Ch02-1 Fundamentals of Medicine

01-1 Clinical decision-making

1 Clinical decision-making

2 • CLINICAL DECISION-MAKING Diagnostic error has been defined as ‘a situation in which the clinician has all the information necessary to make the diagnosis but then makes the wrong diagnosis’. Why does this happen? Studies reveal three main reasons: • knowledge gaps • misinterpretation of diagnostic tests • cognitive biases. Examples of errors in these three categories are shown in Box 1.2. Clearly, clinical knowledge is required for sound clinical reasoning, and an incomplete knowledge base or inadequate experience can lead to diagnostic error. However, this chapter focuses on other elements of clinical reasoning: namely, the interpretation of diagnostic tests, cognitive biases and human factors. Clinical reasoning: definitions ‘Clinical reasoning’ describes the thinking and decision-making processes associated with clinical practice. It is a clinician’s ability to make decisions (often with others) based on all the available clinical information, starting with the history and physical examination. Our understanding of clinical reasoning derives from the fields of education, cognitive psychology and studies of expertise. Figure 1.1 shows the different elements involved in clinical reasoning. Good clinical skills are fundamental, followed by understanding how to use and interpret diagnostic tests. Other essential elements include an understanding of cognitive biases and human factors, and the ability to think about one’s own thinking (which is explained in more detail later). Other key elements of clinical reasoning include patient-centred evidencebased medicine (EBM) and shared decision-making with patients and/or carers. Introduction A great deal of knowledge and skill is required to practise as a doctor. Physicians in the 21st century need to have a comprehensive knowledge of basic and clinical sciences, have good communication skills, be able to perform procedures, work effectively in a team and demonstrate professional and ethical behaviour. But how doctors think, reason and make decisions is arguably their most critical skill. Knowledge is necessary, but not sufficient on its

own for good performance and safe care. This chapter describes the principles of clinical decision-making, or clinical reasoning. The problem of diagnostic error It is estimated that diagnosis is wrong 10–15% of the time in specialties such as emergency medicine, internal medicine and general practice. Diagnostic error is associated with greater morbidity than other types of medical error, and the majority is considered to be preventable. For every diagnostic error there are a number of root causes. Studies of misdiagnosis assign three main categories, shown in Box 1.1; however, errors in clinical reasoning play a significant role in the majority of diagnostic adverse events. Fig. 1.1 Elements of clinical reasoning. (EBM = evidence-based medicine) Clinical reasoning Clinical skills (history and physical examination) Thinking about thinking Patient-centred EBM Shared decision-making Using and interpreting diagnostic tests Understanding cognitive biases and human factors Adapted from Graber M, Gordon R, Franklin N. Reducing diagnostic errors in medicine: what's the goal? Acad Med 2002; 77:981–992. 1.1 Root causes of diagnostic error in studies Error category Examples No fault Unusual presentation of a disease Missing information System error Inadequate diagnostic support Results not available Error-prone processes Poor supervision of inexperienced staff Poor team communication Human cognitive error Inadequate data-gathering Errors in reasoning 1.2 Reasons for errors in clinical reasoning Source of error Examples Knowledge gaps Telling a patient she cannot have biliary colic because she has had her gallbladder removed – gallstones can form in the bile ducts in patients who have had a cholecystectomy Misinterpretation of diagnostic tests Deciding a patient has not had a stroke because his brain scan is normal – computed tomography and even magnetic resonance imaging, especially when performed early, may not identify an infarct Cognitive biases Accepting a diagnosis handed over to you without question (the 'framing effect') instead of asking yourself 'What is the evidence that supports this diagnosis?'

value, the greater the probability). Similarly, an LR of less than 1 decreases the probability of disease. LRs are developed against a

diagnostic standard (e.g. in the case of meningitis, lumbar puncture results), so do not exist for all clinical findings. LRs illustrate how an individual clinical finding changes the probability of a disease. For example, in a person presenting with headache and fever, the clinical finding of nuchal rigidity (neck stiffness) may carry little weight in deciding

whether to perform a lumbar puncture because LRs do not determine the prior probability of disease; they reflect only how a single clinical finding changes it. Clinicians have to take all the available information from the history and physical examination into account. If the overall clinical probability is high to begin with, a clinical finding

with an LR of around 1 does not change this. ‘Evidence-based history and examination’ is a term used to describe how clinicians incorporate knowledge about the prevalence and diagnostic weight of clinical findings into their history and physical examination. This is important because an estimate of clinical probability is vital in

decision-making and the interpretation of diagnostic tests. Use and interpretation of diagnostic tests There is no such thing as a perfect diagnostic test. Test results give us test probabilities, not real probabilities. Test results have to be interpreted because they are affected by the following: • how ‘normal’ is defined • factors other than disease •

operating characteristics • sensitivity and specificity • prevalence of disease in the population. Normal values Most tests provide quantitative results (i.e. a value on a continuous numerical scale). In order to classify quantitative results as normal or abnormal, it is necessary to define a cut-off point. Many quantitative measurements in

populations have a Gaussian or ‘normal’ distribution. By convention, the normal range is defined as those values that encompass 95% of the population, or 2 standard deviations above and below the mean. This means that 2.5% of the normal population will have values above, and 2.5% will have values below the normal range. For this

reason, it is more appropriate to talk about the ‘reference range’ rather than the ‘normal range’ (Fig. 1.3). Test results in abnormal populations also have a Gaussian distribution, with a different mean and standard deviation. In some diseases there is no overlap between results from the abnormal and normal population. However, in

many diseases there is overlap; in these circumstances, the greater the difference between the test result and the limits of the reference range, the higher the chance that the person has a disease. However, there are also situations in medicine when ‘normal’ is abnormal and ‘abnormal’ is normal. For example, a normal $PaCO_2$ in

the context of a severe asthma attack is abnormal and means the patient has life-threatening asthma. A low ferritin in a young menstruating woman is not considered to be a disease at all. Normal, to some extent, is therefore arbitrary. Clinical skills and decision-making Even with major advances in medical technology, the history

remains the most important part of the clinical decision-making process. Studies show that physicians make a diagnosis in 70–90% of cases from the history alone. It is important to remember that a good history is gathered not only from the patient but also, if necessary (and with consent if required), from all available sources: for example,

paramedic and emergency department notes, eye-witnesses, relatives and/or carers. Clinicians need to be aware of the diagnostic usefulness of clinical features in the history and examination. For example, students are taught that meningitis presents with the following features: • headache • fever • meningism (photophobia,

nuchal rigidity). However, the frequency with which patients present with certain features and the diagnostic weight of each feature are important in clinical reasoning. For example, many patients with meningitis do not have classical signs of meningeal irritation (Kernig’s sign, Brudzinski’s sign and nuchal rigidity). In one prospective

study, they had likelihood ratios of around 1, meaning they carried little diagnostic weight (Fig. 1.2). Likelihood ratios (LR) are clinical diagnostic weights. An LR of greater than 1 increases the probability of disease (the higher the Fig. 1.2 Likelihood ratio (LR) of Kernig’s sign, Brudzinski’s sign and nuchal rigidity in the clinical diagnosis of meningitis. LR

probability of finding in patients disease probabilit

with y of finding in patients disease without LRs are also used for diagnostic tests; here a physical examination finding can be considered a diagnostic test. Data from Thomas KE, Hasbun R, Jekel J, Quagliarello VJ. The diagnostic accuracy of Kernig’s sign, Brudzinski’s sign, and nuchal rigidity in adults with suspected meningitis. Clin Infect Dis 2002; 35:46–52. Change in probability of disease

Infinity Zero

0.5 0.2 0.1

- 45%
- 30%
- 15% No change No change Kernig’s sign Brudzinski’s sign Nuchal rigidity Increase probability LR Decrease probability – 15% – 30% – 45%

4 • CLINICAL DECISION-MAKING Sensitivity and specificity Diagnostic tests have characteristics termed ‘sensitivity’ and ‘specificity’. Sensitivity is the ability to detect true positives; specificity is the ability to detect true negatives. Even a very good test, with 95% sensitivity, will miss 1 in 20 people with the disease. Every test therefore has ‘false positives’ and ‘false negatives’ (Box 1.4). A very sensitive test will detect most disease but generate abnormal findings in healthy people. A negative result will therefore reliably exclude disease but a positive result does not mean the disease is present – it means further evaluation is required. On the other hand, a very specific test may miss significant pathology but is likely to establish the diagnosis beyond doubt when the result is positive. All tests differ in their sensitivity and specificity, and clinicians require a working knowledge of the tests they use in this respect. In choosing how a test is used to guide decision-making there is a trade-off between sensitivity versus specificity. For example, defining an exercise electrocardiogram (p. 449) as abnormal if there is at least 0.5 mm of ST depression would ensure that very few cases of coronary artery disease are missed but would generate many false-positive results (high sensitivity, low specificity). On the other hand, a cut-off point of 2.0 mm of ST depression would detect most cases of important coronary artery disease with far fewer false positives. This trade-off is illustrated by the receiver operating characteristic curve of the test (Fig. 1.4). An extremely important concept is this: the probability that a person has a disease depends on the pre-test probability, and the sensitivity and specificity of the test. For example, imagine that an elderly lady has fallen and hurt her left hip. On examination, Factors other than disease that influence test results A number of factors other than disease influence test results: • age • ethnicity • pregnancy • sex • spurious (in vitro) results. Box 1.3 gives some examples. Operating characteristics Tests are also subject to operating characteristics. This refers to the way the test is performed. Patients need to be able to comply fully with some tests, such as spirometry (p. 569), and if they cannot, then the test result will be affected. Some tests are very dependent on the skill

of the operator and are also affected by the patient's body habitus and clinical state; ultrasound of the heart and abdomen are examples. A common mistake is when doctors refer to a test result as 'no abnormality detected' when, in fact, the report describes a technically difficult and incomplete scan that should more accurately be described as 'non-diagnostic'. Some conditions are paroxysmal. For example, around half of patients with epilepsy have a normal standard electroencephalogram (EEG). A normal EEG therefore does not exclude epilepsy. On the other hand, around 10% of patients who do not have epilepsy have epileptiform discharges on their EEG. This is referred to as an 'incidental finding'. Incidental findings are common in medicine, and are increasing in incidence with the greater availability of more sensitive tests. Test results should always be interpreted in the light of the patient's history and physical examination. Fig. 1.3 Normal distribution and reference range. For many tests, the frequency distribution of results in the normal healthy population (red line) is a symmetrical bell-shaped curve. The mean ± 2 standard deviations (SD) encompasses 95% of the normal population and usually defines the 'reference range'; 2.5% of the normal population have values above, and 2.5% below, this range (shaded areas). For some diseases (blue line), test results overlap with the normal population or even with the reference range. For other diseases (green line), tests may be more reliable because there is no overlap between the normal and abnormal population. Normal population Number of people having each value Abnormal populations Mean – 2SD Mean

- 2SD Mean 'Reference range' Value 1.3 Examples of factors other than disease that influence test results Factor Examples Age Creatinine is lower in old age (due to relatively lower muscle mass) – an older person can have a significantly reduced eGFR rate with a 'normal' creatinine Ethnicity Healthy people of African ancestry have lower white cell counts Pregnancy Several tests are affected by late pregnancy, due to the effects of a growing fetus, including: Reduced urea and creatinine (haemodilution) Iron deficiency anaemia (increased demand) Increased alkaline phosphatase (produced by the placenta) Raised D-dimer (physiological changes in the coagulation system) Mild respiratory alkalosis (physiological maternal hyperventilation) ECG changes (tachycardia, left axis deviation) Sex Males and females have different reference ranges for many tests, e.g. haemoglobin Spurious (in vitro) results A spurious high potassium is seen in haemolysis and in thrombocytosis ('pseudohyperkalaemia') (ECG = electrocardiogram; eGFR = estimated glomerular filtration rate, a better estimate of renal function than creatinine)

new information depends on what you believed beforehand. In other words, the interpretation of a test result depends on the probability of disease before the test. Prevalence of disease Consider this problem that was posed to a group of Harvard doctors: if a test to detect a disease whose prevalence is 1 : 1000 has a false-positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms and signs? Take a moment to work this out. In this problem, we have removed clinical probability and are only considering prevalence. The answer is at the end of the chapter. Predictive values combine sensitivity, specificity and prevalence. Sensitivity and specificity are characteristics of the test; the population does not change this. However, as doctors, we are interested in the question, 'What is the probability that a person with a positive test actually has the disease?' This is illustrated in Box 1.5. Post-test probability and predictive values are different. Posttest

probability is the probability of a disease after taking into account new information from a test result. Bayes' Theorem can be used to calculate post-test probability for a patient in any population. The pre-test probability of disease is decided by the doctor; it is a judgement based on information gathered prior to ordering the test. Predictive value is the proportion of patients with a test result who have the disease (or no disease) and is calculated from a table of results in a specific population (see Box 1.5). It is not possible to transfer this value to a different population. This is important to realise because published information about the performance of diagnostic tests may not apply to different populations. In deciding the pre-test probability of disease, clinicians often neglect to take prevalence into account and this distorts their estimate of probability. To estimate the probability of disease in a patient more accurately, clinicians should anchor on the prevalence of disease in the subgroup to which the patient belongs and then adjust to take the individual factors into account. the hip is extremely painful to move and she cannot stand. However, her hip X-rays are normal. Does she have a fracture? The sensitivity of plain X-rays of the hip performed in the emergency department for suspected hip fracture is around 95%. A small percentage of fractures are therefore missed. If our patient has (or is at risk of) osteoporosis, has severe pain on hip movement and cannot bear weight on the affected side, then the clinical probability of hip fracture is high. If, on the other hand, she is unlikely to have osteoporosis, has no pain on hip movement and is able to bear weight, then the clinical probability of hip fracture is low. Doctors are continually making judgements about whether something is true, given that something else is true. This is known as 'conditional probability'. Bayes' Theorem (named after English clergyman Thomas Bayes, 1702–1761) is a mathematical way to describe the post-test probability of a disease by combining pre-test probability, sensitivity and specificity. In clinical practice, doctors are not able to make complex mathematical calculations for every decision they make. In practical terms, the answer to the question of whether there is a fracture is that in a high-probability patient a normal test result does not exclude the condition, but in a low-probability patient it makes it very unlikely. This principle is illustrated in Figure 1.5. Sox and colleagues (see 'Further information') state a fundamental assertion, which they describe as a profound and subtle principle of clinical medicine: the interpretation of Fig. 1.4 Receiver operating characteristic graph illustrating the trade-off between sensitivity and specificity for a given test. The curve is generated by 'adjusting' the cut-off values defining normal and abnormal results, calculating the effect on sensitivity and specificity and then plotting these against each other. The closer the curve lies to the top left-hand corner, the more useful the test. The red line illustrates a test with useful discriminant value and the green line illustrates a less useful, poorly discriminant test. 1.0 0.8 0.6 0.4 0.2 0.01.0 0.8 0.6 Specificity 0.4 0.2

Sensitivity 1.4 Sensitivity and specificity Disease No disease Positive test A B (True positive) (False positive) Negative test C D (False negative) (True negative) Sensitivity = A/(A+C) × 100 Specificity = D/(D+B) × 100 1.5 Predictive values: 'What is the probability that a person with a positive test actually has the disease?' Disease No disease Positive test A B (True positive) (False positive) Negative test C D (False negative) (True negative) Positive predictive value = A/(A+B) × 100 Negative predictive value = D/(D+C) × 100 Dealing with uncertainty Clinical findings are imperfect and diagnostic tests are imperfect. It is important to recognise that clinicians frequently deal with uncertainty. By expressing uncertainty as probability, new information from diagnostic tests can be incorporated more accurately. However, subjective estimates of probability can sometimes be unreliable. As the section on cognitive biases will demonstrate (see below), intuition can be a source of error.

6 • CLINICAL DECISION-MAKING an understanding of the prevalence of disease in the particular care setting or the population to which the patient belongs. Cognitive biases Advances in cognitive psychology in recent decades have demonstrated that human thinking and decision-making are prone to error. Cognitive biases are subconscious errors that lead to inaccurate judgement and illogical interpretation of information. They are prevalent in everyday life; as the famous saying goes, ‘to err is human.’ Take a few moments to look at this simple puzzle. Do not try to solve it mathematically but listen to your intuition: A bat and ball cost £1.10. The bat costs £1 more than the ball. How much does the ball cost? The answer is at the end of the chapter. Most people get the answer to this puzzle wrong. Two things are going on: one is that humans have two distinct types of processes when it comes to thinking and decision-making – termed ‘type 1’ and ‘type 2’ thinking. The other is that the human brain is wired to jump to conclusions sometimes or to miss things that are obvious. British psychologist and patient safety pioneer James Knowing the patient’s true state is often unnecessary in clinical decision-making. Sox and colleagues (see ‘Further information’) argue that there is a difference between knowing that a disease is present and acting as if it were present. The requirement for diagnostic certainty depends on the penalty for being wrong. Different situations require different levels of certainty before starting treatment. How we communicate uncertainty to patients will be discussed later in this chapter (p. 10). The treatment threshold combines factors such as the risks of the test, and the risks versus benefits of treatment. The point at which the factors are all evenly weighed is the threshold. If a test or treatment for a disease is effective and low-risk (e.g. giving antibiotics for a suspected urinary tract infection), then there is a lower threshold for going ahead. On the other hand, if a test or treatment is less effective or high-risk (e.g. starting chemotherapy for a malignant brain tumour), then greater confidence is required in the clinical diagnosis and potential benefits of treatment first. In principle, if a diagnostic test will not change the management of the patient, then careful consideration should be given to whether it is necessary to do the test at all. In summary, test results shift our thinking, but rarely give a ‘yes’ or a ‘no’ answer in terms of a diagnosis. Sometimes tests shift the probability of disease by less than we realise. Pre-test probability is key, and this is derived from the history and physical examination, combined with a sound knowledge of medicine and Fig. 1.5 The interpretation of a test result depends on the probability of the disease before the test is carried out. In the example shown, the test being carried out has a sensitivity of 95% and a specificity of 85%. Patient A has very characteristic clinical findings, which make the pre-test probability of the condition for which the test is being used very high – estimated as 90%. Patient B has more equivocal findings, such that the pre-test probability is estimated as only 50%. If the result in Patient A is negative, there is still a significant chance that he has the condition for which he is being tested; in Patient B, however, a negative result makes the diagnosis very unlikely. Patient A 90% chance of having the disease before the test is done 34.6% chance of having the disease if the test is negative

% probability of having the disease

98.3% chance of having the disease if the test is positive Patient B 50% chance of having the disease before the test is done 86.4% chance of having the disease if the test is positive 5.6% chance of having the disease if the test is negative

Cognitive biases • 7

was found beside her at home. Her observations show she has a Glasgow Coma Scale score of 10/15, heart rate 100 beats/ min, blood pressure 100/60 mmHg, respiratory rate 14 breaths/ min, oxygen saturations 98% on air and temperature 37.5°C. Already your mind has reached a working diagnosis. It fits a pattern (type 1 thinking). You think she has taken an overdose. At this point you can stop to think about your thinking (rational override in Fig. 1.6): 'What is the evidence for this diagnosis? What else could it be?' On the other hand, imagine being asked to assess a patient who has been admitted with syncope. There are several different causes of syncope and a systematic approach is required to reach a diagnosis (type 2 thinking). However, you recently heard about a case of syncope due to a leaking abdominal aortic aneurysm. At the end of your assessment, following evidence-based guidelines, it is clear the patient can be discharged. Despite this, you decide to observe the patient overnight 'just in case' (irrational override in Fig. 1.6). In this example, your intuition is actually availability bias (when things are at the forefront of your mind), which has significantly distorted your estimate of probability. Common cognitive biases in medicine Figure 1.7 illustrates the common cognitive biases prevalent in medical practice. Biases often work together; for example, in Reason said that, 'Our propensity for certain types of error is the price we pay for the brain's remarkable ability to think and act intuitively – to sift quickly through the sensory information that constantly bombards us without wasting time trying to work through every situation anew.' This property of human thinking is highly relevant to clinical decision-making. Type 1 and type 2 thinking Studies of cognitive psychology and functional magnetic resonance imaging demonstrate two distinct types of processes when it comes to decision-making: intuitive (type 1) and analytical (type 2). This has been termed 'dual process theory'. Box 1.6 explains this in more detail. Psychologists estimate that we spend 95% of our daily lives engaged in type 1 thinking – the intuitive, fast, subconscious mode of decision-making. Imagine driving a car, for example; it would be impossible to function efficiently if every decision and movement were as deliberate, conscious, slow and effortful as in our first driving lesson. With experience, complex procedures become automatic, fast and effortless. The same applies to medical practice. There is evidence that expert decision-making is well served by intuitive thinking. The problem is that although intuitive processing is highly efficient in many circumstances, in others it is prone to error. Clinicians use both type 1 and type 2 thinking, and both types are important in clinical decision-making. When encountering a problem that is familiar, clinicians employ pattern recognition and reach a working diagnosis or differential diagnosis quickly (type 1 thinking). When encountering a problem that is more complicated, they use a slower, systematic approach (type 2 thinking). Both types of thinking interplay – they are not mutually exclusive in the diagnostic process. Figure 1.6 illustrates the interplay between type 1 and type 2 thinking in clinical practice. Errors can occur in both type 1 and type 2 thinking; for example, people can apply the wrong rules or make errors in their application while using type 2 thinking. However, it has been argued that the common cognitive biases encountered in medicine tend to occur when clinicians are engaged in type 1 thinking. For example, imagine being asked to see a young woman who is drowsy. She is handed over to you as a 'probable overdose' because she has a history of depression and a packet of painkillers Fig. 1.6 The interplay between type 1 and type 2 thinking in the diagnostic process. Adapted from Croskerry P. A universal model of diagnostic reasoning. Acad Med 2009; 84:1022–1028. Experience Context Ambient conditions Education Training Logical competence Clinical presentation Recognised Not recognised Type 2 processes Type 1 processes Cognitive biases more likely Irrational override Rational override Working diagnosis 1.6 Type 1 and type 2 thinking Type 1 Type 2 Intuitive, heuristic (pattern recognition) Analytical, systematic Automatic, subconscious Deliberate, conscious Fast, effortless Slow, effortful Low/variable reliability

High/consistent reliability Vulnerable to error Less prone to error Highly affected by context Less affected by context High emotional involvement Low emotional involvement Low scientific rigour High scientific rigour

8 • CLINICAL DECISION-MAKING Fig. 1.7 Common cognitive biases in medicine. Adapted from Croskerry P. Achieving quality in clinical decision-making: cognitive strategies and detection of bias. Acad Emerg Med 2002; 9:1184–1204. Anchoring The common human tendency to rely too heavily on the first piece of information offered (the 'anchor') when making decisions Diagnostic momentum Once a diagnostic label has been attached to a patient (by the patient or other health-care professionals), it can gather momentum with each review, leading others to exclude other possibilities in their thinking Premature closure The tendency to close the decisionmaking process prematurely and accept a diagnosis before it, and other possibilities, have been fully explored Ascertainment bias We sometimes see what we expect to see ('self-fulfilling prophecy'). For example, a frequent self-harmer attends the emergency department with drowsiness; everyone assumes he has taken another overdose and misses a brain injury Psych-out error Psychiatric patients who present with medical problems are underassessed, under-examined and under-investigated because problems are presumed to be due to, or exacerbated by, their psychiatric condition Framing effect How a case is presented – for example, in handover – can generate bias in the listener. This can be mitigated by always having 'healthy scepticism' about other people's diagnoses Availability bias Things may be at the forefront of your mind because you have seen several cases recently or have been studying that condition in particular. For example, when one of the authors worked in an epilepsy clinic, all blackouts were possible seizures Hindsight bias Knowing the outcome may profoundly influence the perception of past events and decision-making, preventing a realistic appraisal of what actually occurred – a major problem in learning from diagnostic error Search satisficing We may stop searching because we have found something that fits or is convenient, instead of systematically looking for the best alternative, which involves more effort Base rate neglect The tendency to ignore the prevalence of a disease, which then distorts Bayesian reasoning. In some cases, clinicians do this deliberately in order to rule out an unlikely but worst-case scenario Omission bias The tendency towards inaction, rooted in the principle of 'first do no harm.' Events that occur through natural progression of disease are more acceptable than those that may be attributed directly to the action of the health-care team Triage-cueing Triage ensures patients are sent to the right department. However, this leads to 'geography is destiny'. For example, a diabetic ketoacidosis patient with abdominal pain and vomiting is sent to surgery. The wrong location (surgical ward) stops people thinking about medical causes of abdominal pain and vomiting Commission bias The tendency towards action rather than inaction, on the assumption that good can come only from doing something (rather than 'watching and waiting') Overconfidence bias The tendency to believe we know more than we actually do, placing too much faith in opinion instead of gathered evidence Unpacking principle Failure to 'unpack' all the available information may mean things are missed. For example, if a thorough history is not obtained from either the patient or carers (a common problem in geriatric medicine), diagnostic possibilities may be discounted Confirmation bias The tendency to look for confirming evidence to support a theory rather than looking for disconfirming evidence to refute it, even if the latter is clearly present. Confirmation bias is common when a patient has been seen first by another doctor Posterior probability Our estimate of the likelihood of disease may be unduly influenced by what has gone on before for a particular patient. For example, a patient who has been extensively investigated for headaches presents with a severe headache, and serious causes are discounted

Visceral bias The influence of either negative or positive feelings towards patients, which can affect our decisionmaking

