation of vomit
- Hypoxaemia

Differential diagnosis

Convulsions must be differentiated from blue breath holding attacks, which usually begin at 9–18 months. Immediately after a frustrating or painful experience infants cry vigorously and suddenly hold their breath, become cyanosed, and in the most severe cases lose consciousness. Rarely, their limbs become rigid and there may be a few clonic movements lasting a few seconds. Respiratory movements begin again and infants gain consciousness immediately. The attacks diminish with age with no specific treatment. The mother may be helped to manage these extremely frightening episodes by being told that the child will not die and that she should handle each attack consistently by putting the child on his side.

Rigors may occur in any acute febrile illness, but there is no loss of consciousness.

Box 24.3 Breath holding attack

Pain or frustration
? Breath holding attack

Febrile convulsions

A febrile convulsion is a fit occurring in a child aged from six months to five years, precipitated by fever arising from infection outside the nervous system in a child who is otherwise neurologically normal.

Convulsions with fever include any convulsion in a child of any age with fever of any cause. Among children who have convulsions with fever are those with pyogenic or viral meningitis, encephalitis, or cerebral palsy with intercurrent infections. Children who have a prolonged fit or who have not completely recovered within one hour should be suspected of having one of these conditions.

Most of the fits that occur between the ages of six months and five years are simple febrile convulsions and have an excellent prognosis.

Box 24.4 Simple febrile convulsions

All the following:

- < 20 minutes
- No focal features
- Six months to five years
- No developmental or neurological abnormalities
- Not repeated in the same episode
- Complete recovery within one hour

By arbitrary definition, in simple febrile convulsions the fit lasts less than 20 minutes, there are no focal features, and the child is aged between six months and five years and has been developing normally.

Often fever is recognised only when a convulsion has already occurred. Febrile convulsions are usually of the tonic-clonic type. The objective of emergency treatment is the prevention of a prolonged fit (lasting over 20 minutes), which may be followed by permanent brain damage, epilepsy, and developmental delay.

An electroencephalogram (EEG) is not a guide to diagnosis, treatment, or prognosis.

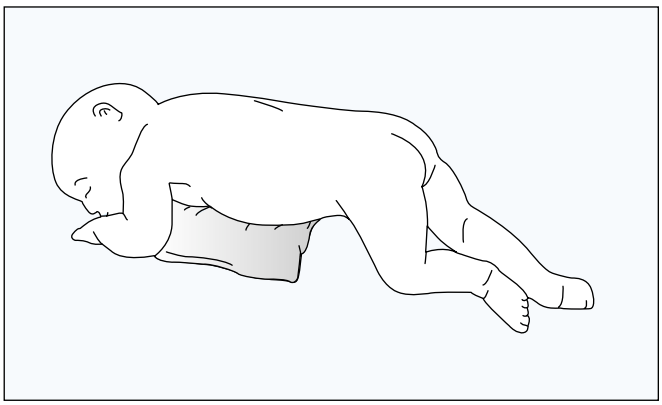


Figure 24.1 Position should be on the side or prone with the head to one side.

Emergency treatment

A child who has fever should have all his clothes removed and should be covered with a sheet only. He should be nursed on his side or prone with his head to one side because vomiting with aspiration is a constant hazard.

Rectal diazepam (0.5 mg/kg) produces an effective blood concentration of anticonvulsant within 10 minutes. The most convenient preparation resembles a toothpaste tube (Stesolid). Early admission to hospital or transfer to the intensive care unit should be considered if a second dose of anticonvulsant is needed.

Figure 24.2 Rectal diazepam tube.

All children who have had a first febrile convulsion should be admitted to hospital to exclude meningitis and to educate the parents, as many fear that their child is dying during the fit. Physical examination at this stage usually does not show a cause for the fever, but a specimen of urine should be examined in the laboratory to exclude infection and a blood glucose test should be performed. Blood should be taken for blood culture and plasma glucose and calcium estimations. Most of these children have a generalised viral infection with viraemia. A febrile convulsion may occur in roseola at the onset and three days later the rash appears. Occasionally, acute otitis media is present, in which case an antibiotic is indicated, but most children with febrile convulsions do not need an antibiotic. A pupuric rash suggests meningococcal septicaemia and the need for penicillin to be given immediately either intravenously or intramuscularly (see p. 91).

Lumbar puncture – A lumbar puncture should be performed if the child is under 18 months old or any of the following are present:

- signs of meningism such as neck stiffness
- drowsiness, irritability, or systemic illness
- complex convulsion that contains any feature that does not conform with the definitions of a simple convulsion.

Ideally, the decision should be taken by an experienced doctor, who may decide on clinical grounds that lumbar puncture is unnecessary even in a younger child, but when in doubt the investigation should be performed. The doctor deciding not to undertake a lumbar puncture should review the patient personally within a few hours. Children less than two years of age may have meningitis with *no* neck stiffness or other specific signs.

A child who has had severe vomiting or is in coma must be examined by an experienced doctor before lumbar puncture because of the risk of coning.

Management of fever – There is no evidence that antipyretic treatment influences the recurrence of febrile convulsions, but fever should be treated to promote the comfort of the child and to prevent dehydration. The child's clothes should be taken off and he should be covered with a sheet only.

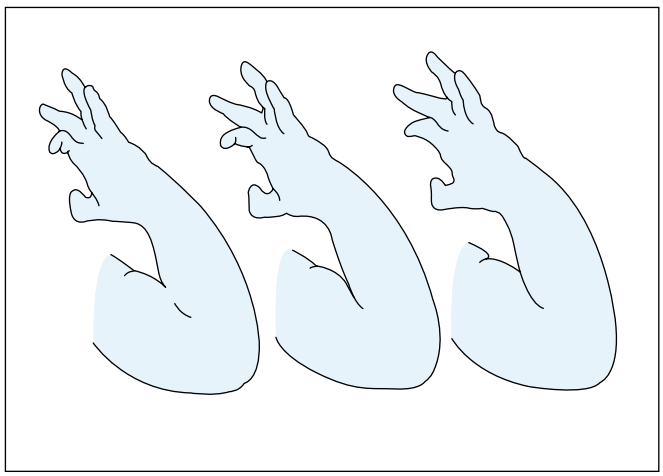


Figure 24.3 Repeated similar movements indicate a convulsion.