Reducing errors in clinical decision-making • 9

• adopting ‘cognitive debiasing strategies’ • using clinical prediction rules and other decision aids • engaging in effective team communication. Cognitive debiasing strategies There are some simple and established techniques that can be used to avoid cognitive biases and errors in clinical decision-making. History and physical examination Taking a history and performing a physical examination may seem obvious, but these are sometimes carried out inadequately. This is the ‘unpacking principle’: failure to unpack all the available information means things can be missed and lead to error. Problem lists and differential diagnosis Once all the available data from history, physical examination and (sometimes) initial test results are available, these need to be synthesised into a problem list. The ability to identify key clinical data and create a problem list is a key step in clinical reasoning. Some problems (e.g. low serum potassium) require action but not necessarily a differential diagnosis. Other problems (e.g. vomiting) require a differential diagnosis. The process of generating a problem list ensures nothing is missed. The process of generating a differential diagnosis works against anchoring on a particular diagnosis too early, thereby avoiding search satisficing and premature closure (see Fig. 1.7). Mnemonics and checklists These are used frequently in medicine in order to reduce reliance on fallible human memory. ABCDE (airway, breathing, circulation, disability, exposure/examination) is probably the most successful checklist in medicine, used during the assessment and treatment of critically ill patients (ABCDE is sometimes prefixed with ‘C’ for ‘control of any obvious problem’; see p. 188). Checklists ensure that important issues have been considered and completed, especially under conditions of complexity, stress or fatigue. Red flags and ROWS (‘rule out worst case scenario’) These are strategies that force doctors to consider serious diseases that can present with common symptoms. Red flags in back pain are listed in Box 24.19 (p. 996). Considering and investigating for possible pulmonary embolism in patients who overconfidence bias (the tendency to believe we know more than we actually do), too much faith is placed in opinion instead of gathered evidence. This bias can be augmented by the availability bias and finally by commission bias (the tendency towards action rather than inaction) – sometimes with disastrous results. The mark of a well-calibrated thinker is the ability to recognise what mode of thinking is being employed and to anticipate and recognise situations in which cognitive biases and errors are more likely to occur. Human factors ‘Human factors’ is the science of the limitations of human performance, and how technology, the work environment and team communication can adapt for this to reduce diagnostic and other types of error. Analysis of serious adverse events in clinical practice shows that human factors and poor team communication play a significant role when things go wrong. Research shows that many errors are beyond an individual’s conscious control and are precipitated by many factors. The discipline of human factors seeks to understand interactions between: • people and tasks or technology • people and their work environment • people in a team. An understanding of these interactions makes it easier for health-care professionals, who are committed to ‘first do no harm,’ to work in the safest way possible. For example, performance is adversely affected by factors such as poorly designed processes and equipment, frequent interruptions and fatigue. The areas of the brain required for type 2 processing are most affected by things like fatigue and cognitive overload, and the brain reverts to type 1 processing to conserve cognitive energy. Figure 1.8 illustrates some of the internal and external factors that affect human judgement and decision-making. Various experiments

demonstrate that we focus our attention to filter out distractions. This is advantageous in many situations, but in focusing on what we are trying to see we may not notice the unexpected. In a team context, what is obvious to one person may be completely missed by someone else. Safe and effective team communication therefore requires us never to assume, and to verbalise things, even though they may seem obvious. Reducing errors in clinical decision-making Knowledge and experience do not eliminate errors. Instead, there are a number of ways in which we can act to reduce errors in clinical decision-making. Examples are: Fig. 1.8 Factors that affect our judgement and decision-making. Type 1 thinking = fast, intuitive, subconscious, low-effort. Error Type 1 thinking/ conservation of cognitive effort Cognitive and affective biases Internal factors Knowledge Training Beliefs and values Emotions Sleep/fatigue Stress Physical illness Personality type External factors Interruptions Cognitive overload Time pressure Ambient conditions Insufficient data Team factors Patient factors Poor feedback

 with dual antiplatelet therapy and low-molecular-weight heparin as recommended in clinical guidelines? As this chapter has described, clinicians frequently deal with uncertainty/probability. Clinicians need to be able to explain risks and benefits of treatment in an accurate and understandable way. Providing the relevant statistics is seldom sufficient to guide decision-making because a patient's perception of risk may be influenced by irrational factors as well as individual values. Research evidence provides statistics but these can be confusing. Terms such as 'common' and 'rare' are nebulous. Whenever possible, clinicians should quote numerical information using consistent denominators (e.g. '90 out of 100 patients who have this operation feel much better, 1 will die during the operation and 2 will suffer a stroke'). Visual aids can be used to present complex statistical information (Fig. 1.9). How uncertainty is conveyed to patients is important. Many studies demonstrate a correlation between effective clinician– patient communication and improved health outcomes. If patients feel they have been listened to and understand the problem and proposed treatment plan, they are more likely to follow the plan and less likely to re-attend. Clinical decision-making: putting it all together The following is a practical example that brings together many of the concepts outlined in this chapter: A 25-year-old woman presents with right-sided pleuritic chest pain and breathlessness. She reports that she had an upper present with pleuritic chest pain and breathlessness is a common example of ruling out a worst-case scenario, as pulmonary embolism can be fatal if missed. Red flags and ROWS help to avoid cognitive biases such as the 'framing effect' and 'premature closure'. Newer strategies to avoid cognitive biases and errors in decisionmaking are emerging. These involve explicit training in clinical reasoning and human factors. In theory, if doctors are aware of the science of human thinking and decision-making, then they are more able to think about their thinking, understand situations in which their decision-making may be affected, and take steps to mitigate this. Using clinical prediction rules and other decision aids A clinical prediction rule is a statistical model of the diagnostic process. When clinical prediction rules are matched against the opinion of experts, the model usually outperforms the experts, because it is applied consistently in each case. However, it is important that clinical prediction rules are used correctly – that is, applied to the patient population that was used to create the rule. Clinical prediction rules force a scientific assessment of the patient's symptoms, signs and other data to develop a numerical probability of a disease or an outcome. They help clinicians to estimate probability more accurately. A good example of a clinical prediction rule to estimate pre-test probability is the Wells score in suspected deep vein thrombosis (see Box 10.15, p. 187). Other commonly used clinical prediction rules predict outcomes and therefore guide the management plan. These include the GRACE score in acute coronary

syndromes (see Fig. 16.62, p. 494) and the CURB-65 score in communityacquired pneumonia (see Fig. 17.32, p. 583). Effective team communication Effective team communication and proper handovers are vital for safe clinical care. The SBAR system of communication has been recommended by the UK's Patient Safety First campaign. It is a structured way to communicate about a patient with another health-care professional (e.g. during handover or when making a referral) and increases the amount of relevant information being communicated in a shorter time. It is illustrated in Box 1.7. In increasingly complex health-care systems, patients are looked after by a wide variety of professionals, each of whom has access to important information required to make clinical decisions. Strict hierarchies are hazardous to patient safety if certain members of the team are not able to speak up. Patient-centred evidence-based medicine and shared decision-making 'Patient-centred evidence-based medicine' refers to the application of best-available research evidence while taking individual patient factors into account; these include both clinical and non-clinical factors (e.g. the patient's social circumstances, values and wishes). For example, a 95-year-old man with dementia and a recent gastrointestinal bleed is admitted with an inferior myocardial infarction. He is clinically well. Should he be treated From Royal College of Physicians of London. National Early Warning Score: standardising the assessment of illness severity in the NHS. Report of a working party. RCP, July 2012; www.rcplondon.ac.uk/projects/outputs/national-earlywarning-score-news (accessed March 2016). 1.7 The SBAR system of communicating SBAR Example (a telephone call to the Intensive Care team) Situation I am [name] calling from [place] about a patient with a NEWS of 10. Background [Patient's name], 30-year-old woman, no past medical history, was admitted last night with community-acquired pneumonia. Since then her oxygen requirements have been steadily increasing. Assessment Her vital signs are: blood pressure 115/60 mmHg, heart rate 120 beats/min, temperature 38°C, respiratory rate 32 breaths/min, oxygen saturations 89% on 15 L via reservoir bag mask. An arterial blood gas shows pH 7.3 (H+ 50 nmol/L), PaCO2 4.0 kPa (30 mmHg), PaO2 7 kPa (52.5 mmHg), standard bicarbonate 14 mmol/L. Chest X-ray shows extensive right lower zone consolidation. Recommendation Please can you come and see her as soon as possible? I think she needs admission to Intensive Care. (NEWS = National Early Warning Score; a patient with normal vital signs scores 0)

< 500 ng/mL). A normal chest X-ray is a common finding in pulmonary embolism. Several studies have shown that the D-dimer assay has at least 95% sensitivity in acute pulmonary embolism but it has a low specificity. A very sensitive test will detect most disease but generate abnormal findings in healthy people. On the other hand, a negative result virtually, but not completely, excludes the disease. It is important at this point to realise that a raised D-dimer result does not mean this patient has a pulmonary embolism; it just means that we have not been able to exclude it. Since pulmonary embolism is a potentially fatal condition we need to rule out the worst-case scenario (ROWS), and the next step is therefore to arrange further imaging. What kind of imaging depends on individual patient characteristics and what is available. Treatment threshold The treatment threshold combines factors such as the risks of the test, and the risks versus benefits of treatment. A CT pulmonary angiogram (CTPA) could be requested for this patient, although in some circumstances ventilation–perfusion single-photon emission computed tomography (V□ /Q□ SPECT, p. 620) may be a more suitable alternative. However, what if the scan cannot be performed until the next day? Because pulmonary embolism is potentially fatal and the risks of treatment in this case are low, the patient should be started on treatment while awaiting the scan. Post-test

probability The patient’s scan result is subsequently reported as ‘no pulmonary embolism’. Combined with the low pre-test probability, this scan result reliably excludes pulmonary embolism. Cognitive biases Imagine during this case that the patient had been handed over to you as ‘nothing wrong – probably a pulled muscle’. Cognitive biases (subconscious tendencies to respond in a certain way) would come into play, such as the ‘framing effect’, ‘confirmation bias’ and ‘search satisficing’. The normal clinical examination might confirm the diagnosis of musculoskeletal pain in your mind, despite the examination being entirely consistent with pulmonary embolism and despite the lack of history and examination findings (e.g. chest wall tenderness) to support the diagnosis of musculoskeletal chest pain. Human factors Imagine that, after you have seen the patient, a nurse hands you some blood forms and asks you what tests you would like to request on ‘this lady’. You request blood tests including a D-dimer on the wrong patient. Luckily, this error is intercepted. Reducing cognitive error The diagnosis of pulmonary embolism can be difficult. Clinical prediction rules (e.g. modified Wells score), guidelines (e.g. from the UK’s National Institute for Health and Care Excellence, or NICE) and decision aids (e.g. simplified pulmonary embolism severity index, or PESI) are frequently used in combination with the doctor’s opinion, derived from information gathered in the history and physical examination. respiratory tract infection a week ago and was almost back to normal when the symptoms started. The patient has no past medical history and no family history, and her only medication is the combined oral contraceptive pill. On examination, her vital signs are normal (respiratory rate 19 breaths/min, oxygen saturations 98% on air, blood pressure 115/60 mmHg, heart rate 90 beats/min, temperature 37.5°C) and the physical examination is also normal. You have been asked to assess her for the possibility of a pulmonary embolism. (More information on pulmonary embolism can be found on page 619.) Evidence-based history and examination Information from the history and physical examination is vital in deciding whether this could be a pulmonary embolism. Pleurisy and breathlessness are common presenting features of this disease but are also common presenting features in other diseases. There is nothing in the history to suggest an alternative diagnosis (e.g. high fever, productive cough, recent chest trauma). The patient’s vital signs are normal, as is the physical examination. However, the only feature in the history and examination that has a negative likelihood ratio in the diagnosis of pulmonary embolism is a heart rate of less than 90 beats/min. In other words, the normal physical examination findings (including normal oxygen saturations) carry very little diagnostic weight. Deciding pre-test probability The prevalence of pulmonary embolism in 25-year-old women is low. We anchor on this prevalence and then adjust for individual patient factors. This patient has no major risk factors for pulmonary embolism. To assist our estimate of pre-test probability, we could use a clinical prediction rule: in this case, the modified Wells score for pulmonary embolism, which would give a score of 3 (low probability – answering yes only to the criterion ‘PE is the number one diagnosis, an alternative is less likely’). Interpreting test results Imagine the patient went on to have a normal chest X-ray and blood results, apart from a raised D-dimer of 900 (normal Fig. 1.9 Visual portrayal of benefits and risks. The image refers to an operation that is expected to relieve symptoms in 90% of patients, but cause stroke in 2% and death in 1%. From Edwards A, Elwyn G, Mulley A. Explaining risks: turning numerical data into meaningful pictures. BMJ 2002; 324:827–830, reproduced with permission from the BMJ Publishing Group. Feel better No difference Stroke Dead

 The distinctive mark of this easy puzzle is that it evokes an answer that is intuitive, appealing – and wrong. Do the math, and you will see.’ The correct answer is 5p. Further information Books and journal articles Cooper N, Frain J (eds). ABC of clinical

reasoning. Oxford: Wiley–Blackwell; 2016. Kahneman D. Thinking, fast and slow. Harmondsworth: Penguin; 2012. McGee S. Evidence-based physical diagnosis, 3rd edn. Philadelphia: Saunders; 2012. Scott IA. Errors in clinical reasoning: causes and remedial strategies. BMJ 2009; 338:b186. Sox H, Higgins MC, Owens DK. Medical decision making, 2nd edn. Chichester: Wiley–Blackwell; 2013. Trowbridge RL, Rencic JJ, Durning SJ. Teaching clinical reasoning. Philadelphia: American College of Physicians; 2015. Vincent C. Patient safety. Edinburgh: Churchill Livingstone; 2006.
Websites chfg.org UK Clinical Human Factors Group. clinical-reasoning.org Clinical reasoning resources. creme.org.uk UK Clinical Reasoning in Medical Education group. improvediagnosis.org Society to Improve Diagnosis in Medicine. vassarstats.net/index.html Suite of calculators for statistical computation (Calculator 2 is a calculator for predictive values and likelihood ratios). Person-centred EBM and information given to patient The patient is treated according to evidence-based guidelines that apply to her particular situation. Tests alone do not make a diagnosis and at the end of this process the patient is told that the combination of history, examination and test results mean she is extremely unlikely to have a pulmonary embolism. Viral pleurisy is offered as an alternative diagnosis and she is reassured that her symptoms are expected to settle over the coming days with analgesia. She is advised to re-present to hospital if her symptoms suddenly get worse. Answers to problems Harvard problem (p. 5) Almost half of doctors surveyed said 95%, but they neglected to take prevalence into account. If 1000 people are tested, there will be 51 positive results: 50 false positives and 1 true positive. The chance that a person found to have a positive result actually has the disease is 1/51 or 2%. Bat and ball problem (p. 6) This puzzle is from the book, Thinking, Fast and Slow, by Nobel laureate Daniel Kahneman (see ‘Further information’). He writes, ‘A number came to your mind. The number, of course, is 10p.

02-2 Clinical therapeutics and good prescribing

2 Clinical therapeutics and good prescribing

 strength of the chemical bond. Some drug–receptor interactions are irreversible, either because the affinity is so strong or because the drug modifies the structure of its molecular target. • Selectivity describes the propensity for a drug to bind to one target rather than another. Selectivity is a relative term, not to be confused with absolute specificity. It is common for drugs targeted at a particular subtype of receptor to exhibit some effect at other subtypes. For example, β-adrenoceptors can be subtyped on the basis of their responsiveness to the endogenous agonist noradrenaline (norepinephrine): the concentration of noradrenaline required to cause bronchodilatation (via β2-adrenoceptors) is ten times higher than that required to cause tachycardia (via β1-adrenoceptors). ‘Cardioselective’ β-blockers have anti-anginal effects on the heart (β1) but may still cause bronchospasm in the lung (β2) and are contraindicated for asthmatic patients. • Agonists bind to a receptor to produce a conformational change that is coupled to a biological response. As agonist concentration increases, so does the proportion of receptors occupied, and hence the biological effect. Partial agonists activate the receptor but cannot produce a maximal signalling effect equivalent to that of a full agonist, even when all available receptors are occupied. • Antagonists bind to a receptor but do not produce the conformational change that initiates an intracellular signal. A competitive antagonist competes with endogenous ligands to occupy receptor-binding sites, with the resulting antagonism depending on the relative affinities and concentrations of drug and ligand. Non-competitive antagonists inhibit the effect of an agonist by mechanisms other than direct competition for receptor binding with the agonist (e.g. by affecting post-receptor signalling). Dose–response relationships Plotting the logarithm of drug dose against drug response typically produces a sigmoidal dose–response curve (Fig. 2.2). Progressive increases in drug dose (which, for most

drugs, is proportional to the plasma drug concentration) produce increasing Prescribing medicines is the major tool used by doctors to restore or preserve the health of patients. Medicines contain drugs (the specific chemical substances with pharmacological effects), either alone or in combination with additional drugs, in a formulation mixed with other ingredients. The beneficial effects of medicines must be weighed against their cost and potential adverse drug reactions and interactions. The latter two factors are sometimes caused by injudicious prescribing decisions and by prescribing errors. The modern prescriber must meet the challenges posed by the increasing number of drugs and formulations available and of indications for prescribing them, and the greater complexity of treatment regimens followed by individual patients ('polypharmacy', a particular challenge in the ageing population). The purpose of this chapter is to elaborate on the principles and practice that underpin good prescribing (Box 2.1). Fig. 2.1 Pharmacokinetics and pharmacodynamics. Dosage regimen Plasma concentration Concentration at the site of action Pharmacological effects Pharmacokinetics 'what the body does to a drug' Monitoring Measure plasma drug concentration 'what a drug does to the body' Monitoring Measure clinical effects Time Concentration Pharmacodynamics Concentration Effect *These steps in particular take the patient's views into consideration to establish a therapeutic partnership (shared decision-making to achieve 'concordance'). 2.1 Steps in good prescribing • Make a diagnosis • Consider factors that might influence the patient's response to therapy (age, concomitant drug therapy, renal and liver function etc.) • Establish the therapeutic goal* • Choose the therapeutic approach* • Choose the drug and its formulation (the 'medicine') • Choose the dose, route and frequency • Choose the duration of therapy • Write an unambiguous prescription (or 'medication order') • Inform the patient about the treatment and its likely effects • Monitor treatment effects, both beneficial and harmful • Review/alter the prescription Principles of clinical pharmacology Prescribers need to understand what the drug does to the body (pharmacodynamics) and what the body does to the drug (pharmacokinetics) (Fig. 2.1). Although this chapter is focused on the most common drugs, which are synthetic small molecules, the same principles apply to the increasingly numerous 'biological' therapies (sometimes abbreviated to 'biologics') now in use, which include peptides, proteins, enzymes and monoclonal antibodies (see Box 4.2, p. 65). Pharmacodynamics Drug targets and mechanisms of action Modern drugs are usually discovered by screening compounds for activity either to stimulate or to block the function of a specific molecular target, which is predicted to have a beneficial effect in a particular disease (Box 2.2). Other drugs have useful but less selective chemical properties, such as chelators (e.g. for treatment of iron or copper overload), osmotic agents (used as diuretics in cerebral oedema) or general anaesthetics (that alter the biophysical properties of lipid membranes). The following characteristics of the interaction of drugs with receptors illustrate some of the important determinants of the effects of drugs: • Affinity describes the propensity for a drug to bind to a receptor and is related to the 'molecular fit' and the

Fig. 2.2 Dose–response curve. The green curve represents the beneficial effect of the drug. The maximum response on the curve is the Emax and the dose (or concentration) producing half this value (Emax/2) is the ED50 (or EC50). The red curve illustrates the dose–response relationship for the most important adverse effect of this drug. This occurs at much higher doses; the ratio between the ED50 for the adverse effect and that for the beneficial effect is the 'therapeutic index', which indicates how much margin there is for prescribers when choosing a dose that will provide beneficial effects without also causing this adverse effect. Adverse effects that occur at doses

above the therapeutic range are normally called ‘toxic effects’, while those occurring within the therapeutic range are ‘side-effects’ and those below it are ‘hyper-susceptibility effects’. Hypersusceptibility Side-effects

0.0001 0.001 0.01 0.1

Therapeutic index 100/0.1 = 1000 Drug dose (mg) Response (% of maximum) Toxic effects Adverse effect ED50 =100 mg Beneficial effect ED50 =0.1 mg Emax ED50 ED50 2.2 Examples of target molecules for drugs Drug target Description Examples Receptors Channel-linked receptors Ligand binding controls a linked ion channel, known as ‘ligand-gated’ (in contrast to ‘voltage-gated’ channels that respond to changes in membrane potential) Nicotinic acetylcholine receptor GABA receptor Sulphonylurea receptor G-protein-coupled receptors (GPCRs) Ligand binding affects one of a family of ‘G-proteins’ that mediate signal transduction either by activating intracellular enzymes (such as adenylate or guanylate cyclase, producing cyclic AMP or GMP, respectively) or by controlling ion channels Muscarinic acetylcholine receptor β-adrenoceptors Dopamine receptors 5-Hydroxytryptamine (5-HT, serotonin) receptors Opioid receptors Kinase-linked receptors Ligand binding activates an intracellular protein kinase that triggers a cascade of phosphorylation reactions Insulin receptor Cytokine receptors Transcription factor receptors Intracellular and also known as ‘nuclear receptors’; ligand binding promotes or inhibits gene transcription and hence synthesis of new proteins Steroid receptors Thyroid hormone receptors Vitamin D receptors Retinoid receptors PPARγ and α receptors Other targets Voltage-gated ion channels Mediate electrical signalling in excitable tissues (muscle and nervous system) Na+ channels Ca2+ channels Enzymes Catalyse biochemical reactions. Drugs interfere with binding of substrate to the active site or of co-factors Cyclo-oxygenase ACE Xanthine oxidase Transporter proteins Carry ions or molecules across cell membranes 5-HT re-uptake transporter Na+/K+ ATPase Cytokines and other signalling molecules Small proteins that are important in cell signalling (autocrine, paracrine and endocrine), especially affecting the immune response Tumour necrosis factors Interleukins Cell surface antigens Block the recognition of cell surface molecules that modulate cellular responses Cluster of differentiation molecules (e.g. CD20, CD80) (ACE = angiotensin-converting enzyme; AMP = adenosine monophosphate; ATPase = adenosine triphosphatase; GABA = γ-aminobutyric acid; GMP = guanosine monophosphate; PPAR = peroxisome proliferator-activated receptor)

16 • CLINICAL THERAPEUTICS AND GOOD PRESCRIBING therapeutic index is usually based on adverse effects that might require dose reduction or discontinuation. For most drugs, the therapeutic index is greater than 100 but there are some notable exceptions with therapeutic indices of less than 10 (e.g. digoxin, warfarin, insulin, phenytoin, opioids). The doses of such drugs have to be titrated carefully for individual patients to maximise benefits but avoid adverse effects. Desensitisation and withdrawal effects Desensitisation refers to the common situation in which the biological response to a drug diminishes when it is given continuously or repeatedly. It may be possible to restore the response by increasing the dose of the drug but, in some cases, the tissues may become completely refractory to its effect. • Tachyphylaxis describes desensitisation that occurs very rapidly, sometimes with the initial dose. This rapid loss of response implies depletion of chemicals that may be necessary for the pharmacological actions of the drug (e.g. a stored neurotransmitter released from a nerve terminal) or receptor phosphorylation. • Tolerance describes a more gradual loss of response to a drug that occurs over days or weeks. This slower change implies changes in receptor numbers or the development of counter-regulatory

physiological changes that offset the actions of the drug (e.g. accumulation of salt and water in response to vasodilator therapy). • Drug resistance is a term normally reserved for describing the loss of effectiveness of an antimicrobial (p. 116) or cancer chemotherapy drug. • In addition to these pharmacodynamic causes of desensitisation, reduced response may be the consequence of lower plasma and tissue drug concentrations as a result of altered pharmacokinetics (see below). When drugs induce chemical, hormonal and physiological changes that offset their actions, discontinuation may allow these changes to cause ‘rebound’ withdrawal effects (Box 2.3). response but only within a relatively narrow range of dose; further increases in dose beyond this range produce little extra effect. The following characteristics of the drug response are useful in comparing different drugs: • Efficacy describes the extent to which a drug can produce a target-specific response when all available receptors or binding sites are occupied (i.e. Emax on the dose–response curve). A full agonist can produce the maximum response of which the receptor is capable, while a partial agonist at the same receptor will have lower efficacy. Therapeutic efficacy describes the effect of the drug on a desired biological endpoint and can be used to compare drugs that act via different pharmacological mechanisms (e.g. loop diuretics induce a greater diuresis than thiazide diuretics and therefore have greater therapeutic efficacy). • Potency describes the amount of drug required for a given response. More potent drugs produce biological effects at lower doses, so they have a lower ED50. A less potent drug can still have an equivalent efficacy if it is given in higher doses. The dose–response relationship varies between patients because of variations in the many determinants of pharmacokinetics and pharmacodynamics. In clinical practice, the prescriber is unable to construct a dose–response curve for each individual patient. Therefore, most drugs are licensed for use within a recommended range of doses that is expected to reach close to the top of the dose–response curve for most patients. However, it is sometimes possible to achieve the desired therapeutic efficacy at doses towards the lower end of, or even below, the recommended range. Therapeutic index The adverse effects of drugs are often dose-related in a similar way to the beneficial effects, although the dose–response curve for these adverse effects is normally shifted to the right (Fig. 2.2). The ratio of the ED50 for therapeutic efficacy and for a major adverse effect is known as the ‘therapeutic index’. In reality, drugs have multiple potential adverse effects, but the concept of 2.3 Examples of drugs associated with withdrawal effects Drug Symptoms Signs Treatment Alcohol Anxiety, panic, paranoid delusions, visual and auditory hallucinations Agitation, restlessness, delirium, tremor, tachycardia, ataxia, disorientation, seizures Treat immediate withdrawal syndrome with benzodiazepines Barbiturates, benzodiazepines Similar to alcohol Similar to alcohol Transfer to long-acting benzodiazepine then gradually reduce dosage Glucocorticoids Weakness, fatigue, decreased appetite, weight loss, nausea, vomiting, diarrhoea, abdominal pain Hypotension, hypoglycaemia Prolonged therapy suppresses the hypothalamic–pituitary–adrenal axis and causes adrenal insufficiency requiring glucocorticoid replacement. Withdrawal should be gradual after prolonged therapy (p. 670) Opioids Rhinorrhoea, sneezing, yawning, lacrimation, abdominal and leg cramping, nausea, vomiting, diarrhoea Dilated pupils Transfer addicts to long-acting agonist methadone Selective serotonin re-uptake inhibitors (SSRIs) Dizziness, sweating, nausea, insomnia, tremor, delirium, nightmares Tremor Reduce SSRIs slowly to avoid withdrawal effects

Parenteral administration These routes avoid absorption via the gastrointestinal tract and first-pass metabolism in the liver: • Intravenous (IV). The IV route enables all of a dose to enter the systemic

circulation reliably, without any concerns about absorption or first-pass metabolism (i.e. the dose is 100% bioavailable), and rapidly achieve a high plasma concentration. It is ideal for very ill patients when a rapid, certain effect is critical to outcome (e.g. benzathine benzylpenicillin for meningococcal meningitis). • Intramuscular (IM). IM administration is easier to achieve than the IV route (e.g. adrenaline (epinephrine) for acute anaphylaxis) but absorption is less predictable and depends on muscle blood flow. • Subcutaneous (SC). The SC route is ideal for drugs that have to be administered parenterally because of low oral bioavailability, are absorbed well from subcutaneous fat, and might ideally be injected by patients themselves (e.g. insulin, heparin). • Transdermal. A transdermal patch can enable a drug to be absorbed through the skin and into the circulation (e.g. oestrogens, nicotine, nitrates). Other routes of administration • Topical application of a drug involves direct administration to the site of action (e.g. skin, eye, ear). This has the advantage of achieving sufficient concentration at this site while minimising systemic exposure and the risk of adverse effects elsewhere. • Inhaled (INH) administration allows drugs to be delivered directly to a target in the respiratory tree, usually the small airways (e.g. salbutamol, beclometasone). However, a significant proportion of the inhaled dose may be absorbed from the lung or is swallowed and can reach the systemic circulation. The most common mode of delivery is the metered-dose inhaler but its success depends on some degree of manual dexterity and timing (see Fig. 17.23, p. 571). Patients who find these difficult may use a ‘spacer’ device to improve drug delivery. A special mode Pharmacokinetics Understanding ‘what the body does to the drug’ (Fig. 2.3) is extremely important for prescribers because this forms the basis on which the optimal route of administration and dose regimen are chosen and explains the majority of inter-individual variation in the response to drug therapy. Drug absorption and routes of administration Absorption is the process by which drug molecules gain access to the blood stream. The rate and extent of drug absorption depend on the route of administration (Fig. 2.3). Enteral administration These routes involve administration via the gastrointestinal tract: • Oral. This is the most common route of administration because it is simple, convenient and readily used by patients to self-administer their medicines. Absorption after an oral dose is a complex process that depends on the drug being swallowed, surviving exposure to gastric acid, avoiding unacceptable food binding, being absorbed across the small bowel mucosa into the portal venous system, and surviving metabolism by gut wall or liver enzymes (‘first-pass metabolism’). As a consequence, absorption is frequently incomplete following oral administration. The term ‘bioavailability’ describes the proportion of the dose that reaches the systemic circulation intact. • Buccal, intranasal and sublingual (SL). These routes have the advantage of enabling rapid absorption into the systemic circulation without the uncertainties associated with oral administration (e.g. organic nitrates for angina pectoris, triptans for migraine, opioid analgesics). • Rectal (PR). The rectal mucosa is occasionally used as a site of drug administration when the oral route is compromised because of nausea and vomiting or unconsciousness (e.g. diazepam in status epilepticus). Fig. 2.3 Pharmacokinetics summary. Most drugs are taken orally, are absorbed from the intestinal lumen and enter the portal venous system to be conveyed to the liver, where they may be subject to first-pass metabolism and/or excretion in bile. Active drugs then enter the systemic circulation, from which they may diffuse (or sometimes be actively transported) in and out of the interstitial and intracellular fluid compartments. Drug that remains in circulating plasma is subject to liver metabolism and renal excretion. Drugs excreted in bile may be reabsorbed, creating an enterohepatic circulation. First-pass metabolism in the liver is avoided if drugs are administered via the buccal or rectal mucosa, or parenterally (e.g. by intravenous injection). I n t e r s t i t i a l

f l u i d Intracellular fluid Kidney Liver Parenteral Mouth Stomach Small intestine Large intestine Rectum Buccal Excretion in urine Excretion in faeces Portal venous system Intestinal wall enzymes Liver enzymes Metabolism Circulating plasma Rectal Oral

18 • CLINICAL THERAPEUTICS AND GOOD PRESCRIBING Drug excretion Excretion is the process by which drugs and their metabolites are removed from the body. Renal excretion is the usual route of elimination for drugs or their metabolites that are of low molecular weight and sufficiently water-soluble to

avoid reabsorption from the renal tubule. Drugs bound to plasma proteins are not filtered by the glomeruli. The pH of the urine is more acidic than that of plasma, so that some drugs (e.g. salicylates) become un-ionised and tend to of inhaled delivery is via a nebulised solution created by using pressurised oxygen or air to break up solutions

and suspensions into small aerosol droplets that can be directly inhaled from the mouthpiece of the device. Drug distribution Distribution is the process by which drug molecules transfer into and out of the blood stream. This is influenced by the drug’s molecular size and lipid solubility, the extent to which it binds to proteins in plasma, its susceptibility to

drug transporters expressed on cell surfaces, and its binding to its molecular target and to other cellular proteins (which can be irreversible). Most drugs diffuse passively across capillary walls down a concentration gradient into the interstitial fluid until the concentration of free drug molecules in the interstitial fluid is equal to that in the

plasma. As drug molecules in the blood are removed by metabolism or excretion, the plasma concentration falls, drug molecules diffuse back from the tissue compartment into the blood and eventually all will be eliminated. Note that this reverse movement of drug away from the tissues will be prevented if further drug doses are administered and

absorbed into the plasma. Volume of distribution The apparent volume of distribution (Vd) is the volume into which a drug appears to have distributed following intravenous injection. It is calculated from the equation Vd D C

where D is the amount of drug given and C0 is the initial plasma concentration (Fig. 2.4A). Drugs that are highly bound to plasma proteins may have a Vd below 10 L (e.g. warfarin, aspirin), while those that diffuse into the interstitial fluid but do not enter cells because they have low lipid solubility may have a Vd between 10 and 30 L (e.g. gentamicin, amoxicillin). It is an ‘apparent’ volume because those drugs that are lipid-soluble and highly tissue-bound may have a Vd of greater than 100 L (e.g. digoxin, amitriptyline). Drugs with a larger Vd have longer half-lives (see below), take longer to reach steady state on repeated administration and are eliminated more slowly from the body following discontinuation. Drug elimination Drug metabolism Metabolism is the process by which drugs are chemically altered from a lipid-soluble form suitable for absorption and distribution to a more water-soluble form that is necessary for excretion. Some drugs, known as ‘prodrugs’, are inactive in the form in which they are administered but are converted to an active metabolite in vivo. Phase I metabolism involves oxidation, reduction or hydrolysis to make drug molecules suitable for phase II reactions or for excretion. Oxidation is by far the most common form of phase I reaction and chiefly involves members of the cytochrome P450 family of

membrane-bound enzymes in the endoplasmic reticulum of hepatocytes. Phase II metabolism involves combining phase I metabolites with an endogenous substrate to form an inactive conjugate that is much more water-soluble. Reactions include glucuronidation, sulphation, acetylation, methylation and conjugation with glutathione. This is necessary to enable renal excretion, because lipid-soluble metabolites will simply diffuse back into the body after glomerular filtration (p. 349). Fig. 2.4 Drug concentrations in plasma following single and multiple drug dosing. A In this example of first-order kinetics following a single intravenous dose, the time period required for the plasma drug concentration to halve (half-life, t1/2) remains constant throughout the elimination process. B After multiple dosing, the plasma drug concentration rises if each dose is administered before the previous dose has been entirely cleared. In this example, the drug's half-life is 30 hours, so that with daily dosing the peak, average and trough concentrations steadily increase as drug accumulates in the body (black line). Steady state is reached after approximately 5 half-lives, when the rate of elimination (the product of concentration and clearance) is equal to the rate of drug absorption (the product of rate of administration and bioavailability). The long half-life in this example means that it takes 6 days for steady state to be achieved and, for most of the first 3 days of treatment, plasma drug concentrations are below the therapeutic range. This problem can be overcome if a larger loading dose (red line) is used to achieve steady-state drug concentrations more rapidly. Time (hours) A constant fraction of drug is cleared in unit time t1/2= 8 hours C0 Plasma drug concentration

A Loading dose Dose Dose Dose Dose Dose Dose Dose Subtherapeutic Dose interval = 24 hours Time (days) Plasma drug concentration

Therapeutic range Adverse effects t1/2= 30 hours B

means that the effects of a new prescription, or dose titration, for a drug with a long half-life (e.g. digoxin – 36 hours) may not be known for a few days. In contrast, drugs with a very short half-life (e.g. dobutamine – 2 minutes) have to be given continuously by infusion but reach a new steady state within minutes. For drugs with a long half-life, if it is unacceptable to wait for 5 half-lives until concentrations within the therapeutic range are achieved, then an initial 'loading dose' can be given that is much larger than the maintenance dose and equivalent to the amount of drug required in the body at steady state. This achieves a peak plasma concentration close to the plateau concentration, which can then be maintained by successive maintenance doses. 'Steady state' actually involves fluctuations in drug concentrations, with peaks just after administration followed by troughs just prior to the next administration. The manufacturers of medicines recommend dosing regimens that predict that, for most patients, these oscillations result in troughs within the therapeutic range and peaks that are not high enough to cause adverse effects. The optimal dose interval is a compromise between convenience for the patient and a constant level of drug exposure. More frequent administration (e.g. 25 mg 4 times daily) achieves a smoother plasma concentration profile than 100 mg once daily but is much more difficult for patients to sustain. A solution to this need for compromise in dosing frequency for drugs with half-lives of less than 24 hours is the use of 'modified-release' formulations. These allow drugs to be absorbed more slowly from the gastrointestinal tract and reduce the oscillation in plasma drug concentration profile, which is especially important for drugs with a low therapeutic index (e.g.

levodopa). Inter-individual variation in drug responses Prescribers have numerous sources of guidance about how to use drugs appropriately (e.g. dose, route, frequency, duration) for many conditions. However, this advice is based on average dose–response data derived from observations in many individuals. When applying this information to an individual patient, prescribers must take account of inter-individual variability in response. Some of this variability is predictable and good prescribers are able to anticipate it and adjust their prescriptions accordingly to maximise the chances of benefit and minimise harm. Inter-individual variation in responses also mandates that effects of treatment should be monitored (p. 34). Some inter-individual variation in drug response is accounted for by differences in pharmacodynamics. For example, the beneficial natriuresis produced by the loop diuretic furosemide is often significantly reduced at a given dose in patients with renal impairment, while delirium caused by opioid analgesics is more likely in the elderly. Differences in pharmacokinetics more commonly account for different drug responses, however. Examples of factors influencing the absorption, metabolism and excretion of drugs are shown in Box 2.4. It is hoped that a significant proportion of the inter-individual variation in drug responses can be explained by studying genetic differences in single genes (‘pharmacogenetics’; Box 2.5) or the effects of multiple gene variants (‘pharmacogenomics’). The aim is to identify those patients most likely to benefit from specific treatments and those most susceptible to adverse effects. In this way, it may be possible to select drugs and dose regimens for individual patients to maximise the benefit-to-hazard ratio (‘personalised medicine’). be reabsorbed. Alkalination of the urine can hasten excretion (e.g. after a salicylate overdose; p. 138). For some drugs, active secretion into the proximal tubule lumen, rather than glomerular filtration, is the predominant mechanism of excretion (e.g. methotrexate, penicillin). Faecal excretion is the predominant route of elimination for drugs with high molecular weight, including those that are excreted in the bile after conjugation with glucuronide in the liver, and any drugs that are not absorbed after enteral administration. Molecules of drug or metabolite that are excreted in the bile enter the small intestine, where they may, if they are sufficiently lipid-soluble, be reabsorbed through the gut wall and return to the liver via the portal vein (see Fig. 2.3). This recycling between the liver, bile, gut and portal vein is known as ‘enterohepatic circulation’ and can significantly prolong the residence of drugs in the body. Elimination kinetics The net removal of drug from the circulation results from a combination of drug metabolism and excretion, and is usually described as ‘clearance’, i.e. the volume of plasma that is completely cleared of drug per unit time. For most drugs, elimination is a high-capacity process that does not become saturated, even at high dosage. The rate of elimination is therefore directly proportional to the drug concentration because of the ‘law of mass action’, whereby higher drug concentrations will drive faster metabolic reactions and support higher renal filtration rates. This results in ‘first-order’ kinetics, when a constant fraction of the drug remaining in the circulation is eliminated in a given time and the decline in concentration over time is exponential (Fig. 2.4A). This elimination can be described by the drug’s half-life (t1/2), i.e. the time taken for the plasma drug concentration to halve, which remains constant throughout the period of drug elimination. The significance of this phenomenon for prescribers is that the effect of increasing doses on plasma concentration is predictable – a doubled dose leads to a doubled concentration at all time points. For a few drugs in common use (e.g. phenytoin, alcohol), elimination capacity is exceeded (saturated) within the usual dose range. This is called ‘zero-order’ kinetics. Its significance for prescribers is that, if the rate of administration exceeds the maximum rate of elimination, the drug will accumulate progressively, leading to serious toxicity. Repeated dose regimens The goal of therapy is usually to maintain drug concentrations within the therapeutic range (see Fig. 2.2) over several days (e.g. antibiotics) or even for months or years

(e.g. antihypertensives, lipid-lowering drugs, thyroid hormone replacement therapy). This goal is rarely achieved with single doses, so prescribers have to plan a regimen of repeated doses. This involves choosing the size of each individual dose and the frequency of dose administration. As illustrated in Figure 2.4B, the time taken to reach drug concentrations within the therapeutic range depends on the half-life of the drug. Typically, with doses administered regularly, it takes approximately 5 half-lives to reach a ‘steady state’ in which the rate of drug elimination is equal to the rate of drug administration. This applies when starting new drugs and when adjusting doses of current drugs. With appropriate dose selection, steady-state drug concentrations will be maintained within the therapeutic range. This is important for prescribers because it

20 • CLINICAL THERAPEUTICS AND GOOD PRESCRIBING 2.5 Examples of pharmacogenetic variations that influence drug response Genetic variant Drug affected Clinical outcome Pharmacokinetic Aldehyde dehydrogenase-2 deficiency Ethanol Elevated blood acetaldehyde causes facial flushing and increased heart rate in ~50% of Japanese, Chinese and other Asian populations Acetylation Isoniazid, hydralazine, procainamide Increased responses in slow acetylators, up to 50% of some populations Oxidation (CYP2D6) Nortriptyline Increased risk of toxicity in poor metabolisers Codeine Reduced responses with slower conversion of codeine to more active morphine in poor metabolisers, 10% of European populations Increased risk of toxicity in ultra-fast metabolisers, 3% of Europeans but 40% of North Africans Oxidation (CYP2C18) Proguanil Reduced efficacy with slower conversion to active cycloguanil in poor metabolisers Oxidation (CYP2C9) Warfarin Polymorphisms known to influence dosages Oxidation (CYP2C19) Clopidogrel Reduced enzymatic activation results in reduced antiplatelet effect Sulphoxidation Penicillamine Increased risk of toxicity in poor metabolisers Human leucocyte antigen (HLA)-B*1502 Carbamazepine Increased risk of serious dermatological reactions (e.g. Stevens–Johnson syndrome) for 1 in 2000 in Caucasian populations (much higher in some Asian countries) Pseudocholinesterase deficiency Suxamethonium (succinylcholine) Decreased drug inactivation leads to prolonged paralysis and sometimes persistent apnoea requiring mechanical ventilation until the drug can be eliminated by alternate pathways; occurs in 1 in 1500 people Pharmacodynamic Glucose-6-phosphate dehydrogenase (G6PD) deficiency Oxidant drugs, including antimalarials (e.g. chloroquine, primaquine) Risk of haemolysis in G6PD deficiency Acute intermittent porphyria Enzyme-inducing drugs Increased risk of an acute attack SLC01B1 polymorphism Statins Increased risk of rhabdomyolysis HLA-B*5701 polymorphism Abacavir Increased risk of skin hypersensitivity reactions HLA-B*5801 polymorphism Allopurinol Increased risk of rashes in Han Chinese HLA-B*1502 polymorphism Carbamazepine Increased risk of skin hypersensitivity reactions in Han Chinese Hepatic nuclear factor 1 alpha (HNF1A) polymorphism Sulphonylureas Increased sensitivity to the blood glucose-lowering effects Human epidermal growth factor receptor 2 (HER2)-positive breast cancer cells Trastuzumab Increased sensitivity to the inhibitory effects on growth and division of the target cancer cells Age • Drug metabolism is low in the fetus and newborn, may be enhanced in young children, and becomes less effective with age • Drug excretion falls with the age-related decline in renal function Sex • Women have a greater proportion of body fat than men, increasing the volume of distribution and half-life of lipid-soluble drugs Body weight • Obesity increases volume of distribution and half-life of lipid-soluble drugs • Patients with higher lean body mass have larger body compartments into which drugs are distributed and may require higher doses Liver function • Metabolism of most drugs depends on several cytochrome P450 enzymes that are impaired in patients with advanced liver disease • Hypoalbuminaemia influences the distribution of drugs that are highly protein-bound Kidney function • Renal disease and the decline in renal

function with ageing may lead to drug accumulation Gastrointestinal function • Small intestinal absorption of oral drugs may be delayed by reduced gastric motility • Absorptive capacity of the intestinal mucosa may be reduced in disease (e.g. Crohn's or coeliac disease) or after surgical resection Food • Food in the stomach delays gastric emptying and reduces the rate (but not usually the extent) of drug absorption • Some food constituents bind to certain drugs and prevent their absorption Smoking • Tar in tobacco smoke stimulates the oxidation of certain drugs Alcohol • Regular alcohol consumption stimulates liver enzyme synthesis, while binge drinking may temporarily inhibit drug metabolism Drugs • Drug–drug interactions cause marked variation in pharmacokinetics (see Box 2.11) 2.4 Patient-specific factors that influence pharmacokinetics

Adverse outcomes of drug therapy The decision to prescribe a drug always involves a judgement of the balance between therapeutic benefits and risk of an adverse outcome. Both prescribers and patients tend to be more focused on the former but a truly informed decision requires consideration of both. Adverse drug reactions Some important definitions for the adverse effects of drugs are: • Adverse event. A harmful event that occurs while a patient is taking a drug, irrespective of whether the drug is suspected of being the cause. • Adverse drug reaction (ADR). An unwanted or harmful reaction that is experienced following the administration of a drug or combination of drugs under normal conditions of use and is suspected to be related to the drug. An ADR will usually require the drug to be discontinued or the dose reduced. • Side-effect. Any effect caused by a drug other than the intended therapeutic effect, whether beneficial, neutral or harmful. The term 'side-effect' is often used interchangeably with 'ADR', although the former usually implies an ADR that occurs during exposure to normal therapeutic drug concentrations (e.g. vasodilator-induced ankle oedema). • Hypersensitivity reaction. An ADR that occurs as a result of an immunological reaction and often at exposure to subtherapeutic drug concentrations. Some of these reactions are immediate and result from the interaction of drug antigens with immunoglobulin E (IgE) on mast cells and basophils, which causes a release of vasoactive biomolecules (e.g. penicillin-related anaphylaxis). 'Anaphylactoid' reactions present similarly but occur through a direct non-immune-mediated release of the same mediators or result from direct complement activation (p. 75). Hypersensitivity reactions may occur via other mechanisms such as antibody-dependent (IgM or IgG), immune complex-mediated or cell-mediated pathways. • Drug toxicity. Adverse effects of a drug that occur because the dose or plasma concentration has risen above the therapeutic range, either unintentionally or intentionally (drug overdose; see Fig. 2.2 and p. 137). • Drug abuse. The misuse of recreational or therapeutic drugs that may lead to addiction or dependence, serious physiological injury (such as liver damage), psychological harm (abnormal behaviour patterns, hallucinations, memory loss) or death (p. 1184). Prevalence of ADRs ADRs are a common cause of illness, accounting in the UK for approximately 3% of consultations in primary care and 7% of emergency admissions to hospital, and affecting around 15% of hospital inpatients. Many 'disease' presentations are eventually attributed to ADRs, emphasising the importance of always taking a careful drug history (Box 2.6). Factors accounting for the rising prevalence of ADRs are the increasing age of patients, polypharmacy (higher risk of drug interactions), increasing availability of over-the-counter medicines, increasing use of herbal or traditional medicines, and the increase in medicines available via the Internet that can be purchased without a prescription from a health-care professional. Risk factors for ADRs are shown in Box 2.7. 2.7 Risk factors for adverse drug reactions Patient factors • Elderly age (e.g. low

physiological reserve) • Gender (e.g. ACE inhibitor-induced cough in women) • Polypharmacy (e.g. drug interactions) • Genetic predisposition (see Box 2.5) • Hypersensitivity/allergy (e.g. β-lactam antibiotics) • Diseases altering pharmacokinetics (e.g. hepatic or renal impairment) or pharmacodynamic responses (e.g. bladder instability) • Adherence problems (e.g. cognitive impairment) Drug factors • Steep dose–response curve (e.g. insulin) • Low therapeutic index (e.g. digoxin, cytotoxic drugs) Prescriber factors • Inadequate understanding of principles of clinical pharmacology • Inadequate knowledge of the patient • Inadequate knowledge of the prescribed drug • Inadequate instructions and warnings provided to patients • Inadequate monitoring arrangements planned (ACE = angiotensin-converting enzyme) 2.6 How to take a drug history Information from the patient (or carer) Use language that patients will understand (e.g. ‘medicines’ rather than ‘drugs’, which may be mistaken for drugs of abuse) while gathering the following information: • Current prescribed drugs, including formulations (e.g. modifiedrelease tablets), doses, routes of administration, frequency and timing, duration of treatment • Other medications that are often forgotten (e.g. contraceptives, over-the-counter drugs, herbal remedies, vitamins) • Drugs that have been taken in the recent past and reasons for stopping them • Previous drug hypersensitivity reactions, their nature and time course (e.g. rash, anaphylaxis) • Previous ADRs, their nature and time course (e.g. ankle oedema with amlodipine) • Adherence to therapy (e.g. ‘Are you taking your medication regularly?’) Information from GP medical records and/or pharmacist • Up-to-date list of medications • Previous ADRs • Last order dates for each medication Inspection of medicines • Drugs and their containers (e.g. blister packs, bottles, vials) should be inspected for name, dosage, and the number of dosage forms taken since dispensed (ADR = adverse drug reaction)

 2.9 DoTS classification of adverse drug reactions Category Example Dose Below therapeutic dose Anaphylaxis with penicillin In the therapeutic dose range Nausea with morphine At high doses Hepatotoxicity with paracetamol Timing With the first dose Anaphylaxis with penicillin Early stages of treatment Hyponatraemia with diuretics On stopping treatment Benzodiazepine withdrawal syndrome Significantly delayed Clear-cell cancer with diethylstilboestrol Susceptibility See patient factors in Box 2.7 (INR = international normalised ratio) 2.8 Drugs that are common causes of adverse drug reactions Drug or drug class Common adverse drug reactions ACE inhibitors (e.g. lisinopril) Renal impairment Hyperkalaemia Antibiotics (e.g. amoxicillin) Nausea Diarrhoea Anticoagulants (e.g. warfarin, heparin) Bleeding Antipsychotics (e.g. haloperidol) Falls Sedation Delirium Aspirin Gastrotoxicity (dyspepsia, gastrointestinal bleeding) Benzodiazepines (e.g. diazepam) Drowsiness Falls β-blockers (e.g. atenolol) Cold peripheries Bradycardia Calcium channel blockers (e.g. amlodipine) Ankle oedema Digoxin Nausea and anorexia Bradycardia Diuretics (e.g. furosemide, bendroflumethiazide) Dehydration Electrolyte disturbance (hypokalaemia, hyponatraemia) Hypotension Renal impairment Insulin Hypoglycaemia NSAIDs (e.g. ibuprofen) Gastrotoxicity (dyspepsia, gastrointestinal bleeding) Renal impairment Opioid analgesics (e.g. morphine) Nausea and vomiting Delirium Constipation (ACE = angiotensin-converting enzyme; NSAID = non-steroidal anti-inflammatory drug) ADRs are important because they reduce quality of life for patients, reduce adherence to and therefore efficacy of beneficial treatments, cause diagnostic confusion, undermine the confidence of patients in their health-care professional(s) and consume health-care resources. Retrospective analysis of ADRs has shown that more than half could have been avoided if the prescriber had taken more care in anticipating the potential hazards of drug therapy. For example, non-steroidal anti-inflammatory drug (NSAID) use accounts for many thousands of

emergency admissions, gastrointestinal bleeding episodes and a significant number of deaths. In many cases, the patients are at increased risk due to their age, interacting drugs (e.g. aspirin, warfarin) or a past history of peptic ulcer disease. Drugs that commonly cause ADRs are listed in Box 2.8. Prescribers and their patients ideally want to know the frequency with which ADRs occur for a specific drug. Although this may be well characterised for more common ADRs observed in clinical trials, it is less clear for rarely reported ADRs when the total numbers of reactions and patients exposed are not known. The words used to describe frequency can be misinterpreted by patients but widely accepted meanings include: very common (10% or more), common (1–10%), uncommon (0.1–1%), rare (0.01–0.1%) and very rare (0.01% or less). Classification of ADRs ADRs have traditionally been classified into two major groups: • Type A ('augmented') ADRs. These are predictable from the known pharmacodynamic effects of the drug and are dose-dependent, common (detected early in drug development) and usually mild. Examples include constipation caused by opioids, hypotension caused by antihypertensives and dehydration caused by diuretics. • Type B ('bizarre') ADRs. These are not predictable, are not obviously dose-dependent in the therapeutic range, are rare (remaining undiscovered until the drug is marketed) and often severe. Patients who experience type B reactions are generally 'hyper-susceptible' because of unpredictable immunological or genetic factors (e.g. anaphylaxis caused by penicillin, peripheral neuropathy caused by isoniazid in poor acetylators). This simple classification has shortcomings, and a more detailed classification based on dose (see Fig. 2.2), timing and susceptibility (DoTS) is now used by those analysing ADRs in greater depth (Box 2.9). The AB classification can be extended as a reminder of some other types of ADR: • Type C ('chronic/continuous') ADRs. These occur only after prolonged continuous exposure to a drug. Examples include osteoporosis caused by glucocorticoids, retinopathy caused by chloroquine, and tardive dyskinesia caused by phenothiazines. • Type D ('delayed') ADRs. These are delayed until long after drug exposure, making diagnosis difficult. Examples include malignancies that may emerge after immunosuppressive treatment post-transplantation (e.g. azathioprine, tacrolimus) and vaginal cancer occurring many years after exposure to diethylstilboestrol. • Type E ('end-of-treatment') ADRs. These occur after abrupt drug withdrawal (see Box 2.3). A teratogen is a drug with the potential to affect the development of the fetus in the first 10 weeks of intrauterine life (e.g. phenytoin, warfarin). The thalidomide disaster in the early 1960s highlighted the risk of teratogenicity and led to mandatory testing of all new drugs. Congenital defects in a live infant or aborted fetus should

of prescribers of a particular drug are issued with questionnaires concerning the clinical outcome for their patients, and the collection of population statistics. Many health-care systems routinely collect patient-identifiable data on prescriptions (a surrogate marker of exposure to a drug), health-care events (e.g. hospitalisation, operations, new clinical diagnoses) and other clinical data (e.g. haematology, biochemistry). As these records are linked, with appropriate safeguards for confidentiality and data protection, they are providing a much more powerful mechanism for assessing both the harms and benefits of drugs. All prescribers will inevitably see patients experiencing ADRs caused by prescriptions written by themselves or their colleagues. It is important that these are recognised early. In addition to the features in Box 2.10, features that should raise suspicion of an ADR and the need to respond (by drug withdrawal, dosage reduction or reporting to the regulatory authorities) include: • concern expressed by a patient that a drug has

harmed them • abnormal clinical measurements (e.g. blood pressure, temperature, pulse, blood glucose and weight) or laboratory results (e.g. abnormal liver or renal function, low haemoglobin or white cell count) while on drug therapy • new therapy started that could be in response to an ADR (e.g. omeprazole, allopurinol, naloxone) • the presence of risk factors for ADRs (see Box 2.7). Drug interactions A drug interaction has occurred when the administration of one drug increases or decreases the beneficial or adverse responses to another drug. Although the number of potential interacting drug combinations is very large, only a small number are common in clinical practice. Important drug interactions are most likely to occur when the affected drug has a low therapeutic index, steep dose–response curve, high first-pass or saturable metabolism, or a single mechanism of elimination. Mechanisms of drug interactions Pharmacodynamic interactions occur when two drugs produce additive, synergistic or antagonistic effects at the same drug target (e.g. receptor, enzyme) or physiological system (e.g. electrolyte excretion, heart rate). These are the most common interactions in clinical practice and some important examples are given in Box 2.11. Pharmacokinetic interactions occur when the administration of a second drug alters the concentration of the first at its site of action. There are numerous potential mechanisms: • Absorption interactions. Drugs that either delay (e.g. anticholinergic drugs) or enhance (e.g. prokinetic drugs) gastric emptying influence the rate of rise in plasma concentration of other drugs but not the total amount of drug absorbed. Drugs that bind to form insoluble complexes or chelates (e.g. aluminium-containing antacids binding with ciprofloxacin) can reduce drug absorption. • Distribution interactions. Co-administration of drugs that compete for protein binding in plasma (e.g. phenytoin and diazepam) can increase the unbound drug concentration, but the effect is usually short-lived due to increased elimination and hence restoration of the pre-interaction equilibrium. provoke suspicion of an ADR and a careful exploration of drug exposures (including self-medication and herbal remedies). Detecting ADRs – pharmacovigilance Type A ADRs become apparent early in the development of a new drug. By the time a new drug is licensed and launched on to a possible worldwide market, however, a relatively small number of patients (just several hundred) may have been exposed to it, meaning that rarer but potentially serious type B ADRs may remain undiscovered. Pharmacovigilance is the process of detecting (‘signal generation’) and evaluating ADRs in order to help prescribers and patients to be better informed about the risks of drug therapy. Drug regulatory agencies may respond to this information by placing restrictions on the licensed indications, reducing the recommended dose range, adding special warnings and precautions for prescribers in the product literature, writing to all health-care professionals or withdrawing the product from the market. Voluntary reporting systems allow health-care professionals and patients to report suspected ADRs to the regulatory authorities. A good example is the ‘Yellow Card’ scheme that was set up in the UK in response to the thalidomide tragedy. Reports are analysed to assess the likelihood that they represent a true ADR (Box 2.10). Although voluntary reporting is a continuously operating and effective early-warning system for previously unrecognised rare ADRs, its weaknesses include low reporting rates (only 3% of all ADRs and 10% of serious ADRs are ever reported), an inability to quantify risk (because the ratio of ADRs to prescriptions is unknown), and the influence of prescriber awareness on likelihood of reporting (reporting rates rise rapidly following publicity about potential ADRs). More systematic approaches to collecting information on ADRs include ‘prescription event monitoring’, in which a sample 2.10 TREND analysis of suspected adverse drug reactions Factor Key question Comment Temporal relationship What is the time interval between the start of drug therapy and the reaction? Most ADRs occur soon after starting treatment and within hours in the case of anaphylactic reactions Re-challenge What happens when the patient is re-challenged with the drug? Re-challenge is rarely

possible because of the need to avoid exposing patients to unnecessary risk Exclusion Have concomitant drugs and other non-drug causes been excluded? ADR is a diagnosis of exclusion following clinical assessment and relevant investigations for non-drug causes Novelty Has the reaction been reported before? The suspected ADR may already be recognised and mentioned in the SPC approved by the regulatory authorities De-challenge Does the reaction improve when the drug is withdrawn or the dose is reduced? Most, but not all, ADRs improve on drug withdrawal, although recovery may be slow (SPC = summary of product characteristics)

24 • CLINICAL THERAPEUTICS AND GOOD PRESCRIBING consequences of drug–drug interactions by taking a careful drug history (see Box 2.6) before prescribing additional drugs, only prescribing for clear indications, and taking special care when prescribing drugs with a narrow therapeutic index (e.g. warfarin). When prescribing an interacting drug is unavoidable, good prescribers will seek further information and anticipate the potential risk. This will allow them to provide special warnings for the patient and arrange for monitoring, either of the clinical effects (e.g. coagulation tests for warfarin) or of plasma concentration (e.g. digoxin). Medication errors A medication error is any preventable event that may lead to inappropriate medication use or patient harm while the medication is in the control of the health-care professional or patient. Errors may occur in prescribing, dispensing, preparing solutions, administration or monitoring. Many ADRs are considered in retrospect to have been ‘avoidable’ with more care or forethought; in other words, an adverse event considered by one prescriber to be an unfortunate ADR might be considered by another to be a prescribing error. Medication errors are very common. Several thousand medication orders are dispensed and administered each day in a medium-sized hospital. Recent UK studies suggest that • Metabolism interactions. Many drugs rely on metabolism by different isoenzymes of cytochrome P450 (CYP) in the liver. CYP enzyme inducers (e.g. phenytoin, rifampicin) generally reduce plasma concentrations of other drugs, although they may enhance activation of prodrugs. CYP enzyme inhibitors (e.g. clarithromycin, cimetidine, grapefruit juice) have the opposite effect. Enzyme induction effects usually take a few days to manifest because of the need to synthesise new CYP enzyme, in contrast to the rapid effects of enzyme inhibition. • Excretion interactions. These primarily affect renal excretion. For example, drug-induced reduction in glomerular filtration rate (e.g. diuretic-induced dehydration, angiotensin-converting enzyme (ACE) inhibitors, NSAIDs) can reduce the clearance and increase the plasma concentration of many drugs, including some with a low therapeutic index (e.g. digoxin, lithium, aminoglycoside antibiotics). Less commonly, interactions may be due to competition for a common tubular organic anion transporter (e.g. methotrexate excretion may be inhibited by competition with NSAIDs). Avoiding drug interactions Drug interactions are increasing as patients are prescribed more medicines (polypharmacy). Prescribers can avoid the adverse 2.11 Common drug interactions Mechanism Object drug Precipitant drug Result Pharmaceutical* Chemical reaction Sodium bicarbonate Calcium gluconate Precipitation of insoluble calcium carbonate Pharmacokinetic Reduced absorption Tetracyclines Calcium, aluminium, and magnesium salts Reduced tetracycline absorption Reduced protein binding Phenytoin Aspirin Increased unbound and reduced total phenytoin plasma concentration Reduced metabolism: CYP3A4 Amiodarone Grapefruit juice Cardiac arrhythmias because of prolonged QT interval (p. 476) Warfarin Clarithromycin Enhanced anticoagulation CYP2C19 Phenytoin Omeprazole Phenytoin toxicity CYP2D6 Clozapine Paroxetine Clozapine toxicity Xanthine oxidase Azathioprine Allopurinol Azathioprine toxicity Monoamine oxidase Catecholamines Monoamine oxidase inhibitors Hypertensive crisis due to monoamine toxicity Increased metabolism (enzyme induction) Ciclosporin St John’s wort Loss of

immunosuppression Reduced renal elimination Lithium Diuretics Lithium toxicity Methotrexate NSAIDs Methotrexate toxicity Pharmacodynamic Direct antagonism at same receptor Opioids Naloxone Reversal of opioid effects used therapeutically Salbutamol Atenolol Inhibits bronchodilator effect Direct potentiation in same organ system Benzodiazepines Alcohol Increased sedation ACE inhibitors NSAIDs Increased risk of renal impairment Indirect potentiation by actions in different organ systems Digoxin Diuretics Digoxin toxicity enhanced because of hypokalaemia Warfarin Aspirin, NSAIDs Increased risk of bleeding because of gastrotoxicity and antiplatelet effects Diuretics ACE inhibitors Blood pressure reduction (may be therapeutically advantageous) because of the increased activity of the renin–angiotensin system in response to diuresis *Pharmaceutical interactions are related to the formulation of the drugs and occur before drug absorption. (ACE = angiotensin-converting enzyme; NSAID = non-steroidal anti-inflammatory drug)