Paracetamol is the preferred antipyretic and adequate fluid should be given.

Failure to respond to paracetamol is an indication to add ibuprofen to the paracetamol or to give ibuprofen alone.

Anticonvulsant drugs – Rectal diazepam should be used as soon as possible after the onset of the convulsion. The parents should be advised not to give it if the convulsion has stopped.

The indications for long term anticouvulsant prophylaxis have changed and the sole indication – frequent recurrences, which should be treated with phenobarbitone – is rare. There is no evidence that in the minority of children who later develop epilepsy the prophylactic use of anticonvulsant drugs would have prevented it.

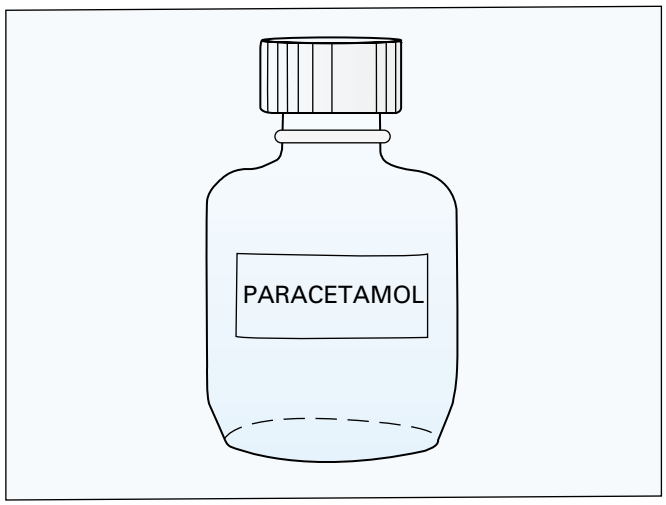


Figure 24.4 Paracetamol is the preferred antipyretic.

Immunisation

As routine immunisation is given to children 2–4 months old, this schedule is usually completed before febrile convulsions occur. Babies having convulsions with fever aged under six months should be assessed by a paediatrician. Children who have febrile convulsions before immunisation against diphtheria, pertussis, and tetanus because the immunisation has been delayed should be immunised after their parents have been instructed about the management of fever and the use of rectal diazepam.

Measles, mumps, and rubella immunisation should be given as usual to children who have had febrile convulsions, with advice about the management of fever to the parents. Rectal diazepam should be made available for use should a convulsion occur.

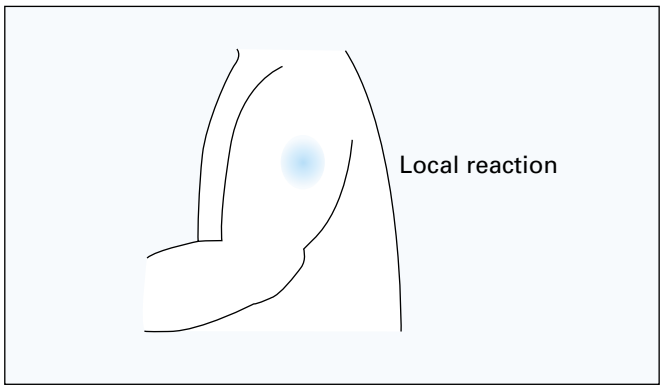


Figure 24.5 Immunisation may be followed by a febrile convulsion.

Prognosis

Unless there is clinical doubt about the child's current developmental or neurological state, parents should be told that prognosis for development is excellent. The risk of subsequent epilepsy after a single febrile convulsion with no complex features is about 1%. With each additional complex feature the risk rises to 13% in those children with two or more complex features. Only about 1% of children with febrile convulsions are in this group.

The risk of having further febrile convulsions is about 30%. This risk increases in younger infants and is about 50% in infants aged under one year at the time of their first convulsion. A history of febrile convulsions in a first degree relative is also associated with a risk of recurrence of about 50%. A complex convulsion or a family history of epilepsy is probably associated with an increase in the risk of further febrile convulsions.

Box 24.5 Features of complex febrile convulsions

- Lasting > 20 minutes
- Focal
- More than one on the same day
- Developmental or neurological abnormalities

Information for parents

Information for parents should include:

- an explanation of the nature of febrile convulsions, including information about the prevalence and prognosis
- instructions about the management of fever, the management of a convulsion, and the use of rectal diazepam
- reassurance.

Verbal and written advice should be given (see p. 96).

Basic life support in the community

Basic life support (BLS) should be given immediately whenever an infant stops breathing, even if no specialised equipment is available. It may be life saving.

Box 24.6 ABC of basic life support

- Airway
- Breathing
- Circulation

Advice to parents: Febrile convulsions

Your child has had a febrile convulsion. We know it was a very frightening experience for you. You may have thought that your child was dead or dying, as many parents think that when they first see a febrile convulsion. Febrile convulsions are not as serious as they appear.

What is a febrile convulsion?
It is an attack brought on by fever in a child aged between six months and five years.

What is a convulsion?
A convulsion is an attack in which the child becomes unconscious and usually stiff, with jerking of the arms and legs. It is caused by unusual electrical activity of the brain. The words convulsion, fit, and seizure have the same meaning.

What shall I do if my child has another convulsion?
Lay him on his side, with his head on the same level or slightly lower than the body. Note the time. Do not try to force anything into his mouth. Do not slap or shake the child.
The hospital may give you medicine to insert into your child's bottom. This is called rectal diazepam. If the convulsion has not stopped by the time that you have found the tube, insert it into the child's bottom and express the contents of the tube. This treatment should stop the convulsion within 10 minutes. If it does not, take your child to the hospital. You may need to dial 999 to obtain an ambulance. Let your doctor know what has happened.

About one child in 30 will have had a febrile convulsion by the age of five years.

Is it epilepsy?
No. The word epilepsy is applied to fits without fever, usually in older children and adults.

Do febrile convulsions lead to epilepsy?
Rarely; 99 out of 100 children with febrile convulsions never have convulsions after they reach school age, and never have fits without fever.