Adverse outcomes of drug therapy • 25

7–9% of hospital prescriptions contain an error, and most are written by junior doctors. Common prescribing errors in hospitals include omission of medicines (especially failure to prescribe regular medicines at the point of admission or discharge, i.e. ‘medicines reconciliation’), dosing errors, unintentional prescribing and poor use of documentation (Box 2.12). Most prescription errors result from a combination of failures by the individual prescriber and the health-service systems in which they work (Box 2.13). Health-care organisations increasingly encourage reporting of errors within a ‘no-blame culture’ so that they can be subject to ‘root cause analysis’ using human error theory (Fig. 2.5). Prevention is targeted at the factors in Box 2.13, and can be supported by prescribers communicating and cross-checking with colleagues (e.g. when calculating doses adjusted for body weight, or planning appropriate monitoring after drug administration), and by health-care systems providing clinical pharmacist support (e.g. for checking the patient’s previous medications and current prescriptions) and electronic prescribing (which avoids errors due to illegibility or serious dosing mistakes, and may be combined with a clinical decision support system to take account of patient characteristics and drug history, and provide warnings of potential contraindications and drug interactions). 2.13 Causes of prescribing errors Systems factors • Working hours of prescribers (and others) • Patient throughput • Professional support and supervision by colleagues • Availability of information (medical records) • Design of prescription forms • Distractions • Availability of decision support • Checking routines (e.g. clinical pharmacy) • Reporting and reviewing of incidents Prescriber factors Knowledge • Clinical pharmacology principles • Drugs in common use • Therapeutic problems commonly encountered • Knowledge of workplace systems Skills • Taking a good drug history • Obtaining information to support prescribing • Communicating with patients • Numeracy and calculations • Prescription writing Attitudes • Coping with risk and uncertainty • Monitoring of prescribing • Checking routines 2.12 Hospital prescribing errors Error type Approximate % of total Omission on admission

Underdose

Overdose

Strength/dose missing

Omission on discharge

Administration times incorrect/missing

Duplication

Product or formulation not specified

Incorrect formulation

No maximum dose

Unintentional prescribing

No signature

Clinical contraindication

Incorrect route

No indication

Intravenous instructions incorrect/missing

Drug not prescribed but indicated

Drug continued for longer than needed

Route of administration missing

Start date incorrect/missing

Risk of drug interaction < 0.5 Controlled drug requirements incorrect/missing < 0.5 Daily dose divided incorrectly < 0.5 Significant allergy < 0.5 Drug continued in spite of adverse effects < 0.5 Premature discontinuation < 0.5 Failure to respond to out-of-range drug level < 0.5 Fig. 2.5 Human error theory. Unintended errors may occur because the prescriber fails to complete the prescription correctly (a slip; e.g. writes the dose in ‘mg’ not ‘micrograms’) or forgets part of the action that is important for success (a lapse; e.g. forgets to co-prescribe folic acid with methotrexate); prevention requires the system to provide appropriate checking routines. Intended errors occur when the prescriber acts incorrectly due to lack of knowledge (a mistake; e.g. prescribes atenolol for a patient with known severe asthma because of ignorance about the contraindication); prevention must focus on training the prescriber. Planned action Prescribing Intended action Correct action Intended outcome Unintended action Lapse Slip Wrong plan selected (Causes include poor training and lack of experience) Correct plan known but not executed (Causes include workload, time pressures, distractions) Prescription not as intended Prescriber unaware Prescription incomplete or forgotten Prescriber may remember Violation Mistake Prescription as intended but written based on the wrong principles or lack of knowledge Prescriber unaware Deliberate deviations from standard practice Prescriber aware

26 • CLINICAL THERAPEUTICS AND GOOD PRESCRIBING cell lines, molecular cloning and purification processes. After the patent for the originator product expires, other manufacturers may develop similar products (‘biosimilars’) that share similar pharmacological actions but are not completely identical. ‘Biosimilars’ are considered distinct from ‘generic’ medications, as complex biological molecules are more susceptible to differences in manufacturing processes than conventional small-molecule-type pharmaceuticals. The number of new drugs produced by the pharmaceutical industry has declined in recent years. The traditional approach of targeting membrane-bound receptors and enzymes with small molecules (see Box 2.2) is now giving way to new targets, such as complex second-messenger systems, cytokines, nucleic acids and cellular networks. These require novel therapeutic agents, which present new challenges for ‘translational medicine’, the discipline of converting scientific discoveries into a useful medicine with a well-defined benefit–risk profile (Box 2.15). Licensing new medicines New drugs are given a ‘market authorisation’, based on the evidence of quality, safety and efficacy presented by the manufacturer. The regulator not only will approve the drug but also will take great care to ensure that the accompanying information reflects the evidence that has been presented. The summary of product characteristics (SPC), or ‘label’, provides detailed information about indications, dosage, adverse effects, warnings, monitoring and so on. If approved, drugs can be made available with different levels of restriction: • Controlled drug (CD). These drugs are subject to strict legal controls on supply and possession, usually due to their abuse potential (e.g. opioid analgesics). 2.14 Clinical development of new drugs Phase I • Healthy volunteers (20–80) • These involve initial single-dose, ‘first-into-man’ studies, followed by repeated-dose studies. They aim to establish the basic pharmacokinetic and pharmacodynamic properties, and short-term safety • Duration: 6–12 months Phase II • Patients (100–200) • These investigate clinical effectiveness (‘proof of concept’), safety and dose–response relationship, often with a surrogate clinical endpoint, in the target patient group to determine the optimal dosing regimen for larger confirmatory studies • Duration: 1–2 years Phase III • Patients (100s–1000s) • These are large, expensive clinical trials that confirm safety and efficacy in the target patient population, using relevant clinical endpoints. They may be placebo-controlled studies or comparisons with other active compounds • Duration: 1–2 years Phase IV • Patients (100s–1000s) • These are undertaken after the medicine has been marketed for its first indication to evaluate new indications, new doses or formulations, long-term safety or cost-effectiveness Responding when an error is discovered All prescribers will make errors. When they do, their first duty is to protect the patient’s safety. This will involve a clinical review and the taking of any steps that will reduce harm (e.g. remedial treatment, monitoring, recording the event in the notes, informing colleagues). Patients should be informed if they have been exposed to potential harm. For errors that do not reach the patient, it is the prescriber’s duty to report them, so that others can learn from the experience and take the opportunity to reflect on how a similar incident might be avoided in the future. Drug regulation and management Given the powerful beneficial and potentially adverse effects of drugs, the production and use of medicines are strictly regulated (e.g. by the Food and Drug Administration in the USA, Medicines and Healthcare Products Regulatory Agency in the UK, and Central Drugs Standard Control Organisation in India). Regulators are responsible for licensing medicines, monitoring their safety (pharmacovigilance; p. 23), approving clinical trials, and inspecting and maintaining standards of drug development and manufacture. In addition, because of the high costs of drugs and their adverse effects, health-care services must prioritise their use in light of the evidence of their benefit and harm, a process referred to as ‘medicines management’. Drug development and marketing Naturally occurring products have been used to treat illnesses for thousands of years and some remain in common use

today. Examples include morphine from the opium poppy (Papaver somniferum), digitalis from the foxglove (Digitalis purpurea), curare from the bark of a variety of species of South American trees, and quinine from the bark of the Cinchona species. Although plants and animals remain a source of discovery, the majority of new drugs come from drug discovery programmes that aim to identify small-molecule compounds with specific interactions with a molecular target that will induce a predicted biological effect. The usual pathway for development of these small molecules includes: identifying a plausible molecular target by investigating pathways in disease; screening a large library of compounds for those that interact with the molecular target in vitro; conducting extensive medicinal chemistry to optimise the properties of lead compounds; testing efficacy and toxicity of these compounds in vitro and in animals; and undertaking a clinical development programme (Box 2.14). This process typically takes longer than 10 years and may cost up to US$1 billion. Manufacturers have a defined period of exclusive marketing of the drug while it remains protected by an original patent, typically 10–15 years, during which time they must recoup the costs of developing the drug. Meanwhile, competitor companies will often produce similar ‘me too’ drugs of the same class. Once the drug’s patent has expired, ‘generic’ manufacturers may step in to produce cheaper formulations of the drug. Paradoxically, if a generic drug is produced by only one manufacturer, the price may actually rise, sometimes substantially. Newer ‘biological’ products are based on large molecules (e.g. human recombinant antibodies) derived from complex manufacturing processes involving specific

Managing the use of medicines Many medicines meet the three key regulatory requirements of quality, safety and efficacy. Although prescribers are legally entitled to prescribe any of them, it is desirable to limit the choice so that treatments for specific diseases can be focused on the most effective and cost-effective options, prescribers (and patients) gain familiarity with a smaller number of medicines, and pharmacies can concentrate stocks on them. The process of ensuring optimal use of available medicines is known as ‘medicines management’ or ‘quality use of medicines’. It involves careful evaluation of the evidence of benefit and harm from using the medicine, an assessment of cost-effectiveness, and support for processes to implement the resulting recommendations. These activities usually involve both national (e.g. National Institute for Health and Care Excellence (NICE) in the UK) and local organisations (e.g. drug and therapeutics committees). Evaluating evidence The principles of evidence-based medicine are described on page 10. Drugs are often evaluated in high-quality randomised controlled trials, the results of which can be considered by systematic review (Fig. 2.6). Ideally, data are available not only for comparison with placebo but also for ‘head-to-head’ comparison with alternative therapies. Trials are conducted in selected patient populations, however, and are not representative of every clinical scenario, so extrapolation to individual patients is not always straightforward. Other subtle bias may be introduced because of the sources of funding (e.g. pharmaceutical industry) and the interests of the investigators in being involved in research that has a ‘positive’ impact. These biases may be manifest in the way the trials are conducted or in how they are interpreted or reported. A common example of the latter is the difference between relative and absolute risk of clinical events reported in prevention trials. If a clinical event is encountered in the placebo arm at a rate of 1 in 50 patients (2%) but only 1 in 100 patients (1%) in the active treatment arm, then the impact of treatment can be described as either a 50% relative risk reduction or 1% absolute risk reduction. Although the former sounds more impressive, it is the latter that is of more importance

to the • Prescription-only medicine (PoM). These are available only from a pharmacist and can be supplied only if prescribed by an appropriate practitioner. • Pharmacy (P). These are available only from a pharmacist but can be supplied without a prescription. • General sales list (GSL). These medicines may be bought ‘over the counter’ (OTC) from any shop and without a prescription. Although the regulators take great care to agree the exact indications for prescribing a medicine, based on the evidence provided by the manufacturer, there are some circumstances in which prescribers may direct its use outside the terms stated in the SPC (‘off-label’ prescribing). Common situations where this might occur include prescribing outside the approved age group (e.g. prescribing for children) or using an alternative formulation (e.g. administering a medicine provided in a solid form as an oral solution). Other important examples might include prescribing for an indication for which there are no approved medicines or where all of the approved medicines have caused unacceptable adverse effects. Occasionally, medicines may be prescribed when there is no marketing authorisation in the country of use. Examples include when a medicine licensed in another country is imported for use for an individual patient (‘unlicensed import’) or when a patient requires a specific preparation of a medicine to be manufactured (‘unlicensed special’). When prescribing is ‘off-label’ or ‘unlicensed’, there is an increased requirement for prescribers to be able to justify their actions and to inform and agree the decision with the patient. Drug marketing The marketing activities of the pharmaceutical industry are well resourced and are important in the process of recouping the massive costs of drug development. In some countries, such as the USA, it is possible to promote a new drug by direct-toconsumer advertising, although this is illegal in the countries of the European Union. A major focus is on promotion to prescribers via educational events, sponsorship of meetings, advertisements in journals, involvement with opinion leaders, and direct contact by company representatives. Such largesse has the potential to cause significant conflicts of interest and might tempt prescribers to favour one drug over another, even in the face of evidence on effectiveness or cost-effectiveness. 2.15 Novel therapeutic alternatives to conventional small-molecule drugs Approaches Therapeutic indications Challenges Monoclonal antibodies Targeting of receptors or other molecules with relatively specific antibodies Cancer Chronic inflammatory diseases (e.g. rheumatoid arthritis, inflammatory bowel disease) Selectivity of action Complex manufacturing process Small interfering RNA (siRNA) Inhibition of gene expression Macular degeneration Delivery to target Gene therapy Delivery of modified genes that supplement or alter host DNA Cystic fibrosis Cancer Cardiovascular disease Delivery to target Adverse effects of delivery vector (e.g. virus) Stem cell therapy Stem cells differentiate and replace damaged host cells Parkinson’s disease Spinal cord injury Ischaemic heart disease Delivery to target Immunological compatibility Long-term effects unknown

28 • CLINICAL THERAPEUTICS AND GOOD PRESCRIBING Implementing recommendations Many recommendations about drug therapy are included in clinical guidelines written by an expert group after systematic review of the evidence. Guidelines provide recommendations rather than obligations for prescribers and are helpful in promoting more consistent and higher-quality prescribing. They are often written without concern for cost-effectiveness, however, and may be limited by the quality of available evidence. Guidelines cannot anticipate the extent of the variation between individual patients who may, for example, have unexpected contraindications to recommended drugs or choose different priorities for treatment. When deviating from respected national guidance, prescribers should be able to justify their practice. Additional recommendations for prescribing are often implemented locally or imposed by bodies responsible for paying for health care. Most health-care units have a drug and therapeutics committee (or equivalent)

comprised of senior and junior medical staff, pharmacists and nurses, as well as managers (because of the implications of the committee's work for governance and resources). This group typically develops local prescribing policy and guidelines, maintains a local drug formulary and evaluates requests to use new drugs. The local formulary contains a more limited list than any national formulary (e.g. British National Formulary) because the latter lists all licensed medicines that can be prescribed legally, while the former contains only those that the health-care organisation has approved for local use. The local committee may also be involved, with local specialists, in providing explicit protocols for management of clinical scenarios. Prescribing in practice Decision-making in prescribing Prescribing should be based on a rational approach to a series of challenges (see Box 2.1). individual patient. It means that the number of patients that needed to be treated (NNT) for 1 to benefit (compared to placebo) was 100. This illustrates how large clinical trials of new medicines can produce highly statistically significant and impressive relative risk reductions and still predict a very modest clinical impact. Evaluating cost-effectiveness New drugs often represent an incremental improvement over the current standard of care but are usually more expensive. Health-care budgets are limited in every country and so it is impossible to fund all new medicines. This means that very difficult financial decisions have to be taken with due regard to the principles of ethical justice. The main approach taken is cost-effectiveness analysis (CEA), where a comparison is made between the relative costs and outcomes of different courses of action. CEA is usually expressed as a ratio where the denominator is a gain in health and the numerator is the cost associated with the health gain. A major challenge is to compare the value of interventions for different clinical outcomes. One method is to calculate the quality-adjusted life years (QALYs) gained if the new drug is used rather than standard treatment. This analysis involves estimating the 'utility' of various health states between 1 (perfect health) and 0 (dead). If the additional costs and any savings are known, then it is possible to derive the incremental cost-effectiveness ratio (ICER) in terms of cost/QALY. These principles are exemplified in Box 2.16. There are, however, inherent weaknesses in this kind of analysis: it usually depends on modelling future outcomes well beyond the duration of the clinical trial data that are available; it assumes that QALYs gained at all ages are of equivalent value; and the appropriate standard care against which the new drug should be compared is often uncertain. These pharmacoeconomic assessments are challenging and resource-intensive, and are undertaken at national level in most countries, e.g. in the UK by NICE. Fig. 2.6 Systematic review of the evidence from randomised controlled clinical trials. This forest plot shows the effect of warfarin compared with placebo on the likelihood of stroke in patients with atrial fibrillation in five randomised controlled trials that passed the quality criteria required for inclusion in a meta-analysis. For each trial, the purple box is proportionate to the number of participants. The tick marks show the mean odds ratio and the black lines indicate its 95% confidence intervals. Note that not all the trials showed statistically significant effects (i.e. the confidence intervals cross 1.0). However, the meta-analysis, represented by the black diamond, confirms a highly significant statistical benefit. The overall odds ratio is approximately 0.4, indicating a mean 60% risk reduction with warfarin treatment in patients with the characteristics of the participants in these trials. Odds ratio Favours treatment 0.1 0.2 0.5

Favours placebo 2.16 Cost-effectiveness analysis A clinical trial lasting 2 years compares two interventions for the treatment of colon cancer: • Treatment A: standard treatment, cost £1000/year, oral therapy • Treatment B: new treatment, cost £6000/year, monthly intravenous infusions, often followed by a week of nausea. The new treatment (B) significantly increases the average time to progression (18 months versus 12 months) and reduces overall mortality (40%

versus 60%). The health economist models the survival curves from the trial in order to undertake a cost–utility analysis and concludes that: • Intervention A: allows an average patient to live for 2 extra years at a utility 0.7 = 1.4 QALYs (cost £2000) • Intervention B: allows an average patient to live for 3 extra years at a utility 0.6 = 1.8 QALYs (cost £18 000). The health economists conclude that treatment B provides an extra 0.4 QALYs at an extra cost of £16 000, meaning that the ICER = £40 000/QALY. They recommend that the new treatment should not be funded on the basis that their threshold for cost acceptability is £30 000/QALY. (ICER = incremental cost-effectiveness ratio; QALY = quality-adjusted life year)

Excretion Drugs that depend on renal excretion for elimination (e.g. digoxin, aminoglycoside antibiotics) should be avoided in patients with impaired renal function if suitable alternatives exist. Efficacy Prescribers normally choose drugs with the greatest efficacy in achieving the goals of therapy (e.g. proton pump inhibitors rather than H2-receptor antagonists). It may be appropriate, however, to compromise on efficacy if other drugs are more convenient, safer to use or less expensive. Avoiding adverse effects Prescribers should be wary of choosing drugs that are more likely to cause adverse effects (e.g. cephalosporins rather than alternatives for patients allergic to penicillin) or worsen coexisting conditions (e.g. β-blockers as treatment for angina in patients with asthma). Features of the disease This is most obvious when choosing antibiotic therapy, which should be based on the known or suspected sensitivity of the infective organism (p. 116). Severity of disease The choice of drug should be appropriate to disease severity (e.g. paracetamol for mild pain, morphine for severe pain). Coexisting disease This may be either an indication or a contraindication to therapy. Hypertensive patients might be prescribed a β-blocker if they also have left ventricular impairment but not if they have asthma. Avoiding adverse drug interactions Prescribers should avoid giving combinations of drugs that might interact, either directly or indirectly (see Box 2.11). Patient adherence to therapy Prescribers should choose drugs with a simple dosing schedule or easier administration (e.g. the ACE inhibitor lisinopril once daily rather than captopril 3 times daily for hypertension). Cost Prescribers should choose the cheaper drug (e.g. a generic or biosimilar) if two drugs are of equal efficacy and safety. Even if cost is not a concern for the individual patient, it is important to remember that unnecessary expenditure will ultimately limit choices for other prescribers and patients. Sometimes a more costly drug may be appropriate (e.g. if it yields improved adherence). Genetic factors There are already a small number of examples where genotype influences the choice of drug therapy (see Box 2.5). Choosing a dosage regimen Prescribers have to choose a dose, route and frequency of administration (dosage regimen) to achieve a steady-state drug concentration that provides sufficient exposure of the target tissue without producing toxic effects. Manufacturers draw up dosage recommendations based on average observations in many patients but the optimal regimen that will maximise the benefit to harm ratio for an individual patient is never certain. Making a diagnosis Ideally, prescribing should be based on a confirmed diagnosis but, in reality, many prescriptions are based on the balance of probability, taking into account the differential diagnosis (e.g. proton pump inhibitors for post-prandial retrosternal discomfort). Establishing the therapeutic goal The goals of treatment are usually clear, particularly when relieving symptoms (e.g. pain, nausea, constipation). Sometimes the goal is less obvious to the patient, especially when aiming to prevent future events (e.g. ACE inhibitors to prevent hospitalisation and extend life in chronic heart failure). Prescribers should be clear about the therapeutic goal against which they will judge success or failure of

treatment. It is also important to establish that the value placed on this goal by the prescriber is shared by the patient (concordance). Choosing the therapeutic approach For many clinical problems, drug therapy is not absolutely mandated. Having taken the potential benefits and harms into account, prescribers must consider whether drug therapy is in the patient's interest and is preferred to no treatment or one of a range of alternatives (e.g. physiotherapy, psychotherapy, surgery). Assessing the balance of benefit and harm is often complicated and depends on various features associated with the patient, disease and drug (Box 2.17). Choosing a drug For most common clinical indications (e.g. type 2 diabetes, depression), more than one drug is available, often from more than one drug class. Although prescribers often have guidance about which represents the rational choice for the average patient, they still need to consider whether this is the optimal choice for the individual patient. Certain factors may influence the choice of drug: Absorption Patients may find some formulations easier to swallow than others or may be vomiting and require a drug available for parenteral administration. Distribution Distribution of a drug to a particular tissue sometimes dictates choice (e.g. tetracyclines and rifampicin are concentrated in the bile, and lincomycin and clindamycin in bones). Metabolism Drugs that are extensively metabolised should be avoided in severe liver disease (e.g. opioid analgesics). 2.17 Factors to consider when balancing benefits and harms of drug therapy • Seriousness of the disease or symptom • Efficacy of the drug • Seriousness of potential adverse effects • Likelihood of adverse effects • Efficacy of alternative drugs or non-drug therapies • Safety of alternative drugs or non-drug therapies

 Duration Some drugs require a single dose (e.g. thrombolysis post myocardial infarction), while for others the duration of the course of treatment is certain at the outset (e.g. antibiotics). For most, the duration will be largely at the prescriber's discretion and will depend on response and disease progression (e.g. analgesics, antidepressants). For many, the treatment will be long-term (e.g. insulin, antihypertensives, levothyroxine). Involving the patient Patients should, whenever possible, be engaged in making choices about drug therapy. Their beliefs and expectations affect the goals of therapy and help in judging the acceptable benefit/ harm balance when selecting treatments. Very often, patients may wish to defer to the professional expertise of the prescriber. Nevertheless, they play key roles in adherence to therapy and in monitoring treatment, not least by providing early warning of adverse events. It is important for them to be provided with the necessary information to understand the choice that has been made, what to expect from the treatment, and any measurements that must be undertaken (Box 2.20). A major drive to include patients has been the recognition that up to half of the drug doses for chronic preventative therapy are not taken. This is often termed 'non-compliance' but is more appropriately called 'non-adherence', to reflect a less paternalistic view of the doctor–patient relationship; it may or may not be intentional. Non-adherence to the dose regimen reduces the likelihood of benefits to the patient and can be costly in terms Rational prescribing involves treating each prescription as an experiment and gathering sufficient information to amend it if necessary. There are some general principles that should be followed, as described below. Dose titration Prescribers should generally start with a low dose and titrate this slowly upwards as necessary. This cautious approach is particularly important if the patient is likely to be more sensitive to adverse pharmacodynamic effects (e.g. delirium or postural hypotension in the elderly) or have altered pharmacokinetic handling (e.g. renal or hepatic impairment), and when using drugs with a low therapeutic index (e.g. benzodiazepines, lithium, digoxin). There are some exceptions, however. Some drugs must achieve therapeutic concentration quickly because of the

clinical circumstance (e.g. antibiotics, glucocorticoids, carbimazole). When early effect is important but there may be a delay in achieving steady state because of a drug's long half-life (e.g. digoxin, warfarin, amiodarone), an initial loading dose is given prior to establishing the appropriate maintenance dose (see Fig. 2.4). If adverse effects occur, the dose should be reduced or an alternative drug prescribed; in some cases, a lower dose may suffice if it can be combined with another synergistic drug (e.g. the immunosuppressant azathioprine reduces glucocorticoid requirements in patients with inflammatory disease). It is important to remember that the shape of the dose–response curve (see Fig. 2.2) means that higher doses may produce little added therapeutic effect and might increase the chances of toxicity. Route There are many reasons for choosing a particular route of administration (Box 2.18). Frequency Frequency of doses is usually dictated by a manufacturer's recommendation. Less frequent doses are more convenient for patients but result in greater fluctuation between peaks and troughs in drug concentration (see Fig. 2.4). This is relevant if the peaks are associated with adverse effects (e.g. dizziness with antihypertensives) or the troughs are associated with troublesome loss of effect (e.g. anti-Parkinsonian drugs). These problems can be tackled either by splitting the dose or by employing a modified-release formulation, if available. Timing For many drugs the time of administration is unimportant. There are occasionally pharmacokinetic or therapeutic reasons, however, for giving drugs at particular times (Box 2.19). Formulation For some drugs there is a choice of formulation, some for use by different routes. Some are easier to ingest, particularly by children (e.g. elixirs). The formulation is important when writing repeat prescriptions for drugs with a low therapeutic index that come in different formulations (e.g. lithium, phenytoin, theophylline). Even if the prescribed dose remains constant, an alternative formulation may differ in its absorption and bioavailability, and hence plasma drug concentration. These are examples of the small number of drugs that should be prescribed by specific brand name rather than 'generic' international non-proprietary name (INN). 2.18 Factors influencing the route of drug administration Reason Example Only one route possible Dobutamine (IV) Gliclazide (oral) Patient adherence Phenothiazines and thioxanthenes (2 weekly IM depot injections rather than daily tablets, in schizophrenia) Poor absorption Furosemide (IV rather than oral, in severe heart failure) Rapid action Haloperidol (IM rather than oral, in acute behavioural disturbance) Vomiting Phenothiazines (PR or buccal rather than oral, in nausea) Avoidance of first-pass metabolism Glyceryl trinitrate (SL, in angina pectoris) Certainty of effect Amoxicillin (IV rather than oral, in acute chest infection) Direct access to the site of action (avoiding unnecessary systemic exposure) Bronchodilators (INH rather than oral, in asthma) Local application of drugs to skin, eyes etc. Ease of access Diazepam (PR, if IV access is difficult in status epilepticus) Adrenaline (epinephrine) (IM, if IV access is difficult in acute anaphylaxis) Comfort Morphine (SC rather than IV in terminal care) (IM = intramuscular; INH = by inhalation; IV = intravenous; PR = per rectum; SC = subcutaneous; SL = sublingual)

Stopping drug therapy It is also important to review long-term treatment at regular intervals to assess whether continued treatment is required. Elderly patients are keen to reduce their medication burden and are often prepared to compromise on the original goals of long-term preventative therapy to achieve this. Prescribing in special circumstances Prescribing for patients with renal disease Patients with renal impairment are readily identified by having a low estimated glomerular filtration rate (eGFR < 60 mL/min) based on their serum creatinine, age, sex and ethnic group (p. 386). This group includes a large proportion of elderly patients. If a drug (or its active

metabolites) is eliminated predominantly by the kidneys, it will tend to accumulate and so the maintenance dose must be reduced. For some drugs, renal impairment makes patients more sensitive to their adverse pharmacodynamic effects. of wasted medicines and unnecessary health-care episodes. An important reason may be lack of concordance with the prescriber about the goals of treatment. A more open and shared decisionmaking process might resolve any misunderstandings at the outset and foster improved adherence, as well as improved satisfaction with health-care services and confidence in prescribers. Fully engaging patients in shared decision-making is sometimes constrained by various factors, such as limited consultation time and challenges in communicating complex numerical data. Writing the prescription The culmination of the planning described above is writing an accurate and legible prescription so that the drug will be dispensed and administered as planned (see ‘Writing prescriptions’ below). Monitoring treatment effects Rational prescribing involves monitoring for the beneficial and adverse effects of treatment so that the balance remains in favour of a positive outcome (see ‘Monitoring drug therapy’ below). 2.19 Factors influencing the timing of drug therapy Drug Recommended timing Reasons Diuretics (e.g. furosemide) Once in the morning Night-time diuresis undesirable Statins (e.g. simvastatin) Once at night HMG CoA reductase activity is greater at night Antidepressants (e.g. amitriptyline) Once at night Allows adverse effects to occur during sleep Salbutamol Before exercise Reduces symptoms in exercise-induced asthma Glyceryl trinitrate Paracetamol When required Relief of acute symptoms only Regular nitrate therapy (e.g. isosorbide mononitrate) Eccentric dosing regimen (e.g. twice daily at 8 a.m. and 2 p.m.) Reduces development of nitrate tolerance by allowing drug-free period each night Aspirin With food Minimises gastrotoxic effects Alendronate Once in the morning before breakfast, sitting upright Minimises risk of oesophageal irritation Tetracyclines 2 hours before or after food or antacids Divalent and trivalent cations chelate tetracyclines, preventing absorption Hypnotics (e.g. temazepam) Once at night Maximises therapeutic effect and minimises daytime sedation Antihypertensive drugs (e.g. amlodipine) Once in the morning Blood pressure is higher during the daytime (HMG CoA = 3-hydroxy-3-methylglutaryl-coenzyme A) 2.20 What patients need to know about their medicines* Knowledge Comment The reason for taking the medicine How the medicine works Reinforces the goals of therapy How to take the medicine May be important for the effectiveness (e.g. inhaled salbutamol in asthma) and safety (e.g. alendronate for osteoporosis) of treatment What benefits to expect May help to support adherence or prompt review because of treatment failure What adverse effects might occur Discuss common and mild effects that may be transient and might not require discontinuation Mention rare but serious effects that might influence the patient’s consent Precautions that improve safety Explain symptoms to report that might allow serious adverse effects to be averted, monitoring that will be required and potentially important drug–drug interactions When to return for review This will be important to enable monitoring *Many medicines are provided with patient information leaflets, which the patient should be encouraged to read.