Do febrile convulsions cause permanent brain damage?
Almost never. Very rarely, a child who has a very prolonged febrile convulsion lasting half an hour or more may suffer permanent damage from it.

What starts febrile convulsions?
Any illness that causes a high temperature, usually a cold or other virus infection

Will it happen again?
Three out of 10 children who have a febrile convulsion will have another one. The risk of having another febrile convulsion falls rapidly after the age of three years.

Does the child suffer discomfort or pain during a convulsion?
No. The child is unconscious and unaware of what is happening.

What shall I do if my child has fever?
You can take the child's temperature by placing the bulb of the thermometer under his armpit for three minutes with his arm held against his side. Keep him cool by taking off his clothes and reducing the room temperature. Give plenty of fluids to drink. Give children's paracetamol medicine to reduce the temperature. The following doses should be given:

Up to 1 year old	one 5 ml spoonful (120 mg)
Aged 1 to 3 years	two 5 ml spoonfuls (240 mg)
Aged 4 years and over	three 5 ml spoonfuls (360 mg).

Repeat the dose every four hours until the temperature falls to normal, and then every six hours for the next 24 hours.

If the child seems ill or has ear ache or sore throat, let your doctor see him or her in case any other treatment, such as an antibiotic, is needed. Antibiotics are not necessary for most children with fever due to virus infections.

Is regular treatment with tablets or medicine necessary?
Usually not. The doctor will explain to you if your child needs regular medicine.

Adapted from a pamphlet produced by the British Paediatric Association

Assessment of condition

The level of consciousness of the infant should be quickly assessed by gentle pressure on the shoulders or limbs together with a verbal command. Even young infants may open their eyes or move in response to sound if they are not unconscious. Young infants should not be shaken as this may result in brain haemorrhage from trauma caused by the mobile brain hitting the inside of the skull. If the infant is not responsive, then the ABC of basic life support should be followed.

Box 24.7 Sequence of basic life support

- Call for help
- Check environment is safe
- Assess condition
- **Airway**
- **Breathing**
- **Circulation**

Airway

The rescuer can assess whether the infant is breathing by placing their face closely above the infant's face in order to look, listen, and feel for breathing and chest movement. If the infant is not breathing, the head of the infant should be maintained in the neutral position (head in line with body) by gentle pressure with one of the rescuer's hands on the forehead and chin lift or jaw thrust under the angles of the mandible with the other hand. Flexion or overextension of the neck of an infant may cause obstruction of the upper airway by the tongue blocking the pharynx. Attempts to improve an obstructed airway with the rescuer's finger are not recommended and may be dangerous.

Breathing

The rescuer should give five exhaled breaths, each lasting approximately one second into the infant's mouth or mouth and nose. The chest of the infant should be seen to expand with each breath or else the airway is not clear and airway opening manoeuvres should be repeated.

Circulation

After five initial breaths, the pulse should be assessed by palpation of the brachial artery in the medial aspect of the antecubital fossa or femoral artery in the groin. The carotid artery is difficult to palpate in infants because the neck is often short and fat. If the pulse is absent or < 60 beats per minute as assessed over 10 seconds, then external cardiac compressions should be commenced.

External cardiac compression

Cardiac compressions should be applied with the infant lying flat on the back on a firm surface. There are two methods of compressions in small infants.

- Hands encircling the thorax with the thumbs compressing the sternum one finger breadth below an imaginary line drawn between the nipples. This method is only possible in very small infants and is difficult for a lone rescuer to perform along with mouth to mouth with or without nose breaths.
- Two finger technique with the ball (not tips) of two fingers compressing the chest with the fingers again placed one finger breadth below the imaginary line between the nipples.

Whichever method is used, the chest should be compressed by about one-third depth in order to propel blood to the coronary arteries.

Continue CPR

Cardiopulmonary resuscitation (CPR) should be continued with a ratio of five chest compressions to one breath for approximately one minute (20 cycles) before reassessment. It is not necessary to simulate the heart rate of a normally breathing infant and, indeed, it may not be possible to do so in an effective manner. The rescuer should continue CPR until the emergency services arrive and take over resuscitation or if no help has arrived within a few minutes, the infant should be carried quickly to where help can be summoned.

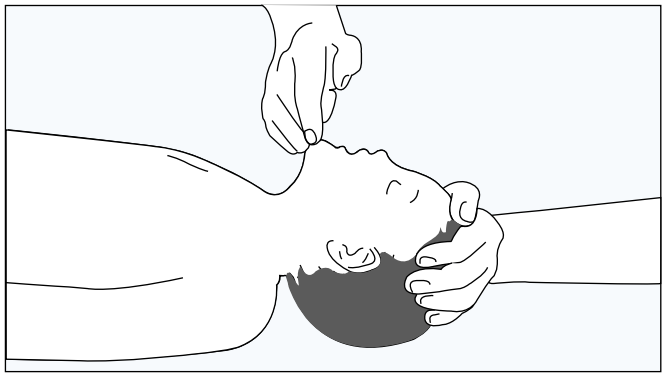


Figure 24.6a Chin lift in an infant.

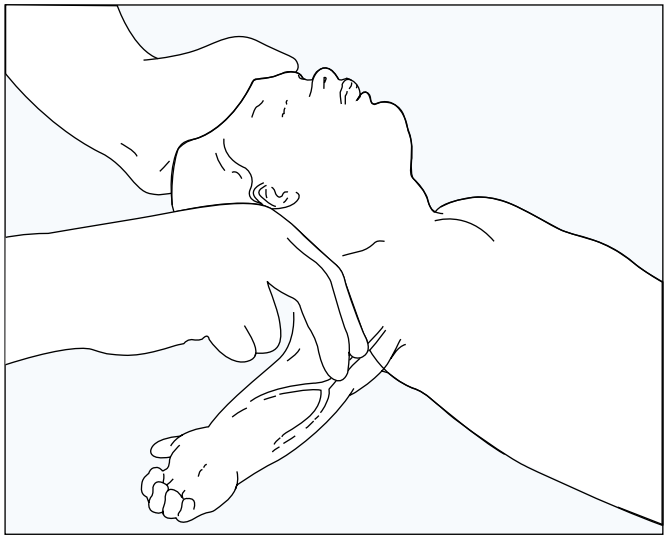


Figure 24.6b Feeling for the brachial pulse.