 Examples of drugs that require extra caution in patients with renal disease are listed in Box 2.21. Prescribing for patients with hepatic disease The liver has a large capacity for drug metabolism and hepatic insufficiency has to be advanced before drug dosages need to be modified. Patients who may have impaired metabolism include those with jaundice, ascites, hypoalbuminaemia, malnutrition or encephalopathy. Hepatic drug clearance may also be reduced in acute hepatitis, in hepatic congestion due to cardiac failure, and in the presence of intrahepatic arteriovenous shunting (e.g. in hepatic cirrhosis). There are no good tests of hepatic drugmetabolising capacity or of biliary excretion, so dosage should be guided

by the therapeutic response and careful monitoring for adverse effects. The presence of liver disease also increases the susceptibility to adverse pharmacological effects of drugs. Some drugs that require extra caution in patients with hepatic disease are listed in Box 2.21. Prescribing for elderly patients The issues around prescribing in the elderly are discussed in Box 2.22. Prescribing for women who are pregnant or breastfeeding Prescribing in pregnancy should be avoided if possible to minimise the risk of adverse effects in the fetus. Drug therapy in pregnancy may, however, be required either for a pre-existing problem (e.g. epilepsy, asthma, hypothyroidism) or for problems that arise during pregnancy (e.g. morning sickness, anaemia, prevention of neural tube defects, gestational diabetes, hypertension). About 35% of women take drug therapy at least once during 2.23 Prescribing in pregnancy • Teratogenesis: a potential risk, especially when drugs are taken between 2 and 8 weeks of gestation (4–10 weeks from last menstrual period). Common teratogens include retinoids (e.g. isotretinoin), cytotoxic drugs, angiotensin-converting enzyme inhibitors, antiepileptics and warfarin. If there is inadvertent exposure, then the timing of conception should be established, counselling given and investigations undertaken for fetal abnormalities. • Adverse fetal effects in late gestation: e.g. tetracyclines may stain growing teeth and bones; sulphonamides displace fetal bilirubin from plasma proteins, potentially causing kernicterus; opioids given during delivery may be associated with respiratory depression in the neonate. • Altered maternal pharmacokinetics: extracellular fluid volume and Vd increase. Plasma albumin falls but other binding globulins (e.g. for thyroid and steroid hormones) increase. Glomerular filtration increases by approximately 70%, enhancing renal clearance. Placental metabolism contributes to increased clearance, e.g. of levothyroxine and glucocorticoids. The overall effect is a fall in plasma levels of many drugs. • In practice: Avoid any drugs unless the risk:benefit analysis is in favour of treating (usually the mother). Use drugs for which there is some record of safety in humans. Use the lowest dose for the shortest time possible. Choose the least harmful drug if alternatives are available. 2.22 Prescribing in old age • Reduced drug elimination: partly due to impaired renal function. • Increased sensitivity to drug effects: notably in the brain (leading to sedation or delirium) and as a result of comorbidities. • More drug interactions: largely as a result of polypharmacy. • Lower starting doses and slower dose titration: often required, with careful monitoring of drug effects. • Drug adherence: may be poor because of cognitive impairment, difficulty swallowing (dry mouth) and complex polypharmacy regimens. Supplying medicines in pill organisers (e.g. dosette boxes or calendar blister packs), providing automatic reminders, and regularly reviewing and simplifying the drug regimen can help. • Some drugs that require extra caution, and their mechanisms: Digoxin: increased sensitivity of Na+/K+ pump; hypokalaemia due to diuretics; renal impairment favours accumulation → increased risk of toxicity. Antihypertensive drugs: reduced baroreceptor function → increased risk of postural hypotension. Antidepressants, hypnotics, sedatives, tranquillisers: increased sensitivity of the brain; reduced metabolism → increased risk of toxicity. Warfarin: increased tendency to falls and injury and to bleeding from intra- and extracranial sites; increased sensitivity to inhibition of clotting factor synthesis → increased risk of bleeding. Clomethiazole, lidocaine, nifedipine, phenobarbital, propranolol, theophylline: metabolism reduced → increased risk of toxicity. Non-steroidal anti-inflammatory drugs: poor renal function → increased risk of renal impairment; susceptibility to gastrotoxicity → increased risk of upper gastrointestinal bleeding. 2.21 Some drugs that require extra caution in patients with renal or hepatic disease Kidney disease Liver disease Pharmacodynamic effects enhanced ACE inhibitors and ARBs (renal impairment, hyperkalaemia) Metformin (lactic acidosis) Spironolactone (hyperkalaemia) NSAIDs (impaired renal function) Sulphonylureas (hypoglycaemia) Insulin (hypoglycaemia) Warfarin (increased anticoagulation

because of reduced clotting factor synthesis) Metformin (lactic acidosis) Chloramphenicol (bone marrow suppression) NSAIDs (gastrointestinal bleeding, fluid retention) Sulphonylureas (hypoglycaemia) Benzodiazepines (coma) Pharmacokinetic handling altered (reduced clearance) Aminoglycosides (e.g. gentamicin) Vancomycin Digoxin Lithium Other antibiotics (e.g. ciprofloxacin) Atenolol Allopurinol Cephalosporins Methotrexate Opioids (e.g. morphine) Phenytoin Rifampicin Propranolol Warfarin Diazepam Lidocaine Opioids (e.g. morphine) (ACE = angiotensin-converting enzyme; ARB = angiotensin receptor blocker; NSAID = non-steroidal anti-inflammatory drug) pregnancy and 6% take drug therapy during the first trimester (excluding iron, folic acid and vitamins). The most commonly used drugs are simple analgesics, antibacterial drugs and antacids. Some considerations when prescribing in pregnancy are listed in Box 2.23.

Hospital discharge (‘to take out’) medicines Most patients will be prescribed a short course of their medicines at discharge. This prescription is particularly important because it usually informs future therapy at the point of transfer of prescribing responsibility to primary care. Great care is required to ensure that this list is accurate. It is particularly important to ensure that any hospital medicines that should be stopped are not included and that those intended to be administered for a short duration only are clearly identified. It is also important for any significant ADRs experienced in hospital to be recorded and any specific monitoring or review identified. Prescribing in primary care Most of the considerations above are equally applicable to primary care (GP) prescriptions. In many health-care systems, community prescribing is electronic, making issues of legibility irrelevant and often providing basic decision support to limit the range of doses that can be written and highlight potential drug interactions. Important additional issues more relevant to GP prescribing are: • Formulation. The prescription needs to carry information about the formulation for the dispensing pharmacist (e.g. tablets or oral suspension). • Amount to be supplied. In the hospital the pharmacist will organise this. Elsewhere it must be specified either as the precise number of tablets or as the duration of treatment. Creams and ointments should be specified in grams and lotions in mL. • Controlled drugs. Prescriptions for ‘controlled’ drugs (e.g. opioid analgesics, with potential for drug abuse) are subject to additional legal requirements. In the UK, they Drugs that are excreted in breast milk may cause adverse effects in the baby. Prescribers should always consult the summary of product characteristics for each drug or a reliable formulary when treating a pregnant woman or breastfeeding mother. Writing prescriptions A prescription is a means by which a prescriber communicates the intended plan of treatment to the pharmacist who dispenses a medicine and to a nurse or patient who administers it. It should be precise, accurate, clear and legible. The two main kinds of prescription are those written, dispensed and administered in hospital and those written in primary care (in the UK by a GP), dispensed at a community pharmacy and self-administered by the patient. The information supplied must include: • the date • the identification details of the patient • the name of the drug • the formulation • the dose • the frequency of administration • the route and method of administration • the amount to be supplied (primary care only) • instructions for labelling (primary care only) • the prescriber’s signature. Prescribing in hospital Although GP prescribing is increasingly electronic, most hospital prescribing continues to be based around the prescription and administration record (the ‘drug chart’) (Fig. 2.7). A variety of charts are in use and prescribers must familiarise themselves with the local version. Most contain the following sections: • Basic patient information: will usually include name, age, date of birth, hospital number and address. These details are often ‘filled in’ using a sticky

addressograph label but this increases the risk of serious error. • Previous adverse reactions/allergies: communicates important patient safety information based on a careful drug history and/or the medical record. • Other medicines charts: notes any other hospital prescription documents that contain current prescriptions being received by the patient (e.g. anticoagulants, insulin, oxygen, fluids). • Once-only medications: for prescribing medicines to be used infrequently, such as single-dose prophylactic antibiotics and other pre-operative medications. • Regular medications: for prescribing medicines to be taken for a number of days or continuously, such as a course of antibiotics, antihypertensive drugs and so on. • 'As required' medications: for prescribing for symptomatic relief, usually to be administered at the discretion of the nursing staff (e.g. antiemetics, analgesics). Prescribers should be aware of the risks of prescription error (Box 2.24 and see Box 2.13), ensure they have considered the rational basis for their prescribing decision described above, and then follow the guidance illustrated in Figure 2.7 in order to write the prescription. It is a basic principle that a prescription will be followed by a judgement as to its success or failure and any appropriate changes made (e.g. altered dosage, discontinuation or substitution). 2.24 High-risk prescribing moments • Trying to amend an active prescription (e.g. altering the dose/ timing) – always avoid and start again • Writing up drugs in the immediate presence of more than one prescription chart or set of notes – avoid • Allowing one's attention to be diverted in the middle of completing a prescription – avoid • Prescribing 'high-risk' drugs (e.g. anticoagulants, opioids, insulin, sedatives) – ask for help if necessary • Prescribing parenteral drugs – take care • Rushing prescribing (e.g. in the midst of a busy ward round) – avoid • Prescribing unfamiliar drugs – consult the formulary and ask for help if necessary • Transcribing multiple prescriptions from an expired chart to a new one – take care to review the rationale for each medicine • Writing prescriptions based on information from another source such as a referral letter (the list may contain errors and some of the medicines may be the cause of the patient's illness) – review the justification for each as if it is a new prescription • Writing up 'to take out' drugs (because these will become the patient's regular medication for the immediate future) – take care and seek advice if necessary • Calculating drug doses – ask a colleague to perform an independent calculation or use approved electronic dose calculators • Prescribing sound-alike or look-alike drugs (e.g. chlorphenamine and chlorpromazine) – take care

34 • CLINICAL THERAPEUTICS AND GOOD PRESCRIBING Fig. 2.7 Example of a hospital prescription and administration record ('drug chart'). A Front page. The correct identification of the patient is critical to reducing the risk of an administration error. This page also clearly identifies other prescriptions charts in use and previous adverse reactions to drugs to minimise the risk of repeated exposure. Note also the codes employed by the nursing staff to indicate reasons why drugs may not have been administered. The patient's name and date of birth should be written on each page of the chart. The patient's weight and height may be required to calculate safe doses for many drugs with narrow therapeutic indices. B 'Once-only medicines'. This area is used for prescribing medicines that are unlikely to be repeated on a regular basis. Note that the prescriber has written the names of all drugs legibly in block capitals. The generic international non-proprietary name (INN) should be used in preference to the brand name (e.g. write 'SIMVASTATIN', not 'ZOCOR'). The only exceptions are when variation occurs in the properties of alternative branded formulations (e.g. modified-release preparations of drugs such as lithium, theophylline, phenytoin and nifedipine) or when the drug is a combination product with no generic name (e.g. Kliovance). The only acceptable abbreviations for drug dose units are 'g' and 'mg'. 'Units' (e.g. of insulin or heparin) and 'micrograms' must always be written in full, never as 'U' or 'µg' (nor 'mcg', nor 'ug').

For liquid preparations write the dose in mg; ‘mL’ can be written only for a combination product (e.g. Gaviscon liquid) or if the strength is not expressed in weight (e.g. adrenaline (epinephrine) 1 in 1000). Use numbers/figures (e.g. 1 or ‘one’) to denote use of a sachet/enema but avoid prescribing numbers of tablets without specifying their strength. Always include the dose of inhaled drugs in addition to stating numbers of ‘puffs’, as strengths can vary. Widely accepted abbreviations for route of administration are: intravenous – ‘IV’; intramuscular – ‘IM’; subcutaneous – ‘SC’; sublingual – ‘SL’; per rectum – ‘PR’; per vaginam – ‘PV’; nasogastric – ‘NG’; inhaled – ‘INH’; and topical – ‘TOP’. ‘ORAL’ is preferred to per oram – ‘PO’. Care should be taken in specifying ‘RIGHT’ or ‘LEFT’ for eye and ear drops. The prescriber should sign and print their name clearly so that they can be identified by colleagues. The prescription should be dated and have an administration time. The nurse who administered the prescription has signed to confirm that the dose has been administered. OTHER MEDICINES CHARTS CODES FOR NON-ADMINISTRATION OF PRESCRIBED MEDICINE PREVIOUS ADVERSE REACTIONS (This must be completed before prescribing on this chart) Hospital/Ward: Consultant: Name of patient: Hospital number: (Attach printed label here) D.O.B.: Weight: Date Date Time Medicine (approved name) Dose Route Time given Given by Prescriber – sign and print If a dose is not administered as prescribed, intial and enter a code in the column with a circle drawn round the code according to the reason as shown below. Inform the responsible doctor of the appropriate timescale.

1. Patient refuses
2. Patient not present
3. Medicines not available – CHECK ORDERED
4. Asleep/drowsy
5. Administration route not available – CHECK FOR ALTERNATIVE
6. Vomiting/nausea
7. Time varied on doctor’s instructions
8. Once-only/as-required medicine given
9. Dose withheld on doctor’s instructions
10. Possible adverse reaction/side-effect Type of chart Medicine Description of reaction Completed by Date Height: If rewritten, date: DISCHARGE PRESCRIPTION PRESCRIPTION AND ADMINISTRATION RECORD Standard Chart ONCE-ONLY MEDICINES Date completed:– Completed by:– A B must contain the address of the patient and prescriber (not necessary on most hospital forms), the form and the strength of the preparation, and the total quantity of the preparation/number of dose units in both words and figures. • ‘Repeat prescriptions’. A large proportion of GP prescribing involves ‘repeat prescriptions’ for chronic medication. These are often generated automatically, although the prescriber remains responsible for regular review and for ensuring that the benefit-to-harm ratio remains favourable. Monitoring drug therapy Prescribers should measure the effects of the drug, both beneficial and harmful, to inform decisions about dose titration (up or down), discontinuation or substitution of treatment. Monitoring can be achieved subjectively by asking the patient about symptoms or, more objectively, by measuring a clinical effect. Alternatively, if the pharmacodynamic effects of the drug are difficult to assess, the plasma drug concentration may be measured, on the basis that it will be closely related to the effect of the drug (see Fig. 2.2).

REGULAR MEDICINES AS-REQUIRED THERAPY C D Drug (approved name) Dose Date Time

Prescriber–sign and print Notes Start date Pharmacy Route Drug (approved name) Dose Prescriber–sign and print Notes Start date Pharmacy Route Drug (approved name) Dose and frequency Prescriber–sign and print Start date Indication/notes Pharmacy Route Date Time Dose Initials Date Time Dose Initials Drug (approved name) Dose and frequency Prescriber–sign and print Start date Indication/notes Pharmacy Route Date Time Dose Initials Date Time Dose Initials Drug (approved name) Dose Prescriber–sign and print Notes Start date Pharmacy Route C 'Regular medicines'. This area is used for prescribing medicines that are going to be given regularly. In addition to the name, dose and route, a frequency of administration is required for each medicine. Widely accepted Latin abbreviations for dose frequency are: once daily – 'OD'; twice daily – 'BD'; 3 times daily – 'TDS'; 4 times daily – 'QDS'; as required – 'PRN'; in the morning – 'OM' (omni mane); at night – 'ON' (omni nocte); and immediately – 'stat'. The hospital chart usually requires specific times to be identified for regular medicines that coincide with nursing drug rounds and these can be circled. If treatment is for a known time period, cross off subsequent days when the medicine is not required. The 'notes' box can be used to communicate additional important information (e.g. whether a medicine should be taken with food, type of inhaler device used, and anything else that the drug dispenser should know). State here the times for peak/trough plasma levels for drugs requiring therapeutic monitoring. Prescriptions should be discontinued by drawing a vertical line at the point of discontinuation, horizontal lines through the remaining days on the chart, and diagonal lines through the drug details and administration boxes. This action should be signed and dated and a supplementary note written to explain it (e.g. describing any adverse effect). In this example, amlodipine has been discontinued because of ankle oedema. There is room for the ward pharmacist to sign to indicate that the prescription has been reviewed and that a supply of the medicine is available. The administration boxes allow the nurse to sign to confirm that the dose has been given. Note that these boxes also allow for recording of reasons for non-administration (in this example '2' indicates that the patient was not present on the ward at the time) and the prevention of future administration by placing an 'X' in the box. D 'As-required medicines'. These prescriptions leave the administration of the drug to the discretion of the nursing staff. The prescription must describe clearly the indication, frequency, minimal time interval between doses, and maximum dose in any 24-hour period (in this case, the maximum daily dose of paracetamol is 4 g). Fig. 2.7, cont'd Clinical and surrogate endpoints Ideally, clinical endpoints are measured directly and the drug dosage titrated to achieve the therapeutic goal and avoid toxicity (e.g. control of ventricular rate in a patient with atrial fibrillation). Sometimes this is impractical because the clinical endpoint is a future event (e.g. prevention of myocardial infarction by statins or resolution of a chest infection with antibiotics); in these circumstances, it may be possible to select a 'surrogate' endpoint that will predict success or failure. This may be an intermediate step in the pathophysiological process (e.g. serum cholesterol as a surrogate for risk of myocardial infarction) or a

 Interpreting the result A target range is provided for many drugs, based on average thresholds for therapeutic benefit and toxicity. Inter-individual variability means that these can be used only as a guide. For instance, in a patient who describes symptoms that could be consistent with toxicity but has a drug concentration in the top half of the target range, toxic effects should still be suspected. Another important consideration is that some drugs are heavily protein-bound (e.g. phenytoin) but only the unbound drug is

pharmacologically active. Patients with hypoalbuminaemia may therefore have a therapeutic or even toxic concentration of unbound drug, despite a low ‘total’ concentration. Further information Websites bnf.org The British National Formulary (BNF) is a key reference resource for UK NHS prescribers, with a list of licensed drugs, chapters on prescribing in renal failure, liver disease, pregnancy and during breastfeeding, and appendices on drug interactions. cochrane.org The Cochrane Collaboration is a leading international body that provides evidence-based reviews (around 7000 so far). evidence.nhs.uk NHS Evidence provides a wide range of health information relevant to delivering quality patient care. icp.org.nz The Interactive Clinical Pharmacology site is designed to increase understanding of principles in clinical pharmacology. medicines.org.uk/emc/ The electronic Medicines Compendium (eMC) contains up-to-date, easily accessible information about medicines licensed by the UK Medicines and Healthcare Products Regulatory Agency (MHRA) and the European Medicines Agency (EMA). nice.org.uk The UK National Institute for Health and Care Excellence makes recommendations to the UK NHS on new and existing medicines, treatments and procedures. who.int/medicines/en/ The World Health Organisation Essential Medicines and Pharmaceutical Policies. measurement that follows the pathophysiology, even if it is not a key factor in its progression (e.g. serum C-reactive protein as a surrogate for resolution of inflammation in chest infection). Such surrogates are sometimes termed ‘biomarkers’. Plasma drug concentration The following criteria must be met to justify routine monitoring by plasma drug concentration: • Clinical endpoints and other pharmacodynamic (surrogate) effects are difficult to monitor. • The relationship between plasma concentration and clinical effects is predictable. • The therapeutic index is low. For drugs with a high therapeutic index, any variability in plasma concentrations is likely to be irrelevant clinically. Some examples of drugs that fulfil these criteria are listed in Box 2.25. Measurement of plasma concentration may be useful in planning adjustments of drug dose and frequency of administration; to explain an inadequate therapeutic response (by identifying subtherapeutic concentration or incomplete adherence); to establish whether a suspected ADR is likely to be caused by the drug; and to assess and avoid potential drug interactions. Timing of samples in relation to doses The concentration of drug rises and falls during the dosage interval (see Fig. 2.4B). Measurements made during the initial absorption and distribution phases are unpredictable because of the rapidly changing concentration, so samples are usually taken at the end of the dosage interval (a ‘trough’ or ‘pre-dose’ concentration). This measurement is normally made in steady state, which usually takes five half-lives to achieve after the drug is introduced or the dose changed (unless a loading dose has been given). 2.25 Drugs commonly monitored by plasma drug concentration Drug Half-life (hrs)* Comment Digoxin

Steady state takes several days to achieve. Samples should be taken 6 hrs post dose. Measurement is useful to confirm the clinical impression of toxicity or non-adherence but clinical effectiveness is better assessed by ventricular heart rate. Risk of toxicity increases progressively at concentrations > 1.5 µg/L, and is likely at concentrations > 3.0 µg/L (toxicity is more likely in the presence of hypokalaemia) Gentamicin

Measure pre-dose trough concentration (should be < 1 µg/mL) to ensure that accumulation (and the risk of nephrotoxicity and ototoxicity) is avoided; see Fig. 6.18 (p. 122) Levothyroxine

> “ 120 Steady state may take up to 6 weeks to achieve (p. 640) Lithium

Steady state takes several days to achieve. Samples should be taken 12 hrs post dose. Target range 0.4–1 mmol/L Phenytoin

Measure pre-dose trough concentration (should be 10–20 mg/L) to ensure that accumulation is avoided. Good correlation between concentration and toxicity. Concentration may be misleading in the presence of hypoalbuminaemia Theophylline (oral)

Steady state takes 2–3 days to achieve. Samples should be taken 6 hrs post dose. Target concentration is 10–20 mg/L but its relationship with bronchodilator effect and adverse effects is variable Vancomycin

Measure pre-dose trough concentration (should be 10–15 mg/L) to ensure clinical efficacy and that accumulation and the risk of nephrotoxicity are avoided (p. 123) *Half-lives vary considerably with different formulations and between patients.

03-3 Clinical genetics

3 Clinical genetics

Clinical genetics K Tatton-Brown DR FitzPatrick The fundamental principles of genomics 38 The packaging of genes: DNA, chromatin and chromosomes 38 From DNA to protein 38 Non-coding RNA 40 Cell division, differentiation and migration 40 Cell death, apoptosis and senescence 41 Genomics, health and disease 42 Classes of genetic variant 42 Consequences of genomic variation 44 Normal genomic variation 45 Constitutional genetic disease 46 Somatic genetic disease 50 Interrogating the genome: the changing landscape of genomic technologies 51 Looking at chromosomes 51 Looking at genes 52 Genomics and clinical practice 56 Genomics and health care 56 Treatment of genetic disease 58 Ethics in a genomic age 59

38 • CLINICAL GENETICS via the production of messenger ribonucleic acid (mRNA) to the production of proteins. The human genome contains over 20 000 genes, although many of these are inactive or silenced in different cell types, reflecting the variable gene expression responsible for cell-specific characteristics. The central dogma is the pathway describing the basic steps of protein production: transcription, splicing, translation and protein modification (Fig. 3.2). Although this is now recognised as an over-simplification (contrary to this linear relationship, a single gene will often encode many different proteins), it remains a useful starting point to explore protein production. Transcription: DNA to messenger RNA Transcription describes the production of ribonucleic acid (RNA) from the DNA template. For transcription to commence, an enzyme called RNA polymerase binds to a segment of DNA at the start of the gene: the promoter. Once bound, RNA polymerase moves along one strand of DNA, producing an RNA molecule complementary to the DNA template. In protein-coding genes this is known as messenger RNA (mRNA). A DNA sequence close to the end of the gene, called the polyadenylation signal, acts as a signal for termination of the RNA transcript (Fig. 3.3). We have entered a genomic era. Powerful new technologies are driving forward transformational change in health care. Genetic sequencing has evolved from the targeted sequencing of a single gene to the parallel sequencing of multiple genes. In addition to improving the chances of identifying a genetic cause of rare diseases, these technologies are increasingly directing therapies and, in the future, are likely to be used in the diagnosis and prevention of common diseases such as diabetes. In this chapter we explore the fundamentals of genomics, the basic principles underlying these new genomic technologies and how the data generated can be applied safely for patient benefit. We will review the use of genomic technology across a breadth of medical specialties, including obstetrics, paediatrics, oncology and infectious disease, and consider how health care is likely to be transformed by technology over the coming decade. Finally, we will consider the ethical impact that these technologies are likely to have, both for the individual and for their wider family. The fundamental principles of genomics The packaging of genes: DNA, chromatin and chromosomes Genes are functional units encoded in double-stranded deoxyribonucleic acid (DNA), packaged as

chromosomes and located in the nucleus of the cell: a membrane-bound compartment found in all cells except erythrocytes and platelets (Fig. 3.1). DNA consists of a linear sequence of just four bases: adenine (A,) cytosine (C), thymine (T) and guanine (G.) It forms a ‘double helix’, a twisted ladder-like structure formed from two complementary strands of DNA joined by hydrogen bonds between bases on the opposite strand that can form only between a C and a G base and an A and a T base. It is this feature of DNA that enables faithful DNA replication and is the basis for many of the technologies designed to interrogate the genome: when the DNA double helix ‘unzips’, one strand can act as a template for the creation of an identical strand. A single copy of the human genome comprises approximately 3.1 billion base pairs of DNA, wound around proteins called histones. The unit consisting of 147 base pairs wrapped around four different histone proteins is called the nucleosome. Sequences of nucleosomes (resembling a string of beads) are wound and packaged to form chromatin: tightly wound, densely packed chromatin is called heterochromatin and open, less tightly wound chromatin is called euchromatin. The chromatin is finally packaged into the chromosomes. Humans are diploid organisms: the nucleus contains two copies of the genome, visible microscopically as 23 chromosome pairs (known as the karyotype). Chromosomes 1 through to 22 are known as the autosomes and consist of identical chromosomal pairs. The 23rd ‘pair’ of chromosomes are the two sex chromosomes: females have two X chromosomes and males an X and Y chromosome. A normal female karyotype is therefore written as 46,XX and a normal male is 46,XY. From DNA to protein Genes are functional elements on the chromosome that are capable of transmitting information from the DNA template Fig. 3.1 The packaging of DNA, genes and chromosomes. From bottom to top: the double helix and the complementary DNA bases; chromatin; and a normal female chromosome pattern – the karyotype. DNA helix Histones Chromatin Chromosome Normal female karyotype Nucleosome A T G A C G G A T T A C T G C C T A