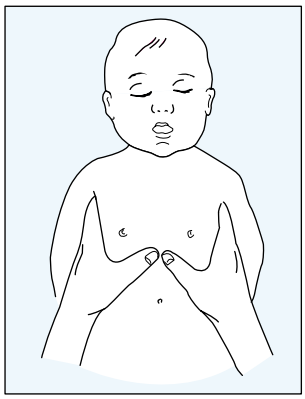


Figure 24.6c External cardiac compression.

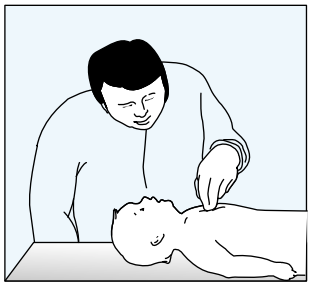


Figure 24.6d External cardiac compression.

25 Crying babies

Apart from the subtle method of eye to eye communication, babies have no other way of signalling their needs to their mothers than by crying. Babies who are quiet or "good" may be abnormal or ill. As crying is a non-specific call for help, obvious needs should be investigated. In most cases no remediable cause is found and the problem may resolve spontaneously when infants are about three months old. By then, however, parents may have become exhausted from lack of sleep and marital discord may follow while the infant remains fresh. As well as suggesting any appropriate management, the family doctor should encourage parents to take one or two evenings off together each week to visit friends or go to a film. Parents often worry that the infant is suffering from lack of sleep and wrongly ascribe poor appetite or frequent colds to this cause.

Sound spectrography has confirmed the clinical impression that cries of cerebral irritability, pain, and hunger are different in quality, but this method is not available to most clinicians. During the first month infants do not shed tears when they cry and even after the age of about six months many infants cry at night without shedding tears.

The doctor needs to take a full history, including details of the pattern of crying, when the problem began, and measures taken to resolve it. It should be possible to determine whether the infant has always needed little sleep or whether he has developed a habit of crying in order to get into his parents' comfortable bed. The doctor should also explore the reason why the parents have sought advice at this stage. The mother should be asked about any change in the house, where the infant sleeps, and who looks after him during the day. Illnesses in the child or family and marital and social backgrounds also need considering. A physical examination usually shows no abnormality, but occasionally there may be signs of acute otitis media.

Difficulty in getting to sleep may be avoided by starting a bedtime ritual in infancy. A warm bath followed by being wrapped in particular blankets may later be replaced by the mother or father reading from a book or singing nursery rhymes before the light is turned out. A soft cuddly toy of any type can lie next to the infant from shortly after birth, and seeing this familiar toy again may help to induce sleep.

The parents should be told that during the night babies often open their eyes and move their limbs and heads. They should be asked to resist getting up to see their baby, as the noise of getting out of bed may wake him and he may then remain awake. Babies who do wake may be pacified with a drink and may then fall asleep. The drink is to provide comfort rather than to assuage any thirst.

Parents whose young children sleep a great deal during the day can discourage them from doing this by taking them out shopping or giving them other diversions, and they may then sleep well at night.

Sleep disturbance is a common reaction to the trauma of admission to hospital or moving house, and taking the child into the parents' room for a few weeks may help to reassure the infant that he has not been abandoned.

If the infant is prepared to go to sleep at a certain time but the parents would like to advance it by an hour, they can put him to bed five minutes earlier each night until the planned bedtime is achieved.

Figure 25.1 Crying signals needs.

Figure 25.2 A bath is a bedtime ritual.

Figure 25.3 Father reading from a book.

Hunger and thirst

Babies often cry and become restless just before a feed is due and in the early weeks of life the infant may demand to be fed three hourly or even two hourly. This is perfectly normal. Preterm infants need particularly high milk intakes for "catch-up" growth and they may become less demanding when they reached their final centiles for weight.

The volumes of bottled milk recommended are only average amounts and many infants will therefore need more to prevent excessive hunger. Regular weighing and use of a growth chart to plot the results are the only ways of ensuring that the infant is obtaining the correct amount of feeds. Feeding at regular intervals is apt to cause crying in many babies and demand feeding is preferable. Some infants scream when they approach the breast and the help of an experienced midwife is needed to overcome this problem.

The age at which infants sleep through the night and do not require a night feed varies greatly. It usually occurs when infants weigh about 5 kg and parents may be worried because their neighbour's child of a similar age has already omitted the night feed. Similar competition between mothers over which baby first manages three meals a day may also lead to inadequate feeding, followed by crying.

It is impossible to determine the difference between a cry due to hunger and a cry due to thirst. The only way of solving this problem is to offer infants water twice daily. This can easily be given mid-morning or mid-afternoon. Some infants dislike the taste of boiled tap water so a little fruit juice or half a teaspoon of sugar can be added to 100 ml of water.

Even newborn infants cry to obtain physical contact and the crying stops as soon as they are picked up. It does not spoil babies to pick them up when they want contact. Once they are a few weeks old infants begin to have periods of waking and some will not tolerate being left in a pram or cot by themselves. A harness that allows them to remain in contact with their mother's body will often satisfy these infants. Alternatively, a canvas chair can be used even as early as one month and can be placed in a safe position in the kitchen or wherever the mother is working.

If babies do not stop crying when held, walking with them, rocking them, or taking them for a car ride may stop the crying.

Wrapping infants in a blanket and putting them firmly into a carrycot or a very small cot may also provide contact, although by an impersonal method.

Non-nutritive sucking stops crying, provided babies are not hungry. Putting a dummy into the mouth of a crying baby who is not hungry may be effective as a temporary measure.

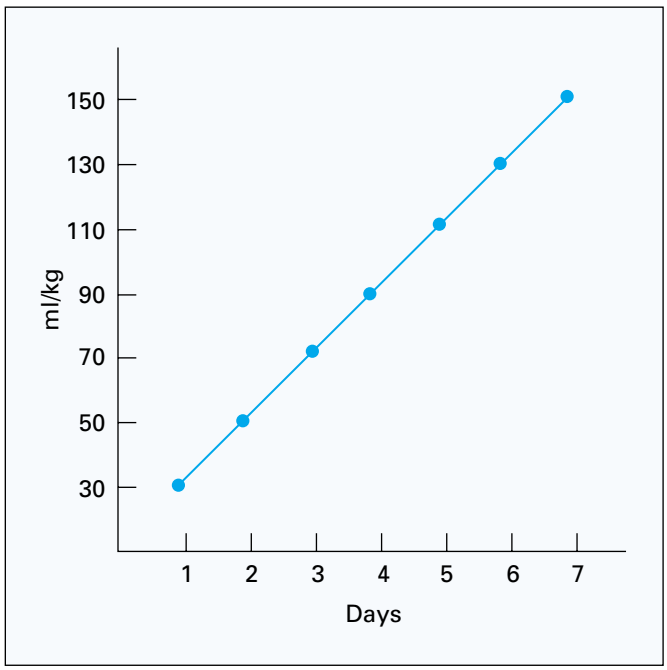


Figure 25.4 Milk intake increases with age.

Figure 25.5 Close contact often stops crying.

Three months' colic and feeding problems

Three months' colic is the name given to a syndrome that usually begins at about two weeks of age and has usually stopped by four months. The infant screams, draws up his legs, and cannot be comforted by milk, being picked up, or having his napkin changed. This usually occurs in the early evening, when the mother is busy getting supper ready, has less time to play with the baby, and may be anxious. Breastfeeding mothers may have less milk at that time of the day.

Some babies swallow excessive air during feeding as a result of milk spurting from the breast or a very small hole in the teat of the bottle. Mothers usually complain that these infants gulp at the beginning of the feed and they have particular difficulty in bringing up the wind. This problem does not occur in

societies where babies are held on the mothers' backs for most of the day. The infant's abdomen is pressed against the mother's back and the infant is kept upright, which is the best position for releasing excessive gastric air.

If the milk spurts from the breast, the first 25 ml or so should be expressed manually and given to the infant later if necessary. If the baby is fed by bottle, the size of the hole in the bottle teat should be checked and made larger with a hot needle if necessary. The infant can be winded most easily by putting him prone with his head higher than his feet, but he must be observed constantly while in this position.

Teeth and teething

Although mothers consider that the eruption of the first tooth is a milestone in development, the age at which this occurs is of no practical importance. The first teeth to appear, at 6–12 months, are the lower incisors.

From the age of a few weeks infants normally put their fingers, and later anything else that comes to hand, into their mouths and mothers often wrongly ascribe this to teething.

Teething produces only teeth. It does not, contrary to common belief, cause convulsions, bronchitis, or napkin rash. Some mothers insist their infants are particularly irritable when they are teething, but it is important to examine the infant to exclude disease such as otitis media or meningitis before accepting the mother's explanation.

Dummies temporarily affect the growth of the mouth but there are no objections to using them. They should not be dipped in honey. Severe dental caries also follows the use of "comforters" or "dinky-feeders", which are filled with fluid containing sugar.

Tetracycline or its derivatives should never be given to children less than eight years of age, as permanent brownish yellow staining of the teeth may occur.

Other causes

Illness – If the infant starts crying persistently, especially if he has not cried like this before, disease should be suspected. Acute otitis media often causes persistent crying at night, but meningitis or a urinary tract infection may have no specific signs. An intussusception is an invagination of a proximal part of the gut into a more distal portion and may result in partial or complete intestinal obstruction. The infant has sudden attacks of screaming and pallor that last for a few minutes and recur every 10–20 minutes. Between attacks the infant appears normal. Bloodstained mucus may be passed rectally. The abdomen is tender only during attacks, but often there is a mass over the course of the colon. If intussusception is suspected the infant should be admitted to hospital (see p. 51).

Mechanical problems – Babies do not seem distressed when their napkins are wet or soiled, though some cry when they are being undressed or changed. An open safety pin is an extremely rare cause of crying.

Fatigue – When young infants have been on a long journey or have been stimulated more than usual by playing with relatives, they may become irritable and not keen to sleep.

Failure of mother–infant adjustment – Crying may be a sign of this failure, which may be related to the character of the infant as much as the response of the mother. The crying may represent the difficulties of each adjusting to the other. If an infant has been crying persistently for several weeks the possibility of severe depression in the mother should be

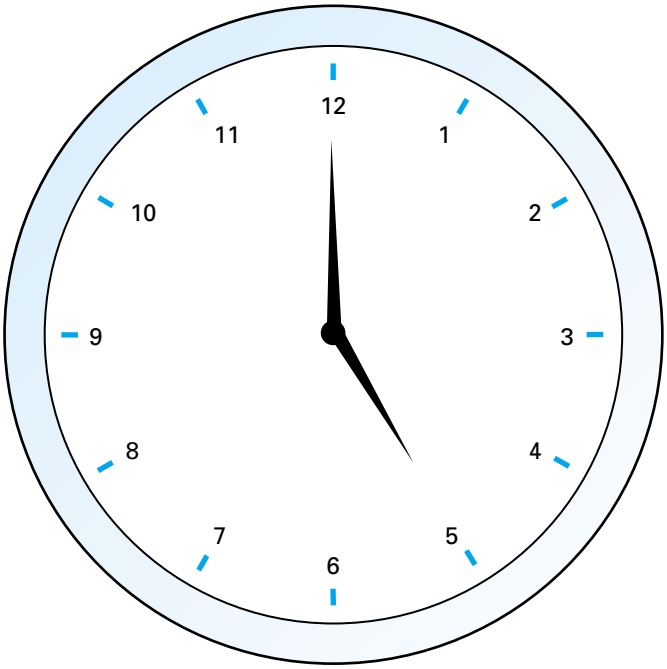


Figure 25.6 Evening colic begins about 5 pm.