Fig. 3.2 The central dogma of protein production. Double-stranded DNA as a template for single-stranded RNA, which codes for the production of a peptide chain of amino acids. Each of these chains has an orientation. For DNA and RNA, this is 5′ to 3′. For peptides, this is N-terminus to C-terminus. DNA 5’CGATTC3’ 3’GCTAAG5’ 5’CGAUUC3’ N_ArgPhe_C RNA Protein Transcription Translation Fig. 3.3 RNA synthesis and its translation into protein. Gene transcription involves binding of RNA polymerase II to the promoter of genes being transcribed with other proteins (transcription factors) that regulate the transcription rate. The primary RNA transcript is a copy of the whole gene and includes both introns and exons, but the introns are removed within the nucleus by splicing and the exons are joined to form the messenger RNA (mRNA). Prior to export from the nucleus, a methylated guanosine nucleotide is added to the 5′ end of the RNA (‘cap’) and a string of adenine nucleotides is added to the 3′ (‘polyA tail’). This protects the RNA from degradation and facilitates transport into the cytoplasm. In the cytoplasm, the mRNA binds to ribosomes and forms a template for protein production. (tRNA = transfer RNA; UTR = untranslated region) Protein product N-term C-term cap Messenger RNA (mRNA) Messenger RNA (mRNA) AAAAA tRNAs Ribosome Nuclear membrane Primary RNA transcript Spliceosome RNA export to cytoplasm cap cap AAAAA PolyA tail 3'UTR 5'UTR Nuclear pore AAAAA Splicing Transcription 3' 5' 3' Sense strand Enhancer Transcription factors RNA polymerase II Exon 1 Intron 2 Intron 1 Exon 2 Exon 3 Exon 1 Exon 2 Exon 3 Promoter Nucleolus Nuclear membrane Gene A Gene B Gene C Active gene RNA Nucleus Translation RNA differs from DNA in three main ways: • RNA is single-stranded. • The sugar residue within the nucleotide is ribose, rather than deoxyribose. • It contains uracil (U) in

place of thymine (T). The activity of RNA polymerase is regulated by transcription factors. These proteins bind to specific DNA sequences at the promoter or to enhancer elements that may be many thousands of base pairs away from the promoter; a loop in the chromosomal DNA brings the enhancer close to the promoter, enabling the bound proteins to interact. The human genome encodes more than 1200 different transcription factors. Mutations within transcription factors, promoters and enhancers can cause disease. For example, the blood disorder alpha-thalassaemia is usually caused by gene deletions (see p. 954 and Box 3.4). However, it can also result from a mutation in an enhancer located more than 100 000 base pairs (bp) from the α-globin gene promoter, leading to greatly reduced transcription. Gene activity, or expression, is influenced by a number of complex interacting factors, including the accessibility of the gene promoter to transcription factors. DNA can be modified by the addition of a methyl group to cytosine molecules (methylation). If DNA methylation occurs in promoter regions, transcription is silenced, as methyl cytosines are usually not available for transcription factor binding. A second mechanism determining promoter accessibility is the structural configuration of chromatin. In open chromatin, called euchromatin, gene promoters are accessible to RNA polymerase and transcription factors; therefore it is transcriptionally active. This contrasts with heterochromatin, which is densely packed and transcriptionally silent. The chromatin configuration is determined by modifications (such as methylation or acetylation) of specific amino acid residues of histone protein tails. Modifications of DNA and histones are termed epigenetic (‘epi-’ meaning ‘above’ the genome), as they do not alter the primary sequence of the DNA code but have biological significance in chromosomal function. Abnormal epigenetic changes are increasingly recognised as important events in the progression of cancer, allowing expression of normally silenced genes that result in cancer cell de-differentiation and proliferation. They also afford therapeutic targets. For instance, the histone deacetylase inhibitor vorinostat has been successfully used to treat cutaneous T-cell lymphoma, due to the re-expression of genes that had

 into the cytoplasm or packaged into vesicles for secretion. The clinical importance of post-translational modification of proteins is shown by the severe developmental, neurological, haemostatic and soft tissue abnormalities that are associated with the many different congenital disorders of glycosylation. Post-translational modifications can also be disrupted by the synthesis of proteins with abnormal amino acid sequences. For example, the most common mutation in cystic fibrosis (ΔF508) results in an abnormal protein that cannot be exported from the ER and Golgi (see Box 3.4). Non-coding RNA Approximately 4500 genes in humans encode non-coding RNAs (ncRNA) rather than proteins. There are various categories of ncRNA, including transfer RNA (tRNA), ribosomal RNA (rRNA), ribozymes and microRNA (miRNA). The miRNAs, which number over 1000, have a role in post-translational gene expression: they bind to mRNAs, typically in the 3′UTR, promoting target mRNA degradation and gene silencing. Together, miRNAs affect over half of all human genes and have important roles in normal development, cancer and common degenerative disorders. This is the subject of considerable research interest at present. Cell division, differentiation and migration In normal tissues, molecules such as hormones, growth factors and cytokines provide the signal to activate the cell cycle: a controlled programme of biochemical events that culminates in cell division. In all cells of the body, except the gametes (the sperm and egg cells, also known as the germ line), mitosis completes cell division, resulting in two diploid daughter cells. In contrast, the sperm and eggs cells complete cell division with meiosis, resulting in four haploid daughter cells (Fig. 3.4). The stages of cell division in the non-germ-line, somatic cells are shown below: • Cells not committed to mitosis are said to be in G0. • Cells

committed to mitosis must go through the preparatory phase of interphase consisting of G1, S and G2: • G1 (first gap): synthesis of the cellular components necessary to complete cell division • S (synthesis): DNA replication producing identical copies of each chromosome called the sister chromatids • G2 (second gap): repair of any errors in the replicated DNA before proceeding to mitosis. • Mitosis (M) consists of four phases: • Prophase: the chromosomes condense and become visible, the centrioles move to opposite ends of the cell and the nuclear membrane disappears. • Metaphase: the centrioles complete their migration to opposite ends of the cell and the chromosomes – consisting of two identical sister chromatids – line up at the equator of the cell. • Anaphase: spindle fibres attach to the chromosome and pull the sister chromatids apart. • Telophase: the chromosomes decondense, the nuclear membrane reforms and two daughter cells – each with 46 chromosomes – are formed. The progression from one phase to the next is tightly controlled by cell-cycle checkpoints. For example, the checkpoint between previously been silenced in the tumour. These genes encode transcription factors that promote T-cell differentiation as opposed to proliferation, thereby causing tumour regression. RNA splicing, editing and degradation Transcription produces an RNA molecule that is a copy of the whole gene, termed the primary or nascent transcript. This nascent transcript then undergoes splicing, whereby regions not required to make protein (the intronic regions) are removed while those segments that are necessary for protein production (the exonic regions) are retained and rejoined. Splicing is a highly regulated process that is carried out by a multimeric protein complex called the spliceosome. Following splicing, the mRNA molecule is exported from the nucleus and used as a template for protein synthesis. Many genes produce more than one form of mRNA (and thus protein) by a process termed alternative splicing, in which different combinations of exons are joined together. Different proteins from the same gene can have entirely distinct functions. For example, in thyroid C cells the calcitonin gene produces mRNA encoding the osteoclast inhibitor calcitonin (p. 634), but in neurons the same gene produces an mRNA with a different complement of exons via alternative splicing that encodes a neurotransmitter, calcitonin-gene-related peptide (p. 772). Translation and protein production Following splicing, the segment of mRNA containing the code that directs synthesis of a protein product is called the open reading frame (ORF). The inclusion of a particular amino acid in the protein is specified by a codon composed of three contiguous bases. There are 64 different codons with some redundancy in the system: 61 codons encode one of the 20 amino acids, and the remaining three codons – UAA, UAG and UGA (known as stop codons) – cause termination of the growing polypeptide chain. ORFs in humans most commonly start with the amino acid methionine. All mRNA molecules have domains before and after the ORF called the 5′ untranslated region (UTR) and 3′UTR, respectively. The start of the 5′UTR contains a cap structure that protects mRNA from enzymatic degradation, and other elements within the 5′UTR are required for efficient translation. The 3′UTR also contains elements that regulate efficiency of translation and mRNA stability, including a stretch of adenine bases known as a polyA tail (see Fig. 3.3). The mRNAs then leave the nucleus via nuclear pores and associate with ribosomes, the sites of protein production (see Fig. 3.3). Each ribosome consists of two subunits (40S and 60S), which comprise non-coding rRNA molecules (see Fig. 3.9, p. 50) complexed with proteins. During translation, a different RNA molecule known as transfer RNA (tRNA) binds to the ribosome. The tRNAs deliver amino acids to the ribosome so that the newly synthesised protein can be assembled in a stepwise fashion. Individual tRNA molecules bind a specific amino acid and ‘read’ the mRNA ORF via an ‘anticodon’ of three nucleotides that is complementary to the codon in mRNA (see Fig. 3.3). A proportion of ribosomes is bound to the membrane of the endoplasmic reticulum (ER), a complex tubular structure that surrounds the nucleus. Proteins synthesised on these ribosomes are

translocated into the lumen of the ER, where they undergo folding and processing. From here, the protein may be transferred to the Golgi apparatus, where it undergoes post-translational modifications, such as glycosylation (covalent attachment of sugar moieties), to form the mature protein that can be exported

(prophase, metaphase, anaphase and telophase) but differs from mitosis in the following ways: • It consists of two separate cell divisions known as meiosis I and meiosis II. • It reduces the chromosome number from the diploid to the haploid number via a tetraploid stage, i.e. from 46 to 92 (MI S) to 46 (MI M) to 23 (MII M) chromosomes, so that when a sperm cell fertilises the egg, the resulting zygote will return to a diploid, 46, chromosome complement. This reduction to the haploid number occurs at the end of meiosis II. • The 92 chromosome stage consists of 23 homologous pairs of sister chromatids, which then swap genetic material, a process known as recombination. This occurs at the end of MI prophase and ensures that the chromosome that a parent passes to his or her offspring is a mix of the chromosomes that the parent inherited from his or her own mother and father. The individual steps in meiotic cell division are similar in males and females. However, the timing of the cell divisions is very different. In females, meiosis begins in fetal life but does not complete until after ovulation. A single meiotic cell division can thus take more than 40 years to complete. As women become older, the separation of chromosomes at meiosis II becomes less efficient. That is why the risk of trisomies (p. 44) due to non-disjunction grows greater with increasing maternal age. In males, meiotic division does not begin until puberty and continues throughout life. In the testes, both meiotic divisions are completed in a matter of days. Cell death, apoptosis and senescence With the exception of stem cells, human cells have only a limited capacity for cell division. The Hayflick limit is the number of divisions a cell population can go through in culture before division stops and enters a state known as senescence. This ‘biological clock’ is of great interest in the study of the normal ageing process. Rare human diseases associated with premature ageing, called progeric syndromes, have been very helpful in identifying the importance of DNA repair mechanisms in senescence (p. 1034). For example, in Werner’s syndrome, a DNA helicase (an enzyme that separates the two DNA strands) is mutated, leading to failure of DNA repair and premature ageing. A distinct mechanism of cell death is seen in apoptosis, or programmed cell death. Apoptosis is an active process that occurs in normal tissues and plays an important role in development, tissue remodelling and the immune response. The signal that triggers apoptosis is specific to each tissue or cell type. This signal activates enzymes, called caspases, which actively destroy cellular components, including chromosomal DNA. This degradation results in cell death, but the cellular corpse contains characteristic vesicles called apoptotic bodies. The corpse is then recognised and removed by phagocytic cells of the immune system, such as macrophages, in a manner that does not provoke an inflammatory response. A third mechanism of cell death is necrosis. This is a pathological process in which the cellular environment loses one or more of the components necessary for cell viability. Hypoxia is probably the most common cause of necrosis. Fig. 3.4 Meiosis and gametogenesis: the main chromosomal stages of meiosis in both males and females. A single homologous pair of chromosomes is represented in different colours. The final step is the production of haploid germ cells. Each round of meiosis in the male results in four sperm cells; in the female, however, only one egg cell is produced, as the other divisions are sequestered on the periphery of the mature egg as peripheral polar bodies. Egg Sperm Father Mother Meiotic cell divisions 1st polar bodies 2nd polar body DNA

replication Sister chromatids Homologous pairing Swapping of genetic material between homologues: Recombination Individual chromosome pair (homologues) Non-disjunction of chromosomes is a common error in human meiosis, resulting in trisomy of individual chromosomes or uniparental disomy (both chromosomes from single parent) G2 and mitosis ensures that all damaged DNA is repaired prior to segregation of the chromosomes. Failure of these control processes is a crucial driver in the pathogenesis of cancer, as discussed on page 1316. • Meiosis is a special, gamete-specific, form of cell division (Fig. 3.4). Like mitosis, meiosis consists of four phases

42 • CLINICAL GENETICS mutation (Box 3.1 and Fig. 3.5). If a multiple of three nucleotides is involved, this is in-frame. If an indel change affects one or two nucleotides within the ORF of a protein-coding gene, this can have serious consequences because the triple nucleotide sequence of the codons is disrupted, resulting in a frameshift mutation. The effect on the gene is typically severe because the amino acid sequence is totally disrupted. Fig. 3.5 Different types of mutation affecting coding exons. A Normal sequence. B A synonymous nucleotide substitution changing the third base of a codon; the resulting amino acid sequence is unchanged. C A missense mutation in which the nucleotide substitution results in a change in a single amino acid from the normal sequence (AAG) encoding lysine to glutamine (CAG). D Insertion of a G residue (boxed) causes a frameshift mutation, completely altering the amino acid sequence downstream. This usually results in a loss-of-function mutation. E A nonsense mutation resulting in a single nucleotide change from a lysine codon (AAG) to a premature stop codon (TAG). Normal Silent polymorphism (no amino acid change) Missense mutation causing Lys–Gln amino acid change ‘G’ insertion causing frameshift mutation Nonsense mutation causing premature termination codon A B C D E 3.1 Classes of genetics variant The classes of genetic variant can be illustrated using the sentence ‘THE FAT FOX WAS ILL COS SHE ATE THE OLD CAT’ Synonymous Silent polymorphism with no amino acid change THE FAT FOX WAS ILL COS SHE ATE THE OLD KAT where the C is replaced with a K but the meaning remains the same Non-synonymous Causing an amino acid change THE FAT BOX WAS ILL COS SHE ATE THE OLD CAT where the F of FOX is replaced by a B and the original meaning of the sentence is lost Stop gain (also called a nonsense mutation) Causing the generation of a premature stop codon THE CAT where the F of FAT is replaced by a C generating a premature stop codon Indel Where bases are either inserted or deleted; disruption of the reading frame is dependent on the number of bases inserted or deleted THE FAT FOX WAS ILL ILL COS SHE ATE THE OLD CAT where the insertion of three bases results in maintenance of the reading frame THE FAT FOX WAW ASI LLC OSS HEA TET HEO LDC AT where the insertion of two bases results in disruption of the reading frame Genomics, health and disease Classes of genetic variant There are many different classes of variation in the human genome, categorised by the size of the DNA segment involved and/or by the mechanism giving rise to the variation. Nucleotide substitutions The substitution of one nucleotide for another is the most common type of genomic variation. Depending on their frequency and functional consequences, these changes are known as point mutations or single nucleotide polymorphisms (SNPs). They occur by misincorporation of a nucleotide during DNA synthesis or by chemical modification of the base. When these substitutions occur within ORFs of a protein-coding gene, they are further classified into: • synonymous – resulting in a change in the codon without altering the amino acid • non-synonymous (also known as a missense mutation) – resulting in a change in the codon and the encoded amino acid • stop gain (or nonsense mutation) – introducing a premature stop codon and resulting in truncation of the protein • splicing – taking place at splice sites that most frequently occur at the junction between an intron and an exon. These different

types of mutation are illustrated in Box 3.1 and examples are shown in Figures 3.5 and 3.6. Insertions and deletions One or more nucleotides may be inserted or lost in a DNA sequence, resulting in an insertion/deletion (indel) polymorphism or

size of the original repeat, in that longer repeats tend to be more unstable. Many microsatellites and minisatellites occur in introns or in chromosomal regions between genes and have no obvious adverse effects. However, some genetic diseases are caused by microsatellite repeats that result in duplication of amino acids within the affected gene product or affect gene expression (Box 3.2). Simple tandem repeat mutations Variations in the length of simple tandem repeats of DNA are thought to arise as the result of slippage of DNA during meiosis and are termed microsatellite (small) or minisatellite (larger) repeats. These repeats are unstable and can expand or contract in different generations. This instability is proportional to the Fig. 3.6 Splice site mutations. A The normal sequence is shown, illustrating two exons, and intervening intron (blue) with splice donor (AG) and splice acceptor sites (GT) underlined. Normally, the intron is removed by splicing to give the mature messenger RNA that encodes the protein. B In a splice site mutation the donor site is mutated. As a result, splicing no longer occurs, leading to read-through of the mRNA into the intron, which contains a premature termination codon downstream of the mutation. Normal Splice site mutation Splice donor site Exon Exon Intron Exon Exon Intron Intron removed by splicing Splice acceptor site mRNA ‘reads through’ intron Abnormal protein with premature stop codon A B 3.2 Diseases associated with triplet and other repeat expansions* Repeat No. of repeats Gene Gene location Inheritance Normal Mutant Coding repeat expansion Huntington’s disease [CAG] 6–34

> 35 Huntingtin 4p16 AD Spinocerebellar ataxia (type 1) [CAG] 6–39 40 Ataxin 6p22–23 AD Spinocerebellar ataxia (types 2, 3, 6, 7) [CAG] Various Various Various Various AD Dentatorubral-pallidoluysian atrophy [CAG] 7–25 49 Atrophin 12p12–13 AD Machado–Joseph disease [CAG] 12–40 67 MJD 14q32 AD Spinobulbar muscular atrophy [CAG] 11–34 40 Androgen receptor Xq11–12 XL recessive Non-coding repeat expansion Myotonic dystrophy [CTG] 5–37 50 DMPK-3′UTR 19q13 AD Friedreich’s ataxia [GAA] 7–22 200 Frataxin-intronic 9q13 AR Progressive myoclonic epilepsy [CCCCGCCCCGCG]4–8 2–3 25 Cystatin B-5′UTR 21q AR Fragile X mental retardation [CGG] 5–52 200 FMR1–5′UTR Xq27 XL dominant Fragile site mental retardation 2 (FRAXE) [GCC] 6–35 200 FMR2 Xq28 XL, probably recessive *The triplet repeat diseases fall into two major groups: those with disease stemming from expansion of [CAG]n repeats in coding DNA, resulting in multiple adjacent glutamine residues (polyglutamine tracts), and those with non-coding repeats. The latter tend to be longer. Unaffected parents usually display ‘pre-mutation’ allele lengths that are just above the normal range. (AD/AR = autosomal dominant/recessive; UTR = untranslated region; XL = X-linked)

 Copy number variations Variation in the number of copies of an individual segment of the genome from the usual diploid (two copies) content can be categorised by the size

of the segment involved. Rarely, individuals may gain (trisomy) or lose (monosomy) a whole chromosome. Such numerical chromosome anomalies most commonly occur by a process known as non-disjunction, where pairs of homologous chromosomes do not separate at meiosis II (p. 40). Common trisomies include Down's syndrome (trisomy 21), Edward's syndrome (trisomy 18) and Patau's syndrome (trisomy 13). Monosomy of the autosomes (present in all the cells, as opposed to in a mosaic distribution) does not occur but Turner's syndrome, in which there is monosomy for the X chromosome, affects approximately 1 in 2500 live births (Box 3.3). Large insertions or deletions of chromosomal DNA also occur and are usually associated with a learning disability and/or congenital malformations. Such structural chromosomal anomalies usually arise as the result of one of two different processes: • non-homologous end-joining • non-allelic homologous recombination. Random double-stranded breaks in DNA are a necessary process in meiotic recombination and also occur during mitosis at a predictable rate. The rate of these breaks is dramatically increased by exposure to ionising radiation. When such breaks take place, they are usually repaired accurately by DNA repair mechanisms within the cell. However, in a proportion of breaks, segments of DNA that are not normally contiguous will be joined ('non-homologous end-joining'). If the joined fragments are from different chromosomes, this results in a translocation. If they are 3.3 Chromosome and contiguous gene disorders Disease Locus Incidence Clinical features Numerical chromosomal abnormalities Down's syndrome (trisomy 21) 47,XY,+21 or 47,XX+21 1 in 800 Characteristic facies, IQ usually < 50, congenital heart disease, reduced life expectancy Edwards' syndrome (trisomy 18) 47,XY,+18 or 47,XX,+18 1 in 6000 Early lethality, characteristic skull and facies, frequent malformations of heart, kidney and other organs Patau's syndrome (trisomy 13) 47,XY,+13 or 47, XX,+13 1 in 15 000 Early lethality, cleft lip and palate, polydactyly, small head, frequent congenital heart disease Klinefelter's syndrome 47,XXY 1 in 1000 Phenotypic male, infertility, gynaecomastia, small testes (p. 660) XYY 47,XYY 1 in 1000 Usually asymptomatic, some impulse control problems Triple X syndrome 47,XXX 1 in 1000 Usually asymptomatic, may have reduced IQ Turner's syndrome 45,X 1 in 5000 Phenotypic female, short stature, webbed neck, coarctation of the aorta, primary amenorrhoea (p. 659) Recurrent deletions, microdeletions and contiguous gene defects Di George/velocardiofacial syndrome 22q11.2 1 in 4000 Cardiac outflow tract defects, distinctive facial appearance, thymic hypoplasia, cleft palate and hypocalcaemia. Major gene seems to be TBX1 (cardiac defects and cleft palate) Prader–Willi syndrome 15q11–q13 1 in 15 000 Distinctive facial appearance, hyperphagia, small hands and feet, distinct behavioural phenotype. Imprinted region, deletions on paternal allele in 70% of cases Angelman's syndrome 15q11–q13 1 in 15 000 Distinctive facial appearance, absent speech, electroencephalogram (EEG) abnormality, characteristic gait. Imprinted region, deletions on maternal allele encompassing UBE3A Williams' syndrome 7q11.23 1 in 10 000 Distinctive facial appearance, supravalvular aortic stenosis, learning disability and infantile hypercalcaemia. Major gene for supravalvular aortic stenosis is elastin Smith–Magenis syndrome 17p11.2 1 in 25 000 Distinctive facial appearance and behavioural phenotype, self-injury and rapid eye movement (REM) sleep abnormalities. Major gene seems to be RAI1 from the same chromosome, this will result in either inversion, duplication or deletion of a chromosomal fragment (Fig. 3.7). Large insertions and deletions may be cytogenetically visible as chromosomal deletions or duplications. If the anomalies are too small to be detected by microscopy, they are termed microdeletions and microduplications. Many microdeletion syndromes have been described and most result from nonallelic homologous recombination between repeats of highly similar DNA sequences, which leads to recurrent chromosome anomalies – and clinical syndromes – occurring in unrelated individuals (Fig. 3.7 and Box 3.3). Consequences of genomic variation The consequence of an individual mutation depends

on many factors, including the mutation type, the nature of the gene product and the position of the variant in the protein. Mutations can have profound or subtle effects on gene and cell function. Variations that have profound effects are responsible for ‘classical’ genetic diseases, whereas those with subtle effects may contribute to the pathogenesis of common disease where there is a genetic component, such as diabetes. • Neutral variants have no effect on quality or type of protein produced. • Loss-of-function mutations result in loss or reduction in the normal protein function. Whole-gene deletions are the archetypal loss-of-function variants but stop-gain or indel mutations (early in the ORF), missense mutations affecting a critical domain and splice-site mutations can also result in loss of protein function.

as a common polymorphism. However, the most frequent is the single nucleotide polymorphism, or SNP (pronounced ‘snip’), describing the substitution of a single base. Polymorphisms and common disease The protective and detrimental polymorphisms associated with common disease have been identified primarily through genome-wide association studies (GWAS, p. 56) and are the basis for many direct-to-consumer tests that purport to determine individual risk profiles for common diseases or traits such as cardiovascular disease, diabetes and even male-pattern baldness! An example is the polymorphism in the gene SLC2A9 that not only explains a significant proportion of the normal population variation in serum urate concentration but also predisposes ‘high-risk’ allele carriers to the development of gout. However, the current reality is that, until we have a more comprehensive understanding of the full genomic landscape and knowledge of the complete set of detrimental and protective polymorphisms, we cannot accurately assess risk. Evolutionary selection Genetic variants play an important role in evolutionary selection, with advantageous variants resulting in positive selection via improved reproductive fitness, and variations that decrease • Gain-of-function mutations result in a gain of protein function. They are typically non-synonymous mutations that alter the protein structure, leading to activation/ alteration of its normal function through causing either an interaction with a novel substrate or a change in its normal function. • Dominant negative mutations are the result of nonsynonymous mutations or in-frame deletions/duplications but may also, less frequently, be caused by triplet repeat expansion mutations. Dominant negative mutations are heterozygous changes that result in the production of an abnormal protein that interferes with the normal functioning of the wild-type protein. Normal genomic variation We each have 5–50 million variants in our genome, occurring approximately every 300 bases. These variants are mostly polymorphisms, arising in more than 1% of the population; they have no or subtle effects on gene and cell function, and are not associated with a high risk of disease. Polymorphisms can occur within exons, introns or the intergenic regions that comprise 98–99% of the human genome. Each of the classes of genetic variant discussed on page 42 is present in the genome Fig. 3.7 Chromosomal analysis and structural chromosomal disorders. A Human chromosomes can be classed as metacentric if the centromere is near the middle, or acrocentric if the centromere is at the end. The bands of each chromosome are given a number, starting at the centromere and working out along the short (p) arm and long (q) arm. Translocations and inversions are balanced structural chromosome anomalies where no genetic material is missing but it is in the wrong order. Translocations can be divided into reciprocal (direct swap of chromosomal material between nonhomologous chromosomes) and Robertsonian (fusion of acrocentric chromosomes). Deletions and duplications can also occur due to non-allelic homologous recombination (illustrated in part B). Deletions are classified as interstitial if they lie

within a chromosome, and terminal if the terminal region of the chromosome is affected. Duplications can be either in tandem (where the duplicated fragment is inserted next to the region that is duplicated and orientated in the correct direction) or inverted (where the duplicated fragment is in the wrong direction). (N = normal; A = abnormal) B A common error of meiotic recombination, known as non-allelic homologous recombination, can occur (right panel), resulting in a deletion on one chromosome and a duplication in the homologous chromosome. The error is induced by tandem repeats in the DNA sequences (green), which can misalign and bind to each other, thereby 'fooling' the DNA into thinking the pairing prior to recombination is correct. Chromosome

Cen p

21.1 21.3 21.2

34.2 q Cen p N N A N A N A N A N A N A A 13 (sa) (st) 11.2

24.2

32.2 q Metacentric Acrocentric Chromosome

Reciprocal translocation Robertsonian translocation Inversions Interstitial Terminal Tandem Inverted Mechanism underlying recurrent deletions and duplication: non-allelic homologous recombination How structural chromosomal anomalies are described Deletions Duplications B A Recombination Normal pairing Abnormal pairing between DNA repeats Deletion Duplication

DNA repeat Maternal chromosome Paternal chromosome

46 • CLINICAL GENETICS history, on both sides of the family, enquiring about details of all medical conditions in family members, consanguinity, dates of birth and death, and any history of pregnancy loss or infant death. The basic symbols and nomenclature used in drawing a pedigree are shown in Figure 3.8. Patterns of disease inheritance Autosomal dominant inheritance Take some time to draw out the following pedigree: Anne is referred to Clinical Genetics to discuss her personal history of colon cancer (she was diagnosed at the age of 46 years) and family history of colon/endometrial cancer: her mother was diagnosed with endometrial cancer at the age of 60 years and her cousin through her healthy maternal aunt was diagnosed with colon cancer in her fifties. Both her maternal grandmother and grandfather died of 'old age'. There is no family history of note on her father's side of the family. He has one brother and both his parents died of old age, in their eighties. Anne has two healthy daughters, aged 12 and 14 years, and a healthy full sister. This family history is typical of an autosomal dominant condition (Fig. 3.8): in this case, a colon/endometrial cancer susceptibility syndrome known as Lynch's syndrome, associated with disruption of one of the mismatch repair genes: MSH2, MSH6, MLH1 and PMS2 (see p. 830 and Box 3.11, p. 57). reproductive fitness becoming excluded through evolution. Given this simple paradigm, it would be tempting to assume that common mutations are all advantageous and all rare mutations are pathogenic. Unfortunately, it is often difficult to classify any common mutation as either advantageous or deleterious – or, indeed, neutral. Mutations that are advantageous in early life and thus enhance reproductive fitness may be deleterious in later life. There may be mutations that are advantageous for survival in particular conditions (e.g. famine or pandemic)

that may be disadvantageous in more benign circumstances by causing a predisposition to obesity or autoimmune disorders. Constitutional genetic disease Familial genetic disease is caused by constitutional mutations, which are inherited through the germ line. However, different mutations in the same gene can have different consequences, depending on the genetic mechanism underlying that disease. About 1% of the human population carries constitutional mutations that cause disease. Constructing a family tree The family tree – or pedigree – is a fundamental tool of the clinical geneticist, who will routinely take a three-generation family Fig. 3.8 Drawing a pedigree and patterns of inheritance. A The main symbols used to represent pedigrees in diagrammatic form. B The main modes of disease inheritance (see text for details). SB Male Clinically affected Deceased individual (with age at death) Separated Consanguinity Clinically affected, several diagnoses Carrier Positive presymptomatic test Monozygotic twins Dizygotic twins Stillbirth (with gestation) Termination Miscarriage (with gestation) Unknown sex Female Partners Recessive inheritance Dominant inheritance Mitochondrial DNA disorder X-linked recessive inheritance Transmission to 50% of offspring independent of gender Consanguinity Affected males related through unaffected females Both sexes affected but only inherited through female meiosis I II III IV I II III IV I II III IV V I II III IV

d. 50 y 30 wk SB 39 wk 16 wk A B

Genomics, health and disease • 47

• Males and females are usually affected in roughly equal numbers (unless the clinical presentation of the condition is gender-specific, such as an inherited susceptibility to breast and/or ovarian cancer). The offspring risk for an individual affected with an autosomal dominant condition is 1 in 2 (or 50%). This offspring risk is true for each pregnancy, since half the affected individual gametes (sperm or egg cells) will contain the affected chromosome/gene and half will contain the normal chromosome/gene. There is a long list of autosomal dominant conditions, some of which are shown in Box 3.4. Features of an autosomal dominant pedigree include: • There are affected individuals in each generation (unless the mutation has arisen de novo, i.e. for the first time in an affected individual). However, variable penetrance and expressivity can influence the number of affected individuals and the severity of disease in each generation. Penetrance is defined as the proportion of individuals bearing a mutated allele who develop the disease phenotype. The mutation is said to be fully penetrant if all individuals who inherit a mutation develop the disease. Expressivity describes the level of severity of each aspect of the disease phenotype. 3.4 Genetic conditions dealt with by clinicians in other specialties Name of condition Gene Reference Autosomal dominant conditions Autosomal dominant polycystic kidney disease (ADPKD) PKD1 (85%), PKD2 (15%) p. 405 Box 15.28, p. 415 Tuberous sclerosis TSC1 TSC2 p. 1264 p. 1264 Marfan’s syndrome FBN1 p. 508 Long QT syndrome KCNQ1 p. 476 Brugada’s syndrome SCN5A p. 477 Neurofibromatosis type 1 NF1 p. 1131 Box 25.77, p. 1132 Neurofibromatosis type 2 NF2 p. 1131 Box 25.77, p. 1132 Hereditary spherocytosis ANK1 p. 947 Vascular Ehlers–Danlos syndrome (EDS type 4) COL3A1 p. 970 Hereditary haemorrhagic telangiectasia ENG, ALK1, GDF2 p. 970 Osteogenesis imperfecta COL1A1, COL1A2 p. 1055 Charcot–Marie–Tooth disease PMP22, MPZ, GJB1 p. 1140 Hereditary neuropathy with liability to pressure palsies PMP22 Autosomal recessive conditions Familial Mediterranean fever MEFV p. 81 Mevalonic aciduria (mevalonate kinase deficiency) MVK p. 81 Autosomal recessive polycystic kidney disease (ARPKD) PKHD1 Box 15.28, p. 415 Kartagener’s syndrome (primary ciliary dyskinesia) DNAI1 Box 17.30, p. 578 Cystic fibrosis CFTR1 p. 580 Box 17.30, p. 578

p. 842 Pendred’s syndrome SLC26A4 p. 650 Congenital adrenal hyperplasia-21 hydroxylase deficiency CYP21A p. 676 Box 18.27, p. 658 Haemochromatosis HFE p. 895 Wilson’s disease ATP7B p. 896 Alpha1-antitrypsin deficiency SERPINA1 p. 897 Gilbert’s syndrome UGT1A1 p. 897 Benign recurrent intrahepatic cholestasis ATP8B1 p. 902 Alpha-thalassaemia HBA1, HBA2 p. 951 p. 954 Beta-thalassaemia HBB p. 951 p. 953 Sickle cell disease HBB p. 951 Spinal muscular atrophy SMN1 p. 1117 X-linked conditions Alport’s syndrome COL4A5 Box 15.28, p. 415 p. 403 Primary agammaglobulinaemia BTK p. 78 Haemophilia A (factor VIII deficiency) F8 p. 971 Haemophilia B (factor IX deficiency) F9 p. 973 Duchenne muscular dystrophy DMD p. 1143 and Box 25.91