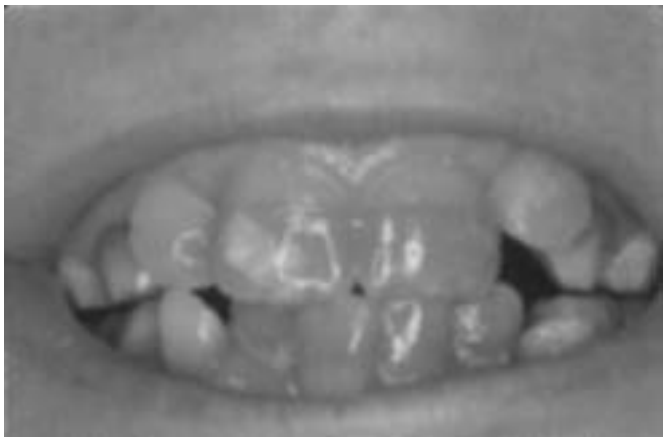


Figure 25.7 Severe discoloration due to tetracycline.

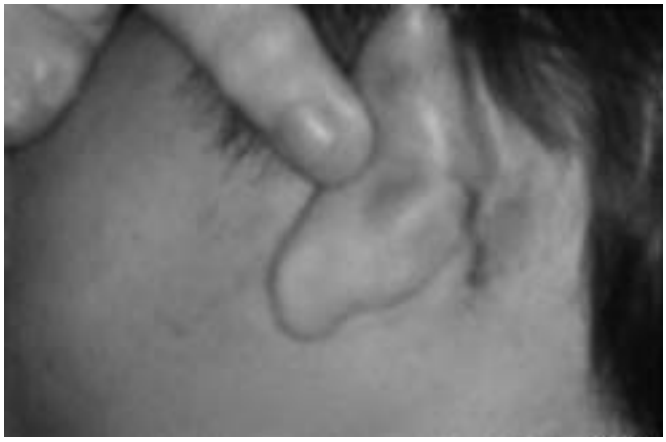


Figure 25.8 Non-accidental injury behind the ear.

considered. This depression could be the cause or the result of the crying. The mother is exhausted by lack of sleep and particularly liable to injure her infant, and both mother and infant may need to be admitted to hospital for diagnosis and management.

For most infants none of the above reasons for crying are present. The parents need reassuring strongly after examination that their infant is healthy and that the problem will resolve spontaneously. Sedation of the infant with chloral hydrate for a week during a difficult period may enable the mother to regain some of her strength. The drug should not be used for longer, since it loses its effect. Only a few doses should be prescribed, as the mother may take them herself or give too many to her infant in a depressive episode.

Special problems after the age of three months

Loneliness is probably the commonest cause of crying after the age of three months. While they are awake during the day some infants will not tolerate being left alone but are happy if they are left in the room where the mother is working. These infants are extremely interested in their surroundings and need the stimulation of things going on around them. Similarly, infants need to be propped up when out in their prams and when they begin to reach out they need simple toys to play with.

Separation from the mother during admission to hospital or when parents go on holiday by themselves may be followed by bouts of crying. An infant can relate to only two or three adults at a time and frequent changes in caretaker may be associated with crying.

After about five months infants may cry when a stranger approaches them or when they sleep in a different room. They may wake suddenly and appear terrified, which might be due to a nightmare, although there is no method of proving this.

Infants who do not sleep after three months

During the first few weeks of life some babies sleep almost continuously for the 24 hours whereas others sleep for only about 12 hours. Many mothers consider that a newborn baby should sleep continuously and do not realise that babies take an interest in what is going on around them. Infants who need little sleep may wake regularly at 2 or 4 am and remain awake for two or three hours. Some are settled by a drink, but others cry persistently after this and the mother may take the infant into her own bed before they go to sleep. Some infants wake in a similar way but are then content to remain awake looking at mobiles above their cots or playing with toys left in the cot.

Behaviour modification and drugs

When infants wake frequently during the night and cry persistently until they are taken into their parents' bed, a plan of action is needed. If there is an obvious cause, such as acute illness or recent admission to hospital, the problem may resolve itself within a few weeks, and at first there need be no change in management. If there is no obvious cause, the parents are asked to keep a record of the infant's sleep pattern for two weeks (see Figure 24.10). This helps to determine where the main problem lies and can be used as a comparison with treatment.

Both parents are seen at the next visit; both need to accept that they must be firm and follow the plan exactly. Behaviour

Figure 25.9 A chair in the kitchen reduces loneliness.

Figure 25.10 Playing with toys suspended across the cot.

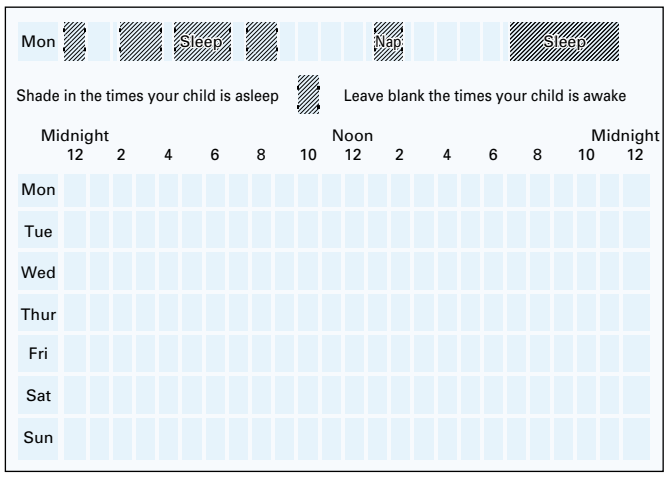


Figure 25.11 Record of sleep pattern.

modification is the only method that produces long term improvement, but it can be combined with drugs initially if the mother is at breaking point.

Behaviour modification separates the mother from the child gradually or abruptly, depending on the parents' and doctor's philosoply. The slow method starts with the mother giving a drink and staying with the child for decreasing lengths of time. In the next stage no drink is given. Then she speaks to the child through the closed door and, finally, does not go to the child at all. The abrupt method consists of letting the child cry it out; he stops after three or four nights. There are an infinite number of variations between these extremes and the temperament of the parents, child, and doctor will determine what is acceptable.

Another approach is to increase the waiting time before going to the infant. In severe cases a written programme of several small changes can be given to the mother and she can be seen again by the health visitor or family doctor after each step has been achieved. The mother will need to be reassured that the infant will not develop a hernia from crying or vomit or choke, and neighbours may be pacified by being told that the child will soon be cured.

Many sleep problems can be resolved without drugs, but some mothers are so exhausted by loss of sleep that they cannot manage a programme of behaviour modification unless the infant receives some preliminary sedation. The most satisfactory drug for this age group is chloral hydrate 30 mg/kg body weight given one hour before going to bed. The full dose is given for two weeks, followed by a half dose for a week; the drug is then given on alternate nights for a week. The objective is to change the pattern of sleeping. A behaviour modification plan is needed during the third and subsequent weeks.

Day	Time to bed	Time to sleep	First problem	What did you do?	Second problem	What did you do?	Time woke up in morning
Mon							
Tue							
Wed							
Thur							
Fri							
Sat							
Sun							

Figure 25.12 Record of action by parents.

		If your child is still crying		
Day	At first episode	Second episode	Third episode	Subsequent episodes
1	5	10	15	15
2	10	15	20	20
3	15	20	25	25
4	20	25	30	30
5	25	30	35	35

Figure 25.13 Number of minutes to wait before going into your child briefly.

26 Non-accidental injury and selected drugs

Non-accidental injury

This section has been restricted to physical abuse, as sexual abuse is rare in infants less than a year of age.

Non-accidental injury inflicted by adults on children is often hard to detect. Symptoms that are difficult to explain may be the result of inflicted injury. In Britain as many as 100 children a year may die from non-accidental injuries.

Non-accidental injury should be suspected, especially in a child aged under three, whenever: (*a*) there has been a delay between the accident occurring and the parents seeking medical help; (*b*) the explanation of the injury is inadequate, discrepant, or too plausible; (*c*) the child or sibling has a history of non-accidental or suspicious injury; (*d*) there is evidence of earlier injury; (*e*) the child has often been brought to the family doctor or accident department for little apparent reason; (*f*) the parents show disturbed behaviour or unusual reactions to the child's injuries or have a history of psychiatric illness; (*g*) the child shows obvious neglect or failure to thrive.

Typical initial injuries are: (*a*) burns, abrasions, or small bruises on the face; (*b*) injuries to the mouth or torn fraenulum of the tongue; (*c*) bruises caused by shaking or rough handling, including finger-shaped bruises; (*d*) subconjunctival or retinal haemorrhages. If the doctor suspects non-accidental injury, the child should be undressed completely and examined fully, with all signs of injury noted. If the doctor's suspicions are not allayed, he or she should ensure the immediate examination of the child by a consultant paediatrician. The family doctor should, if possible, not mention any suspicions to the parents, as they may refuse to allow this examination and it may impair his relationship with them.

If parents refuse to allow their child to be seen or want to remove the child from hospital too soon, the social services department, the police or National Society for the Prevention of Cruelty to Children may seek an emergency protection order, which allows the child to be kept in hospital for eight days.

Children with genuinely accidental injuries will occasionally be admitted unnecessarily, but this should not deter any doctor from admitting a child when there is reasonable doubt about the cause of an injury.

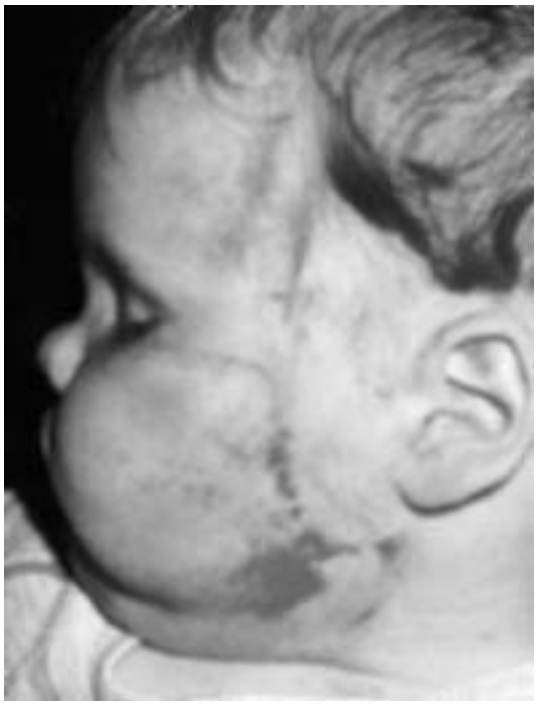


Figure 26.1 Non-accidental injury.

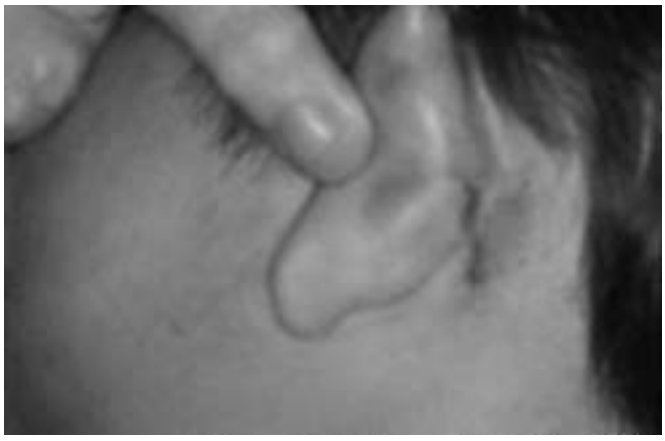


Figure 26.2 Non-accidental injury behind ear.

Preventable causes of death

Six preparations of drugs are especially associated with avoidable deaths in children.

Ampoules of digoxin for use in adults should never be supplied to children's wards. Ampoules of digoxin made especially for children ensure that a fatal overdose cannot be given.

Penicillin given intrathecally has no therapeutic value. Errors are often made in dispensing the correct dose of penicillin for intrathecal use, with fatal results. This route should therefore never be used for penicillin.

Excessive doses of chloramphenicol can easily be given if the special infant preparation is not used.