48 • CLINICAL GENETICS Autosomal recessive inheritance As above, take some time to draw a pedigree representing the following: Mr and Mrs Kent, a non-consanguineous couple, are referred because their son, Jamie, had severe neonatal liver disease. Included among the many investigations that the paediatric hepatologist undertook was testing for α1-antitrypsin deficiency (Box 3.5). Jamie was shown to have the PiZZ phenotype. Testing confirmed both parents as carriers with PiMZ phenotypes. In the family, Jamie has an older sister who has no medical problems. Mr Kent is one of four children with two brothers and a sister and Mrs Kent has a younger brother. Both sets of grandparents are alive and well. There is no family history of α1-antitrypsin deficiency. This family history is characteristic of an autosomal recessive disorder (Fig. 3.8), where both alleles of a gene must be mutated before the disease is manifest in an individual; an affected individual inherits one mutant allele from each of their parents, who are therefore healthy carriers for the condition. An autosomal recessive condition might be suspected in a family where: • Males and females are affected in roughly equal proportions. • Parents are blood related; this is known as consanguinity. Where there is consanguinity, the mutations are usually homozygous, i.e. the same mutant allele is inherited from both parents. • Individuals within one sibship in one generation are affected and so the condition can appear to have arisen ‘out of the blue’. Approximately 1 in 4 children born to carriers of an autosomal recessive condition will be affected. The offspring risk for carrier parents is therefore 25% and the chances of an unaffected child, with an affected sibling, being a carrier is 2/3. Examples of some autosomal recessive conditions, discussed elsewhere in this book, are shown in Box 3.4. X-linked inheritance The following is an exemplar of an X-linked recessive pedigree (Fig. 3.8): Edward has a diagnosis of Duchenne muscular dystrophy (DMD, Box 3.6). His parents had suspected the diagnosis when he was 3 years old because he was not yet walking and there was a family history of DMD: Edward’s maternal uncle had been affected and died at the age of 24 years. Edward’s mother has no additional siblings. After Edward demonstrated a very high creatinine kinase level, the paediatrician also requested genetic testing, which identified a deletion of exons 2–8 of the dystrophin gene. Edward has a younger, healthy sister and grandparents on both sides of the family are well, although the maternal grandmother has recently developed a cardiomyopathy. Edward’s father has an older sister and an older brother who are both well. Genetic diseases caused by mutations on the X chromosome have specific characteristics: • X-linked diseases are mostly recessive and restricted to males who carry the mutant allele. This is because males 3.5 Alpha1-antitrypsin deficiency Inheritance pattern • Autosomal recessive Genetic cause • Two common mutations in the SERPINA1 gene: p.Glu342Lys and p.Glu264Val Prevalence • 1 in 1500–3000 of European ancestry Clinical presentation • Variable presentation from neonatal period through to adulthood • Neonatal period: prolonged jaundice with conjugated hyperbilirubinaemia or (rarely) liver disease • Adulthood: pulmonary emphysema and/or cirrhosis. Rarely, the skin disease, panniculitis, develops Disease mechanism • SERPINA1 encodes α1-antitrypsin, which protects the body from the effects of neutrophil elastase. The

symptoms of α1-antitrypsin deficiency result from the effects of this enzyme attacking normal tissue Disease variants • M variant: if an individual has normal SERPINA1 genes and produces normal levels of α1-antitrypsin, they are said to have an M variant • S variant: p.Glu264Val mutation results in α1-antitrypsin levels reduced to about 40% of normal • Z variant: p.Glu342Lys mutation results in very little α1-antitrypsin • PiZZ: individuals who are homozygous for the p.Glu342Lys mutation are likely to have α1-antitrypsin deficiency and the associated symptoms • PiZS: individuals who are compound heterozygous for p.Glu342Lys and p.Glu264Val are likely to be affected, especially if they smoke, but usually to a milder degree 3.6 Duchenne muscular dystrophy* Inheritance pattern • X-linked recessive Genetic cause • Mutations or deletions encompassing/within the DMD (dystrophin) gene located at Xp21 Prevalence • 1 in 3000–4000 live male births Clinical presentation • Delayed motor milestones • Speech delay • Grossly elevated creatine kinase (CK) levels (in the thousands) • Ambulation is usually lost between the ages of 7 and 13 years • Lifespan is reduced with a mean age of death, usually from respiratory failure, in the mid-twenties • Cardiomyopathy affects almost all boys with Duchenne muscular dystrophy and some female carriers Disease mechanism • DMD encodes dystrophin, a major structural component of muscle • Dystrophin links the internal cytoskeleton to the extracellular matrix Disease variants • Becker muscular dystrophy, although a separate disease, is also caused by mutations in the dystrophin gene • In Duchenne muscular dystrophy, there is no dystrophin protein, whereas in Becker muscular dystrophy there is a reduction in the amount or alteration in the size of the dystrophin protein *See also page 1143.

dinucleotide (NADH) and the reduced form of flavine adenine dinucleotide (FADH2). Both NADH and FADH2 then donate electrons to the respiratory chain. Here these elections are transferred via a complex series of reactions, resulting in the formation of a proton gradient across the inner mitochondrial membrane. The gradient is used by an inner mitochondrial membrane protein, ATP synthase, to produce ATP, which is then transported to other parts of the cell. Dephosphorylation of ATP is used to produce the energy required for many cellular processes. Each mitochondrion contains 2–10 copies of a 16-kilobase (kB) double-stranded circular DNA molecule (mtRNA). This mtDNA contains 13 protein-coding genes, all involved in the respiratory chain, and the ncRNA genes required for protein synthesis within the mitochondria (Fig. 3.9). The mutational rate of mtDNA is relatively high due to the lack of protection by chromatin. Several mtDNA diseases characterised by defects in ATP production have been described. Mitochondria are most numerous in cells with high metabolic demands, such as muscle, retina and the basal ganglia, and these tissues tend to be the ones most severely affected in mitochondrial diseases (Box 3.7). There are many other mitochondrial diseases that are caused by mutations in nuclear genes, which encode proteins that are then imported into the mitochondrion and are critical for energy production, e.g. most forms of Leigh’s syndrome (although Leigh’s syndrome may also be caused by a mitochondrial gene mutation). The inheritance of mtDNA disorders is characterised by transmission from females, but males and females generally are equally affected (see Fig. 3.8). Unlike the other inheritance patterns mentioned above, mitochondrial inheritance has nothing to do with meiosis but reflects the fact that mitochondrial DNA is transmitted by oöcytes: sperm do not contribute mitochondria to the zygote. Mitochondrial disorders tend to be variable in penetrance and expressivity within families, and this is mostly accounted for by the fact that only a proportion of multiple mtDNA molecules within mitochondria contain the causal mutation (the degree of mtDNA

heteroplasmy). Imprinting Several chromosomal regions (loci) have been identified where gene expression is inherited in a parent-of-origin-specific manner; have only one X chromosome, whereas females have two (see Fig. 3.1). However, occasionally, female carriers may exhibit signs of an X-linked disease due to a phenomenon called skewed X-inactivation. All female embryos, at about 100 cells in size, stably inactivate one of their two X chromosomes in each cell. Where this inactivation is random, approximately 50% of the cells will express the genes from one X chromosome and 50% of cells will express genes from the other. Where there is a mutant gene, there is often skewing away from the associated X chromosome, resulting in an unaffected female carrier. However, if, by chance, there is a disproportionate inactivation of the normal X chromosome with skewing towards the mutant allele, then an affected female carrier may be affected (albeit more mildly than males). • The gene can be transmitted from female carriers to their sons: in families with an X-linked recessive condition, there are often a number of affected males related through unaffected females. • Affected males cannot transmit the condition to their sons (but all their daughters would be carriers). The risk of a female carrier having an affected child is 25% or half of her male offspring. Mitochondrial inheritance The mitochondrion is the main site of energy production within the cell. Mitochondria arose during evolution via the symbiotic association with an intracellular bacterium. They have a distinctive structure with functionally distinct inner and outer membranes. Mitochondria produce energy in the form of adenosine triphosphate (ATP). ATP is mostly derived from the metabolism of glucose and fat (Fig. 3.9). Glucose cannot enter mitochondria directly but is first metabolised to pyruvate via glycolysis. Pyruvate is then imported into the mitochondrion and metabolised to acetyl-co-enzyme A (acetyl-CoA). Fatty acids are transported into the mitochondria following conjugation with carnitine and are sequentially catabolised by a process called β-oxidation to produce acetyl-CoA. The acetyl-CoA from both pyruvate and fatty acid oxidation is used in the citric acid (Krebs) cycle – a series of enzymatic reactions that produces CO2, the reduced form of nicotinamide adenine 3.7 The structure of the respiratory chain complexes and the diseases associated with their dysfunction Complex Enzyme nDNA subunits1 mtDNA subunits2 Diseases I NADH dehydrogenase

MELAS, MERRF bilateral striatal necrosis, LHON, myopathy and exercise intolerance, Parkinsonism, Leigh’s syndrome, exercise myoglobinuria, leucodystrophy/myoclonic epilepsy II Succinate dehydrogenase

Phaeochromocytoma, Leigh’s syndrome III Cytochrome bc1 complex

Parkinsonism/MELAS, cardiomyopathy, myopathy, exercise myoglobinuria, Leigh’s syndrome IV Cytochrome c oxidase

Sideroblastic anaemia, myoclonic ataxia, deafness, myopathy, MELAS, MERRF mitochondrial encephalomyopathy, motor neuron disease-like, exercise myoglobinuria, Leigh’s syndrome V ATP synthase

Leigh’s syndrome, NARP, bilateral striatal necrosis 1nDNA subunits. 2mtDNA subunits = number of different protein subunits in each complex that are encoded in the nDNA and mtDNA, respectively. (ATP = adenosine triphosphate; LHON = Leber hereditary optic neuropathy; MELAS = myopathy, encephalopathy, lactic acidosis and stroke-like episodes; MERRF = myoclonic epilepsy and ragged red fibres; mtDNA = mitochondrial DNA; NADH = the reduced form of nicotinamide adenine

dinucleotide; NARP = neuropathy, ataxia and retinitis pigmentosa; nDNA = nuclear DNA)

50 • CLINICAL GENETICS Somatic genetic disease Somatic mutations are not inherited but instead occur during post-zygotic mitotic cell divisions at any point from embryonic development to late adult life. An example of this phenomenon is polyostotic fibrous dysplasia (McCune–Albright syndrome), in Fig. 3.9 Mitochondria. A Mitochondrial structure. There is a smooth outer membrane surrounding a convoluted inner membrane, which has inward projections called cristae. The membranes create two compartments: the inter-membrane compartment, which plays a crucial role in the electron transport chain, and the inner compartment (or matrix), which contains mitochondrial DNA and the enzymes responsible for the citric acid (Krebs) cycle and the fatty acid β-oxidation cycle. B Mitochondrial DNA. The mitochondrion contains several copies of a circular double-stranded DNA molecule, which has a non-coding region, and a coding region that encodes the genes responsible for energy production, mitochondrial transfer RNA (tRNA) molecules and mitochondrial ribosomal RNA (rRNA) molecules. (ATP = adenosine triphosphate; NADH = the reduced form of nicotinamide adenine dinucleotide) C Mitochondrial energy production. Fatty acids enter the mitochondrion conjugated to carnitine by carnitine-palmityl transferase type 1 (CPT I) and, once inside the matrix, are unconjugated by CPT II to release free fatty acids (FFA). These are broken down by the β-oxidation cycle to produce acetyl-co-enzyme A (acetyl-CoA). Pyruvate can enter the mitochondrion directly and is metabolised by pyruvate dehydrogenase (PDH) to produce acetyl-CoA. The acetyl-CoA enters the Krebs cycle, leading to the production of NADH and flavine adenine dinucleotide (reduced form) (FADH2), which are used by proteins in the electron transport chain to generate a hydrogen ion gradient across the inter-membrane compartment. Reduction of NADH and FADH2 by proteins I and II, respectively, releases electrons (e), and the energy released is used to pump protons into the inter-membrane compartment. Coenzyme Q10/ubiquinone (Q) is an intensely hydrophobic electron carrier that is mobile within the inner membrane. As electrons are exchanged between proteins in the chain, more protons are pumped across the membrane, until the electrons reach complex IV (cytochrome oxidase), which uses the energy to reduce oxygen to water. The hydrogen ion gradient is used to produce ATP by the enzyme ATP synthase, which consists of a proton channel and catalytic sites for the synthesis of ATP from ADP. When the channel opens, hydrogen ions enter the matrix down the concentration gradient, and energy is released that is used to make ATP. L

s t r a n d H

s t r a n d Outer membrane Inner membrane NADH NAD I II III Q Cyt C IV NADH FADH2 Fatty acid β-oxidation cycle Citric acid (Krebs) cycle H+ e 2e FADH2 FADH2 Lactate Pyruvate PDH Acetyl-CoA Glucose 22 tRNAs NADH dehydrogenase 7 subunits Cytochrome B/C oxidase 4 subunits 2 ribosomal RNA subunits 2 ATP synthase subunits Intragenic DNA Inner membrane Cristae Matrix Outer membrane FFA CPT I CPT II Carnitine Carnitine-FA ester C A B FFA FAD 2H+ H2O O2 ATP ADP

- Pi H+ H+ ATP synthase Carnitine e e e these are called imprinted loci. Within these loci the paternally inherited gene may be active while the maternally inherited may be silenced, or vice versa. Mutations within imprinted loci lead to an unusual pattern of inheritance where the phenotype is manifest only if inherited from the parent who contributes the transcriptionally active allele. Examples of imprinting disorders are given in Box 3.8.

light or cigarette smoke, or if the cell has defects in DNA repair systems. Cancer is thus a disease that affects the fundamental processes of molecular and cell biology. Interrogating the genome: the changing landscape of genomic technologies Looking at chromosomes The analysis of metaphase chromosomes by light microscopy was the mainstay of clinical cytogenetic analysis for decades, the aim being to detect gain or loss of whole chromosomes (aneuploidy) or large chromosomal segments (> 4 million bp). More recently, genome-wide microarrays (array comparative genomic hybridisation or array CGH) have replaced chromosome analysis, allowing rapid and precise detection of segmental gain or loss of DNA throughout the genome (see Box 3.3). Microarrays consist of grids of multiple wells containing short DNA sequences (reference DNA) that are complementary to known sequences in the genome. Patient and reference DNA are each labelled with a coloured fluorescent dye (generally, patient DNA is labelled with a green fluorescent dye and reference DNA with a red fluorescent dye) and added to the microarray grid. Where there is an equal quantity of patient and reference DNA bound to the spot, this results in yellow fluorescence. Where there is too much patient DNA (representing a duplication of a chromosome region), the spot will be greener; it will be more red (appears orange) where there is 2 : 1 ratio of the control:patient DNA (representing heterozygous deletion of a chromosome region; Fig. 3.10). Array CGH and other array-based approaches can detect small chromosomal deletions and duplications. They are also generally more sensitive than conventional karyotyping at detecting mosaicism (where there are two or more populations of cells, derived from a single fertilised egg, with different genotypes). which a somatic mutation in the GS alpha gene causes constitutive activation of downstream signalling, resulting in focal lesions in the skeleton and endocrine dysfunction (p. 1055). The most important example of human disease caused by somatic mutations is cancer (see Ch. 33). Here, ‘driver’ mutations occur within genes that are involved in regulating cell division or apoptosis, resulting in abnormal cell growth and tumour formation. The two general categories of cancer-causing mutation are gain-of-function mutations in growth-promoting genes (oncogenes) and loss-of-function mutations in growth-suppressing genes (tumour suppressor genes). Whichever mechanism is acting, most tumours require an initiating mutation in a single cell that can then escape from normal growth controls. This cell replicates more frequently or fails to undergo programmed death, resulting in clonal expansion. As the size of the clone increases, one or more cells may acquire additional mutations that confer a further growth advantage, leading to proliferation of these subclones, which may ultimately result in aggressive metastatic cancer. The cell’s complex self-regulating machinery means that more than one mutation is usually required to produce a malignant tumour (see Fig. 33.3, p. 1318). For example, if a mutation results in activation of a growth factor gene or receptor, then that cell will replicate more frequently as a result of autocrine stimulation. However, this mutant cell will still be subject to normal cell-cycle checkpoints to promote DNA integrity in its progeny. If additional mutations in the same cell result in defective cell-cycle checkpoints, however, it will rapidly accumulate further mutations, which may allow completely unregulated growth and/ or separation from its matrix and cellular attachments and/or resistance to apoptosis. As cell growth becomes increasingly dysregulated, cells de-differentiate, lose their response to normal tissue environment and cease to ensure appropriate mitotic chromosomal segregation. These processes combine to generate the classical malignant characteristics of disorganised growth, variable levels of differentiation, and numerical and structural chromosome abnormalities. An increase in somatic mutation rate can occur on exposure to external mutagens, such as ultraviolet 3.8 Imprinting disorders Disorder

Locus Genes Notes Beckwith–Wiedemann syndrome 11p15 CDKN1C, IGF2, H19 Increased growth, macroglossia, hemihypertrophy, abdominal wall defects, ear lobe pits/creases and increased susceptibility to developing childhood tumours Prader–Willi syndrome 15q11–q13 SNRPN, Necdin and others Obesity, hypogonadism and learning disability. Lack of paternal contribution (due to deletion of paternal 15q11–q13, or inheritance of both chromosome 15q11–q13 regions from the mother) Angelman's syndrome (AS) 15q11–q13 UBE3A Severe mental retardation, ataxia, epilepsy and inappropriate laughing bouts. Due to loss-of-function mutations in the maternal UBE3A gene. The neurological phenotype results because most tissues express both maternal and paternal alleles of UBE3A, whereas the brain expresses predominantly the maternal allele Pseudohypoparathyroidism (p. 664) 20q13 GNAS1 Inheritance of the mutation from the mother results in hypocalcaemia, hyperphosphataemia, raised parathyroid hormone (PTH) levels, ectopic calcification, obesity, delayed puberty and shortened 4th and 5th metacarpals (the syndrome known as Albright's hereditary osteodystrophy, AHO). When the mutation is inherited from the father, PTH, calcium and phosphate levels are normal but the other features are present (pseudopseudohypoparathyroidism, p. 664). These differences are due to the fact that, in the kidney (the main target organ through which PTH regulates serum calcium and phosphate), the paternal allele is silenced and the maternal allele is expressed, whereas both alleles are expressed in other tissues.

52 • CLINICAL GENETICS cycle of heating/cooling and denaturation/replication is repeated many times, resulting in the exponential amplification of DNA between primer sites (Fig. 3.11). Gene sequencing In the mid-1970s, a scientist called Fred Sanger pioneered a DNA sequencing technique ('Sanger sequencing') that determined the precise order and nucleotide type (thymine, cytosine, adenine and guanine) in a molecule of DNA. Modern Sanger sequencing uses fluorescently labelled, chain-terminating nucleotides that are sequentially incorporated into the newly synthesised DNA, generating multiple DNA chains of differing lengths. These DNA chains are subject to capillary electrophoresis, which separates them by size, allowing the fragments to be 'read' by a laser and producing a sequence chromatogram that corresponds to the target sequence (Fig. 3.12). Although transformative, Sanger sequencing was difficult and costly to scale, as exemplified by the Human Genome Project, which took 12 years to sequence the entire human genome at a cost approaching 3 billion dollars. Recently, DNA sequencing has been transformed again by a group of technologies collectively known as 'next-generation sequencing' (NGS; Fig. 3.13). This refers to a family of postSanger sequencing technologies that utilise the same five basic principles: • Library preparation: DNA samples are fragmented (by enzyme cleavage or ultrasound) and then modified with a custom adapter sequence. • Amplification: the library fragment is amplified to produce DNA clusters, each originating from a single DNA fragment. Each cluster will act as a single sequencing reaction. • Capture: if an entire genome is being sequenced, this step will not be included. The capture step is required if targeted resequencing is necessary, such as for a panel gene test or an exome (Box 3.9). • Sequencing: each DNA cluster is simultaneously sequenced and the data from each captured; this is known as a 'read' and is usually between 50 and 300 bases long sequenced (see Box 3.10 for a detailed description of the three most commonly used sequencing methods: synthesis, ligation and ion semiconductor sequencing). • Alignment and variant identification: specialised software analyses read sequences and compares the data to a reference template. This is known as 'alignment' or 'mapping' and, although there are 3 billion bases in the However, array-based approaches will not detect balanced chromosome rearrangements where there is no loss or gain of genes/chromosome material, such as balanced

reciprocal translocations, or a global increase in copy number, such as triploidy. The widespread use of array-based approaches has brought a number of challenges for clinical interpretation, including the identification of copy number variants (CNVs) of uncertain clinical significance, CNVs of variable penetrance and incidental findings. A CNV of uncertain clinical significance describes a loss or gain of chromosome material where there are insufficient data to conclude whether or not it is associated with a learning disability and/or medical problems. While this uncertainty can be difficult to prepare families for and can be associated with considerable anxiety, it is likely that there will be greater clarity in the future as we generate larger CNV datasets. A CNV of variable penetrance, also known as a neurosusceptibility locus, describes a chromosome deletion or duplication associated with a lower threshold for manifesting a learning disability or autistic spectrum disorder. CNVs of variable penetrance are therefore identified at greater frequencies among individuals with a learning disability and/or autistic spectrum disorder than in the general population. The current understanding is that additional modifying factors (genetic, environmental or stochastic) must influence the phenotypic expression of these neurosusceptibility loci. Finally, an incidental CNV finding describes a deletion or duplication encompassing a gene or genes that are causative of a phenotype or risk unrelated to the presenting complaint. For instance, if, through the array CGH investigation for an intellectual disability, a deletion encompassing the BRCA1 gene were identified, this would be considered an incidental finding. Looking at genes Gene amplification: polymerase chain reaction The polymerase chain reaction (PCR) is a fundamental laboratory technique that amplifies targeted sections of the human genome for further analyses – most commonly, DNA sequencing. The method utilises thermal cycling: repeated cycles of heating and cooling allow the initial separation of double-stranded DNA into two single strands (known as denaturation), each of which serves as a template during the subsequent replication step, guided by primers designed to anneal to a specific genomic region. This Fig. 3.10 Detection of chromosome abnormalities by comparative genomic hybridisation (CGH). Deletions and duplications are detected by looking for deviation from the 1 : 1 ratio of patient and control DNA in a microarray. Ratios in excess of 1 indicate duplications, whereas ratios below 1 indicate deletions. CGH Patient DNA Label DNA with different fluorescent dyes Mix equimolar amounts of labelled DNA Apply DNA mix to glass slide with high-density array of different DNA probes with known location in the human genome Patient/control ratio = 0.5:1 → deletion of patient DNA Patient/control ratio = 1.5:1 → duplication of patient DNA Patient/control ratio = 1:1 → normal Normal control DNA

human genome, allows the remarkably accurate determination of the genomic origin where a read consists of 25 nucleotides or more. Variants are identified as differences between the read and the reference genome. For instance, if there is a different nucleotide in half the reads at a given position compared to the reference genome, this is likely to represent a heterozygous base substitution. The number of reads that align at a given point is called the ‘depth’ or ‘coverage’. The higher the read depth, the more accurate the variant call. However, in general, a depth of 30 or more reads is generally accepted as producing diagnostic-grade results. Rather than sequencing only one small section of DNA at a time, NGS allows the analysis of many hundreds of thousands of DNA strands in a single experiment and so is also commonly referred to as multiple parallel sequencing technology. Today’s NGS machines can sequence the entire human genome in a single day at a cost approaching 1000 US dollars. NGS capture Although we now have the capability to sequence the entire genome in a single experiment, whole-genome sequencing is not always the

optimal use of NGS. NGS capture refers to the ‘pull-down’ of a targeted region of the genome and may constitute several to several hundred genes associated with a given phenotype (a gene panel), the exons of all known coding genes (an exome), or the exons of all coding genes known to be associated with disease (a clinical exome). Each of these targeted resequencing approaches is associated with a number of advantages and disadvantages (see Box 3.9). In order for NGS to be used for optimal patient benefit, it is essential for the clinician to have a good understanding of which test is the best one to request in any given clinical presentation. Challenges of NGS technologies Genomic technologies have the potential to transform the way that we practise medicine, and ever faster and cheaper DNA sequencing offers increasing opportunities to prevent, diagnose and treat disease. However, genomic technologies are not without their challenges: for instance, storing the enormous quantities of data generated by NGS. While the A, C, T and G of our genomic code could be stored on the memory of a smartphone, huge computers, able to store several petabytes of data (where 1 petabyte is 1 million gigabytes of data), are required to store the information needed to generate each individual’s genome. Even if we can store and handle these huge datasets successfully, we then need to be able to sift through the millions of Fig. 3.11 The polymerase chain reaction (PCR). PCR involves adding a tiny amount of the patient’s DNA to a reaction containing primers (short oligonucleotides 18–21 bp in length, which bind to the DNA flanking the region of interest) and deoxynucleotide phosphates (dATP, dCTP, dGTP, dTTP), which are used to synthesise new DNA and a heat-stable polymerase. The reaction mix is first heated to 95°C, which causes the double-stranded DNA molecules to separate. The reaction is then cooled to 50–60°C, which allows the primers to bind to the target DNA. The reaction is then heated to 72°C, at which point the polymerase starts making new DNA strands. These cycles are repeated 20–30 times, resulting in exponential amplification of the DNA fragment between the primer sites. The resulting PCR products can then be used for further analysis – most commonly, DNA sequencing (see Fig. 3.12). Cool ~60°C DNA sample DNA strands separate Primers bind to DNA DNA replicated Heat 95°C Heat 95°C Cool ~60°C Repeat cycles 20–30 times

PCR cycles Exponential amplification of DNA between primer sites DNA molecules

DNA strands separate Primers bind to DNA Cycle no. 1 DNA replicated Heat ~72°C Heat ~72°C Polymerase

- dNTPs Primers Cycle no. 2

 their interpretation will require input from a genetics expert in the context of the clinical presentation, where an ‘innocent until proven guilty’ approach is often adopted. Finally, if we are to interrogate the entire genome or even the exome, it is foreseeable that we will routinely identify ‘incidental’ or secondary findings – in other words, findings not related to the initial diagnostic question. The UK has so far advocated a conservative approach to incidental findings. Uses of NGS NGS is now frequently used, within diagnostic laboratories, to identify base substitutions and indels (although the latter were Fig. 3.12 Sanger sequencing of DNA, which is very widely used in DNA diagnostics. This is performed using PCR-amplified fragments of DNA corresponding to the gene of interest. The sequencing reaction is carried out with a combination dNTP and fluorescently labelled di-deoxy-dNTP (ddATP, ddTTP, ddCTP and ddGTP), which become incorporated into the newly synthesised DNA, causing termination of the chain at that point. The reaction products are then subject to capillary electrophoresis and the different-sized fragments

are detected by a laser, producing a sequence chromatogram that corresponds to the target DNA sequence. Capillary electrophoresis Largest fragments migrate slowest Smallest fastest Laser fluorescence detector ddTTP ddCTP ddATP ddGTP DNA sample PCR DNA sequence chromatogram Key Fragments detected by laser fluorescence New DNA molecules terminated by incorporation of ddNTP Polymerase

- ddNTPs Primers 3.9 The advantages and disadvantages of whole-genome sequencing, whole-exome sequencing and gene panels Test Advantages Disadvantages Whole-genome sequencing (WGS) The most comprehensive analysis of the genome available More even coverage of genes, allowing better identification of dosage abnormalities Will potentially detect all gene mutations, including intronic mutations More expensive to generate and store Will detect millions of variants in non-coding DNA, which can be very difficult to interpret Associated with a greater risk of identifying incidental findings Shallow sequencing (few reads per gene) and so less sensitive and less able to detect mosaicism Whole-exome sequencing (WES) Cheaper than whole-genome sequencing Analysis is not restricted to only those genes known to cause a given condition Fewer variants detected than in WGS and so easier interpretation Deeper sequencing than WGS increases sensitivity and detection of mosaicism Less even coverage of the genome and so dosage abnormalities are more difficult to detect Less comprehensive analysis (1–2% of the genome) than WGS Increased risk of identifying incidental findings over targeted gene sequencing Gene panels Cost-effective Very deep sequencing, increasing the chances of mosaicism being detected Fewer variants detected and so data easier to interpret As analysis is restricted to known genes, the likelihood of a variant being pathogenic is greatly increased Will only detect variation in genes known to cause a given condition Difficult to add new genes to the panel as they are discovered normal variants to identify the single (or, rarely, several) pathogenic, disease-causing mutation. While this can, to an extent, be achieved through the application of complex algorithms, these take time and considerable expertise to develop and are not infallible. Furthermore, even after these data have been sifted by bioinformaticians, it is highly likely that clinicians will be left with some variants for which there are insufficient data to enable their definitive categorisation as either pathogenic or non-pathogenic. This may be because we simply do not know enough about the gene, because the particular variant has not previously been reported and/or it is identified in an unaffected parent. These variants must be interpreted with caution and, more usually,