Intravenous glucose solution should never be given at a concentration higher than 20%. Sodium bicarbonate solution should never be given intravenously at a concentration higher than 4.2%. Higher concentrations of either of these solutions

Box 26.1 Drugs which are potentially dangerous

- Digoxin
- Penicillin intrathecally
- Chloramphenicol
- Glucose > 20%
- Sodium bicarbonate > 4.2%
- Potassium chloride solution

are severely hypertonic and may be followed by sudden death or permanent brain damage.

A bag of intravenous fluid to which potassium chloride solution has been added should be shaken thoroughly to avoid giving a bolus of potassium chloride solution, which may cause a fatal arrhythmia.

Selected drugs in the newborn

Intestinal absorption is variable and regurgitation of antibiotics common, so the intramuscular or intravenous route should usually be used initially. Intramuscular injections are given into the upper lateral aspect of the thigh. Schemes for rotation of sites are essential to prevent local necrosis and to avoid further injections being given into a relatively avascular area. Drugs are usually given intramuscularly or orally every eight hours and intravenously by slow bolus injection every 12 hours. In preterm infants an adjustment of dose and frequency of administration may be needed and specific texts should be consulted.

Amoxycillin	(oral, IM, or IV)	50–100 mg/kg/24 h
Ampicillin	(oral, IM, or IV)	50–100 mg/kg/24 h
Ceftazidime	(IM or IV)	60–100 mg/kg/24 h
Chloral hydrate	(oral)	30–50 mg/kg per dose
Diazepam	(IM or IV)	200–300 microgram/kg per dose
Flucloxacillin	(oral, IM, or IV)	50–100 mg/kg/24 h
Gentamicin	(IM or IV)	7 mg/kg/24 h (12–18 hourly). Blood concentration *must* be checked
Miconazole gel	(oral)	2.5 ml twice daily
Naloxone	(IM)	60 microgram/kg per dose
Nystatin	(oral)	100 000 units/dose (after feeds)
Paraldehyde	(IM)	0.1–0.2 ml/kg per dose
Benzylpenicillin	(IM or IV)	50–100 mg/kg/24 h
Phenobarbitone	(IM or IV)	15 mg/kg once only followed by 3 mg/kg per dose 12 hourly
Phenytoin	(IV)	15 mg/kg per dose slowly once followed by 3 mg/kg per dose 12 hourly
	(oral)	5 mg/kg/24 h
Sodium ironedetate	(oral)	1 ml twice daily
Triclofos	(oral)	50 mg/kg per dose
Trimethoprim	(oral)	4 mg/kg/24 h

Figure 26.3 Wrapping the infant may prevent spillage of medicine.

Figure 26.4 Special medicine syringes (which have no needles) can be used to measure and give small volumes accurately.

Selected drugs in older infants

Infants treated at home and most of those in hospital need oral drugs three times a day. The drugs should be given before feeds and the infant need not be woken specially for them. Some preparations, particularly the penicillins, have an unpleasant taste and the medicine should not be mixed with food as the infant may then hate both. Syrup, which is a sucrose solution that forms the base of most elixirs, may cause dental caries if it is given regularly for a long time. If only one adult is present to give the medicine, wrapping the infant securely in a blanket may prevent spillage. Special medicine syringes (which have no needles) can be used to measure and give small volumes accurately.

Amoxycillin	50–100 mg/kg/24 h
Ampicillin	100 mg/kg/24 h
Betamethasone	3 mg per dose intramuscularly or intravenously = 100 mg hydrocortisone
Cefaclor	40 mg/kg/24 h
Cefotaxime	100 mg/kg/24 h intramuscularly or intravenously (12 hourly)
Chloral hydrate	30–50 mg/kg/per dose
Diazepam	200–300 microgram/kg orally per dose or slowly intravenously
Erythromycin	30–50 mg/kg/24 h
Flucloxacillin	50–100 mg/kg/24 h
Gentamicin	7 mg/kg/24 h intramuscularly or intravenously
Hydrocortisone	100 mg per dose intramuscularly or intravenously
Nystatin	100 000 units per dose (four hourly)
Paracetamol	75 mg/kg/24 h
Paraldehyde	0.15 ml/kg per dose intramuscularly
Benzylpenicillin	(intramuscularly or intravenously) 100 mg/kg/24 h
Phenoxymethylpenicillin	60 mg/kg/24 h
Phenobarbitone	5 mg/kg/24 h
Phenytoin	5 mg/kg/24 h
Trimethoprim	1–2 mg/kg single daily prophylactic dose 8 mg/kg/24 h
Trimeprazine tartrate	3 mg/kg per dose once daily

Acknowledgements

We thank Blackwell Scientific Publications for allowing us to adapt liberally material that has appeared in *Practical Management of the Newborn, Accident and Emergency Paediatrics,* and *Paediatric Therapeutics* and use the two photographs of resuscitating a newborn.

We also thank Susan Thomason, Maire Sullivan, Richard Bowlby, Joanna Fairclough, Brian Pashley, Jeanette McKenzie, Derek De Witt, and Ann Shields of the department of medical illustration at Northwick Park Hospital for taking most of the photographs. The remaining photographs were supplied as follows:

Early prenatal diagnosis: ultrasound scans Dr H B Meire and Sujata Patel.
Breathing difficulties in the newborn: micrograph of respiratory distress syndrome, Dr G Slavin; apnoea monitor, Graesby.
Some congenital abnormalities: cleft lip, Mr R Saunders.
Routine examination of the newborn: cardiac ultrasound, Miss Varrela Gooch.
Dislocated and dislocatable hips in the newborn: the ultrasound scans were provided by Dr Dennis Remedios and the photograph of the Pavlik harness by Mr L Freedman.

For all photographs apart from those mentioned specifically above, Dr H B Valman and Dr R M Thomas retain the copyright.
We thank Mr R Lamont for providing extensive new material for the chapter on prenatal assessment. We are indebted to the Child Growth Foundation for allowing the use of data to construct the growth chart on p.66, and Dr Judith Wilson for the Harrow version of the personal health record which is used for recording health surveillance data.

Index

Page numbers in **bold** refer to figures in the text: those in *italic* to tables or boxed material