Fig. 3.13 Sequencing by synthesis as used in the Illumina system. (1) Library preparation: DNA is fragmented and specialist adapters are ligated to the fragmented ends. (2) Cluster amplification: the library is loaded to a flow cell and the adapters hybridise to the flow-cell surface. Each bound fragment is hybridised. (3) Sequencing. (4) Alignment and variant interpretation: reads are aligned to a reference sequence using complex software and differences between reference and case genomes are identified. CCGATATCTAGCTTA ATATCTAGC CG TAGC TATCTAGC CCG TAGCTAGCTTA 1 Library preparation 2 Cluster amplification Genomic DNA Fragmentation Adapter ligation Flow cell Amplification 3 Sequencing Reads Reference genome 4 Alignment and variant interpretation G T A C A A 3.10 Next-generation sequencing methods Sequencing by synthesis (Fig. 3.13) • The most

frequently used NGS method • Used in Illumina systems (commonly used in diagnostic laboratories) • Uses fluorescently labelled, terminator nucleotides that are sequentially incorporated into a growing DNA chain • Library DNA samples (fragmented DNA flanked by DNA adapter sequences) are anchored to a flow cell by hybridisation of the DNA adapter sequence to probes on the flow-cell surface • Amplification occurs by washing the flow cell in a mixture containing all four fluorescently labelled terminator nucleotides: A, C, T and G • Once the nucleotide, complementary to the first base of the DNA template, is incorporated, no further nucleotides can be added until the mixture is washed away • The nucleotide terminator is shed and the newly incorporated nucleotide reverts to a regular, non-fluorescent nucleotide that can be extended • The process is then repeated with the incorporation of a second base etc. • Sequencing by synthesis is therefore space- and time-dependent: a sensor will detect the order of fluorescent emissions for each spot on the plate (representing the cluster) and determine the sequence for that read Sequencing by ligation • Used in SOLiD systems • Uses DNA ligase rather than DNA polymerase (as is used in sequencing by synthesis) and short oligonucleotides (as opposed to single nucleotides) • Library DNA samples are washed in a mixture containing oligonucleotide probes representing 4–16 dinucleotide sequences. Only one nucleotide in the probe is fluorescently labelled • The complementary oligo probes will hybridise, using DNA ligase, to the target sequence, initially at a primer annealed to the anchor site and then progressively along the DNA strand • After incorporation of each probe, fluorescence is measured and the dye is cleaved off • Eventually, a new strand is synthesised (composed of a series of the oligo probes) • A new strand is then synthesised but is offset by one nucleotide • The process is repeated a number of times (5 rounds in the SOLiD system), providing overlapping templates that are analysed and a composite of the target sequence determined Ion semiconductor sequencing • When a nucleotide is incorporated into a growing DNA strand, a hydrogen ion is released that can be detected by an alteration in the pH of the solution. This hydrogen ion release forms the basis of ion semiconductor sequencing • Each amplified DNA cluster is located above a semiconductor transistor, capable of detecting differences in the pH of the solution • The DNA cluster is washed in a mixture containing only one type of nucleotide • If the correct nucleotide, complementary to the next base on the DNA template, is in the mixture and incorporated, a hydrogen ion is released and detected • If a homopolymer (sequence of two or more identical nucleotides) is present, this will be detected as a decrease in pH proportionate to the number of identical nucleotides in the sequence initially problematic). The current NGS challenge is to detect large deletions or duplications spanning several hundreds or thousands of bases and therefore exceeding any single read. Increasingly, however, this dosage analysis is being achieved using sophisticated computational methods, negating the need for more traditional technologies such as array CGH. Additional potential uses of NGS include detection of balanced and unbalanced translocations and mosaicism: NGS has proved remarkably sensitive at detecting the latter when there is high read coverage for a given region. Of note, however, NGS is still not able to interrogate the epigenome (and so will not identify conditions caused by a disruption of imprinting, such as Beckwith–Wiedemann, Silver–Russell, Angelman’s and Prader–Willi syndromes) and will not detect triplet repeat expansions such as those that cause Huntington’s disease,

 regions of the genome, and therefore genes, more strongly associated with a given SNP profile and therefore more likely to contribute to the disease under study. Genomics and obstetrics Prenatal genetic testing may be performed where a pregnancy is considered at increased risk of being affected with a genetic condition, either because of the

ultrasound/biochemical screening results or because of the family history. While invasive tests, such as amniocentesis and chorionic villus sampling, have been the mainstay of prenatal diagnosis for many years, they are increasingly being superseded by non-invasive testing of cell-free fetal DNA (cffDNA), originating from placental trophoblasts and detectable in the maternal circulation from 4–5 weeks' gestation; it is present in sufficient quantities for testing by 9 weeks. • Non-invasive prenatal testing (NIPT): the sequencing and quantification, using NGS, of cffDNA chromosome-specific DNA sequences to identify trisomy 13, 18 or 21. The accuracy of NIPT in detecting pregnancy-specific aneuploidy approaches 98%. A false-negative result can occur when there is too little cffDNA (possibly due to early gestation or high maternal body mass index) or when aneuploidy has arisen later in development and is confined to the embryo and not represented in the placenta. False positives can occur with confined placental mosaicism (describing aneuploidy in the placenta, not the fetus) or with an alternative cause of aneuploidy in the maternal circulation, such a cell-free tumour DNA. • Non-invasive prenatal diagnosis (NIPD): the identification of a fetal single-gene defect that either has been paternally inherited or has arisen de novo and so is not identifiable in the maternal genome. Examples of conditions that are currently amenable to NIPD include achondroplasia and the craniosynostoses. Increasingly, however, NIPD is being used for autosomal recessive conditions such as cystic fibrosis, where parents are carriers for different mutations. The free fetal DNA is tested to see whether the paternal mutation is identified and, if absent, the fetus is not affected. If the paternal mutation is identified, however, a definitive invasive test is required to determine whether the maternal mutation has also been inherited and the fetus is affected. Where a genetic diagnosis is known in a family, a couple may opt to undertake pre-implantation genetic diagnosis (PGD). PGD is used as an adjunct to in vitro fertilisation and involves the genetic testing of a single cell from a developing embryo, prior to implantation. Genomics and oncology Until recently, individuals were stratified to genetic testing if they presented with a personal and/or family history suggestive of an inherited cancer predisposition syndrome (Box 3.11). Relevant clinical information included the age of cancer diagnosis and number/type of tumours. For example, the diagnosis of bilateral breast cancer in a woman in her thirties with a mother who had ovarian cancer in her forties is suggestive of BRCA1/2-associated familial breast/ovarian cancer. In many familial cancer syndromes, somatic mutations act together with an inherited mutation to cause specific cancers (p. 50). Familial cancer syndromes may be due to germ-line loss-of-function mutations in tumour suppressor genes encoding DNA repair enzymes or proto-oncogenes. At the cellular level, loss of one copy of a tumour suppressor myotonic dystrophy and fragile X syndrome (see Boxes 3.8 and 3.2). Third-generation sequencing Increasingly, third-generation or single-molecule sequencing is entering the diagnostic arena. As with next- or second-generation sequencing, a number of different platforms are commercially available. One of the most successful is SMRT technology (single-molecule sequencing in real time), developed by Pacific Biosciences. This system utilises a single-stranded DNA molecule (as compared to the amplified clusters used in NGS), which acts as a template for the sequential incorporation, using a polymerase, of fluorescently labelled nucleotides. As each complementary nucleotide is added, the fluorescence (and therefore the identity of the nucleotide) is recorded before it is removed and another nucleotide is added. A key advantage of third-generation sequencing is the long length of the read it generates: in the region of 10–15 kilobases. It is also cheaper than NGS, as fewer reagents are required. Given these inherent advantages, third-generation sequencing is likely to supersede NGS in the near future. Given the confusion surrounding the terminology of NGS and third-generation sequencing, these technologies are increasingly referred to as 'massively parallel sequencing'. Genomics and clinical practice

Genomics and health care Genomics in rare neurodevelopmental disorders Although, by definition, the diagnosis of a rare disorder is made infrequently, rare diseases, when considered together, affect about 3 million people in the UK, the majority of whom are children. NGS has transformed the ability to diagnose individuals affected by a rare disease. Whereas previously, when we were restricted to the sequential analysis of single genes, a clinician would need to make a clinical diagnosis in order to target testing, NGS allows the interrogation of multiple genes in a single experiment. This might be done through a gene panel, a clinical exome or an exome (see Box 3.9 and p. 53), and has increased the diagnostic yield in neurodevelopmental disorders to approximately 30%. Not only does the identification of the genetic cause of a rare disorder potentially provide families with answers, prognostic information and the opportunity to meet and derive support from other affected families but also it can provide valuable information for those couples planning further children and wishing to consider prenatal testing in the future. Genomics and common disease Most common disorders are determined by interactions between a number of genes and the environment. In this situation, the genetic contribution to disease is termed polygenic. Until recently, very little progress had been made in identifying the genetic variants that predispose to common disorders, but this has been changed by the advent of genome-wide association studies. A GWAS typically involves genotyping many (> 500 000) genetic markers (SNPs) spread across the genome in a large group of individuals with the disease and in controls. By comparing the SNP genotypes in cases and controls, it is possible to identify

ever) in some members of these cancer-prone families. In DNA repair diseases, the inherited mutations increase the somatic mutation rate. Autosomal dominant mutations in genes encoding components of specific DNA repair systems are relatively common causes of familial colon cancer and breast cancer (e.g. BRCA1). Increasingly, genetics is moving into the mainstream, becoming integrated into routine oncological care as new gene-specific treatments are introduced. Testing for a genetic predisposition gene does not have any functional consequences, as the cell is protected by the remaining normal copy. However, a somatic mutation affecting the normal allele is likely to occur in one cell at some point during life, resulting in complete loss of tumour suppressor activity and a tumour developing by clonal expansion of that cell. This two-hit mechanism (one inherited, one somatic) for cancer development is known as the Knudson hypothesis. It explains why tumours may not develop for many years (or 3.11 Inherited cancer predisposition syndromes Syndrome name Gene Associated cancers Additional clinical features Birt–Hogg–Dubé syndrome FLCN Renal tumour (oncocytoma, chromophobe (and mixed), renal cell carcinoma) Fibrofolliculoma Trichodiscoma Pulmonary cysts Breast/ovarian hereditary susceptibility BRCA1 BRCA2 Breast carcinoma Ovarian carcinoma Pancreatic carcinoma Prostate carcinoma Cowden’s syndrome PTEN Breast carcinoma Thyroid carcinoma Endometrial carcinoma Macrocephaly Intellectual disability/autistic spectrum disorder Trichilemmoma Acral keratosis Papillomatous papule Thyroid cyst Lipoma Haemangioma Intestinal hamartoma Gorlin’s syndrome/basal cell naevus syndrome PTCH1 Basal cell carcinoma Medulloblastoma Odontogenic keratocyst Palmar or plantar pits Falx calcification Rib abnormalities (e.g. bifid, fused or missing ribs) Macrocephaly Cleft lip/palate Li–Fraumeni syndrome TP53 Sarcoma (e.g. osteosarcoma, chondrosarcoma, rhabdomyosarcoma) Breast carcinoma Brain cancer (esp. glioblastoma) Adrenocortical carcinoma Brain Lynch’s syndrome/ hereditary non-polyposis colon cancer MLH1 MSH2 MSH6 PMS2 Colorectal carcinoma (majority right-sided) Endometrial carcinoma Gastric carcinoma Cholangiocarcinoma Ovarian

carcinoma (esp. mucinous) Multiple endocrine neoplasia 1 MEN1 Parathyroid tumour Endocrine pancreatic tumour Anterior pituitary tumour Lipoma Facial angiofibroma Multiple endocrine neoplasia 2 and 3 (also known as 2a and 2b, respectively) RET Medullary thyroid tumour Phaeochromocytoma Parathyroid tumour Polyposis, familial adenomatous (FAP) APC Colorectal adenocarcinoma (FAP is characterised by thousands of polyps from the second decade; without colectomy, malignant transformation of at least one of these polyps is inevitable) Duodenal carcinoma Hepatoblastoma Desmoid tumour Congenital hypertrophy of the retinal pigment epithelium (CHRPE) Polyposis, MYH-associated MYH (MUTYH) Colorectal adenocarcinoma Duodenal adenocarcinoma Retinoblastoma, familial RB1 Retinoblastoma Osteosarcoma

58 • CLINICAL GENETICS azathioprine, a drug that is used in the treatment of autoimmune diseases and in cancer chemotherapy. Genetic screening for polymorphic variants of TPMT can be useful in identifying patients who have increased sensitivity to the effects of azathioprine and who can be treated with lower doses than normal. Gene therapy and genome editing Replacing or repairing mutated genes (gene therapy) is challenging in humans. Retroviral-mediated ex vivo replacement of the defective gene in bone marrow cells for the treatment of severe combined immune deficiency syndrome (p. 79) has been successful. The major problems with clinical use of virally delivered gene therapy have been oncogenic integration of the exogenous DNA into the genome and severe immune response to the virus. Other therapies for genetic disease include PTC124, a compound that can ‘force’ cells to read through a mutation that results in a premature termination codon in an ORF with the aim of producing a near-normal protein product. This therapeutic approach could be applied to any genetic disease caused by nonsense mutations. The most exciting development in genetics for a generation has been the discovery of accurate, efficient and specific techniques to enable editing of the genome in cells and organisms. This technology is known as CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats and CRISPR-associated) genome editing. It is likely that ex vivo correction of genetic disease will become commonplace over the next few years. In vivo correction is not yet possible and will take much longer to become part of clinical practice. Induced pluripotent stem cells and regenerative medicine Adult stem-cell therapy has been in wide use for decades in the form of bone marrow transplantation. The identification of adult stem cells for other tissues, coupled with the ability to purify and maintain such cells in vitro, now offers exciting therapeutic potential for other diseases. It was recently discovered that many different adult cell types can be trans-differentiated to form cells (induced pluripotent stem cells or iPS cells) with almost all the characteristics of embryonal stem cells derived from the early blastocyst. In mammalian model species, such cells can be taken and used to regenerate differentiated tissue cells, such as in heart and brain. They have great potential both for the development of tissue models of human disease and for regenerative medicine. Pathway medicine The ability to manipulate pathways that have been altered in genetic disease has tremendous therapeutic potential for Mendelian disease, but a firm understanding of both disease pathogenesis and drug action at a biochemical level is required. An exciting example has been the discovery that the vascular pathology associated with Marfan’s syndrome is due to the defective fibrillin molecules causing up-regulation of transforming growth factor (TGF)-β signalling in the vessel wall. Losartan is an antihypertensive drug that is marketed as an angiotensin II receptor antagonist. However, it also acts as a partial antagonist of TGF-β signalling and is effective in preventing aortic dilatation in a mouse model of Marfan’s syndrome, showing promising effects in early human clinical trials. to cancer is therefore moving from the domain of clinical genetics, where it has informed diagnosis, cascade treatment and screening/prophylactic

management, to oncology, where it is informing the immediate management of the patient following cancer diagnosis. This is exemplified by BRCA1 and BRCA2 (BRCA1/2)-related breast cancer. Previously, women with a mutation in either the BRCA1 or BRCA2 gene would have received similar first-line chemotherapy to women with a sporadic breast cancer without a known genetic association. More recently, it has been shown that BRCA1/2 mutation-positive tumours are sensitive to poly ADP ribose polymerase (PARP) inhibitors. PARP inhibitors block the single-strand break-repair pathway. In a BRCA1/2 mutation-positive tumour – with compromised double-strand break repair – the additional loss of the singlestrand break-repair pathway will drive the cell towards apoptosis. Indeed, PARP inhibitors have been shown to be so effective at destroying BRCA1/2 mutation-positive tumour cells, and with such minimal side-effects, that BRCA1/2 gene testing is increasingly determining patient management. It is likely, with a growing understanding of the genomic architecture of tumours, increasing accessibility of NGS and an expanding portfolio of gene-directed therapies, that testing for many of the other inherited cancer susceptibility genes will, in time, move into the mainstream. Genomics in infectious disease NGS technologies are also transforming infectious disease. Given that a microbial genome can be sequenced within a single day at a current cost of less than 100 US dollars, microbiologists are able to identify a causative microorganism and target effective treatment rapidly and accurately. Moreover, microbial genome sequencing enables the effective surveillance of infections to reduce and prevent transmission. Finally, an understanding of the microbial genome will drive the development of vaccines and antibiotics, essential in an era characterised by increasing microbial resistance to established antibiotic agents. Treatment of genetic disease Pharmacogenomics Pharmacogenomics is the science of dissecting the genetic determinants of drug kinetics and effects using information from the human genome. For more than 50 years, it has been appreciated that polymorphic mutations within genes can affect individual responses to some drugs, such as loss-of-function mutations in CYP2D6 that cause hypersensitivity to debrisoquine, an adrenergic-blocking medication formerly used for the treatment of hypertension, in 3% of the population. This gene is part of a large family of highly polymorphic genes encoding cytochrome P450 proteins, mostly expressed in the liver, which determine the metabolism of a host of specific drugs. Polymorphisms in the CYP2D6 gene also determine codeine activation, while those in the CYP2C9 gene affect warfarin inactivation. Polymorphisms in these and other drug metabolic genes determine the persistence of drugs and, therefore, should provide information about dosages and toxicity. With the increasing use of NGS, genetic testing for assessment of drug response is seldom employed routinely, but in the future it may be possible to predict the best specific drugs and dosages for individual patients based on genetic profiling: so-called ‘personalised medicine’. An example is the enzyme thiopurine methyltransferase (TPMT), which catabolises

BRCA1/2 mutations), DCT is undertaken in isolation with no direct access to professional support. Furthermore, in addition to some (common) single-gene mutations, such as the founder BRCA1/2 mutations frequently identified in the Ashkenazi Jewish population and discussed in this example, current DCT packages utilise a series of SNPs to determine an overall risk profile; they evaluate the number of detrimental and protective SNPs for a given disease. However, given that only a minority of the risk SNPs have so far been characterised, this is often inaccurate. Individuals may be falsely reassured that they are not at increased risk of a genetic condition despite a family history suggesting otherwise, resulting in inadequate surveillance and/or management. The ethical

considerations listed in this clinical scenario give just a flavour of some of the issues frequently encountered in clinical genetics. They are not meant to be an exhaustive summary and whole textbooks and meetings are devoted to the discussion of hugely complex ethical issues in genetics. However, a guiding principle is that, although each counselling situation will be unique with specific communication and ethical challenges, a genetic result is permanent and has implications for the whole family, not just the individual. Where possible, therefore, an informed decision regarding genetic testing should be taken by a competent adult following counselling by an experienced and appropriately trained clinician. Further information Books and journal articles Alberts B, Bray D, Hopkin K et al. Essential cell biology, 4th edn. New York: Garland Science; 2013. Firth H, Hurst JA. Oxford desk reference: clinical genetics. Oxford: Oxford University Press; 2005. Read A, Donnai D. New clinical genetics, 2nd edn. Banbury: Scion; 2010. Strachan T, Read A. Human molecular genetics, 4th edn. New York: Garland Science; 2010. Websites bsgm.org.uk British Society for Genetic Medicine; has a report on genetic testing of children. decipher.sanger.ac.uk Excellent, comprehensive genomic database. ensembl.org Annotated genome databases from multiple organisms. futurelearn.com/courses/the-genomics-era Has a Massive Open Online Course on genomics, for which one of the authors of the current chapter is the lead educator. genome.ucsc.edu Excellent source of genomic information. ncbi.nlm.nih.gov Online Mendelian Inheritance in Man (OMIM). ncbi.nlm.nih.gov/books/NBK1116/ Gene Reviews: excellent US-based source of information about many rare genetic diseases. orpha.net/consor/cgi-bin/index.php Orphanet: European-based database on rare disease. Ethics in a genomic age As genomic technology is increasingly moving into mainstream clinical practice, it is essential for clinicians from all specialties to appreciate the complexities of genetic testing and consider whether genetic testing is the right thing to do in a given clinical scenario. To exemplify the ethical considerations associated with genetic testing, it may be helpful to think about them in the context of a clinical scenario. As you read the scenario, try to think what counselling/ethical issues might arise. A 32-year-old woman is referred to discuss BRCA2 testing; she is currently pregnant with her second child (she already has a 2-year-old daughter) and has an identical twin sister. Her mother, a healthy 65-year-old with Ashkenazi Jewish ancestry, participated in direct-to-consumer testing (DCT) for ‘a bit of fun’ and a BRCA2 mutation – common in the Ashkenazi Jewish population – was identified. There is no significant cancer family history of note. Consider the following issues: • Pre-symptomatic/predictive testing: this describes testing for a known familial gene mutation in an unaffected individual (compared with diagnostic testing, where genetic testing is undertaken in an affected individual). Although this could be considered for the unaffected patient, in the current scenario any testing would also have implications for her identical twin sister. This needs to be fully explored with the patient and her sister prior to testing. There is also the potential issue of predictive testing in the patient’s first child. A fundamental tenet in clinical genetics is that predictive genetic testing should be avoided in childhood for adult-onset conditions. This is because, if no benefit to the patient is accrued through childhood testing, it is better to retain the child’s right to decide for herself, when she is old enough, whether she wishes to participate in genetic testing or not. • Prenatal testing: the principles behind predictive genetic testing in childhood can be extended to prenatal testing, i.e. if a pregnancy is being continued, a baby should not be tested for an adult-onset condition that cannot be prevented or treated in childhood. However, prenatal testing itself is hugely controversial and there is much debate as to how severe a condition should be to justify prenatal diagnosis, which would determine ongoing pregnancy decisions. • DCT: while DCT can be interesting and empowering for individuals wishing to find out more about their genetic backgrounds, it also has several drawbacks. Perhaps the main one is that,

unlike face-to-face genetic counselling (which usually precedes any genetic testing, certainly where there are serious health implications for the individual and their family, such as is associated with

7KLVSDJHLQWHQWLRQDOO\OHIWEODQN

04-4 Clinical immunology

4 Clinical immunology

62 • CLINICAL IMMUNOLOGY to protect against infection (Fig. 4.1). Immune defences are normally categorised into the innate immune response, which provides immediate protection against an invading pathogen, and the adaptive or acquired immune response, which takes more time to develop but confers exquisite specificity and long-lasting protection. The innate immune system Innate defences against infection include anatomical barriers, phagocytic cells, soluble molecules such as complement and acute phase proteins, and natural killer cells. The innate immune system recognises generic microbial structures present on non-mammalian tissue and can be mobilised within minutes. A specific stimulus will elicit essentially identical responses in different individuals, in contrast with adaptive antibody and T-cell responses, which vary greatly between individuals. Physical barriers The tightly packed keratinised cells of the skin physically limit colonisation by microorganisms. The hydrophobic oils that are secreted by sebaceous glands further repel water and microorganisms, and microbial growth is inhibited by the skin’s low pH and low oxygen tension. Sweat also contains lysozyme, an enzyme that destroys the structural integrity of bacterial cell walls; ammonia, which has antibacterial properties; and several The immune system has evolved to identify and destroy pathogens while minimising damage to host tissue. Despite the ancient observation that recovery from an infectious disease frequently results in protection against that condition, the existence of the immune system as a functional entity was not recognised until the end of the 19th century. More recently, it has become clear that the immune system not only protects against infection but also regulates tissue repair following injury, and when dysregulated, governs the responses that can lead to autoimmune and auto-inflammatory diseases. Dysfunction or deficiency of the immune response can lead to a wide variety of diseases that may potentially involve every organ system in the body. The aim of this chapter is to provide a general understanding of the immune system, how it contributes to human disease and how manipulation of the immune system can be put to therapeutic use. A review of the key components of the immune response is followed by sections that illustrate the clinical presentation of the most common forms of immune dysfunction: immune deficiency, inflammation, autoimmunity and

allergy. More detailed discussion of individual conditions can be found in the relevant organ-specific chapters of this book. Functional anatomy and physiology The immune system consists of an intricately linked network of lymphoid organs, cells and proteins that are strategically placed Fig. 4.1 Anatomy of the immune system. Adenoids Lymph nodes Tonsils Thoracic duct Spleen Bone marrow Liver Appendix Germinal centre Proliferating B cells after antigen exposure Afferent lymph Paracortex T cells Dendritic cells Cortex B cells in primary lymphoid follicles Efferent lymph Medulla Plasma cells Sinuses with macrophages Blood vessels Capsule Lymph node section Lymphatics Neutrophil Eosinophil Cells of the innate immune system Natural killer cell Basophil Mast cell Monocyte Macrophage T lymphocyte Cells of the adaptive immune system Antigen-presenting cell B lymphocyte Thymus Peyer’s patches in small intestine

the specific soluble proteins and cells of the innate immune system are activated. Phagocytes Phagocytes (‘eating cells’) are specialised cells that ingest and kill microorganisms, scavenge cellular and infectious debris, and produce inflammatory molecules that regulate other components of the immune system. They include neutrophils, monocytes and macrophages, and are particularly important for defence against bacterial and fungal infections. Phagocytes express a wide range of surface receptors, including pattern recognition receptors (PRRs), which recognise pathogen-associated molecular patterns (PAMPs) on invading microorganisms, allowing their identification. The PRRs include Toll-like receptors, nucleotide oligomerisation domain (NOD) protein-like receptors and mannose receptors, whereas the PAMPs they recognise are molecular motifs not present on mammalian cells, including bacterial cell wall components, bacterial DNA and viral double-stranded RNA. While phagocytes can recognise microorganisms through PRRs alone, engulfment of microorganisms is greatly enhanced by opsonisation. Opsonins include acute phase proteins produced by the liver, such as C-reactive protein and complement. Antibodies generated by the adaptive immune system also act as opsonins. They bind both to the pathogen and to phagocyte receptors, acting as a bridge between the two to facilitate phagocytosis (Fig. 4.2). This is followed by intracellular pathogen destruction and downstream activation of pro-inflammatory genes, resulting in the generation of pro-inflammatory cytokines as discussed below. antimicrobial peptides such as defensins. Similarly, the mucous membranes of the respiratory, gastrointestinal and genitourinary tracts provide a physical barrier to infection. Secreted mucus traps invading pathogens, and immunoglobulin A (IgA), generated by the adaptive immune system, prevents bacteria and viruses attaching to and penetrating epithelial cells. As in the skin, lysozyme and antimicrobial peptides within mucosal membranes directly kill invading pathogens, and lactoferrin acts to starve invading bacteria of iron. Within the respiratory tract, cilia directly trap pathogens and contribute to removal of mucus, assisted by physical manœuvres such as sneezing and coughing. In the gastrointestinal tract, hydrochloric acid and salivary amylase chemically destroy bacteria, while normal peristalsis and induced vomiting or diarrhoea assist clearance of invading organisms. The microbiome, which is made up of endogenous commensal bacteria, provides an additional constitutive defence against infection. Approximately 1014 bacteria normally reside at epithelial surfaces in symbiosis with the human host (p. 102). They compete with pathogenic microorganisms for scarce resources, including space and nutrients, and produce fatty acids and bactericidins that inhibit the growth of many pathogens. In addition, recent research has demonstrated that commensal bacteria help to shape the immune response by inducing specific regulatory T cells within the intestine. Eradication of the normal flora with broad-spectrum

antibiotics commonly results in opportunistic infection by organisms such as Clostridium difficile, which rapidly colonise an undefended ecological niche. These constitutive barriers are highly effective, but if external defences are breached by a wound or pathogenic organism, Fig. 4.2 Phagocytosis and opsonisation. Phagocytosis of microbes can be augmented by several opsonins, such as C-reactive protein, antibodies and complement fragments like C3b, which enhance the ability of phagocytic cells to engulf microorganisms and destroy them. Phagocytes also recognise components of microbes, such as lipopolysaccharide, peptidoglycans, DNA and RNA, collectively as pathogen-associated molecular patterns (PAMPs). These activate pattern recognition receptors (PRRs), such as Toll-like receptors and nucleotide oligomerisation domain (NOD)-like receptors, which promote inflammatory gene expression through the nuclear factor kappa beta (NFκB) pathway. Uric acid and other crystals can also promote inflammation through the NOD pathway. Microbes C3b Antibody C-reactive protein Fc receptor Toll-like receptors NOD-like receptors Lipopolysaccharide Bacterial DNA Bacterial RNA Peptidoglycans Crystals NFκB NFκB Lysosome C3b receptor Phagocytic cell Pro-inflammatory gene expression Response genes

 constitute about 5% of leucocytes. From the blood stream they migrate to peripheral tissues, where they differentiate into tissue macrophages and reside for long periods. Specialised populations of tissue macrophages include Kupffer cells in the liver, alveolar macrophages in the lung, mesangial cells in the kidney, and microglial cells in the brain. Macrophages, like neutrophils, are capable of phagocytosis and killing of microorganisms but also play an important role in the amplification and regulation of the inflammatory response (Box 4.1). They are particularly important in tissue surveillance and constantly survey their immediate surroundings for signs of tissue damage or invading organisms. Dendritic cells Dendritic cells are specialised antigen-presenting cells that are present in tissues in contact with the external environment, such as the skin and mucosal membranes. They can also be found in an immature state in the blood. They sample the environment for foreign particles and, once activated, carry microbial antigens to regional lymph nodes, where they interact with T cells and B cells to initiate and shape the adaptive immune response. Cytokines Cytokines are signalling proteins produced by cells of the immune system and a variety of other cell types. More than 100 have been identified. Cytokines have complex and overlapping roles in cellular communication and regulation of the immune response. Subtle differences in cytokine production, particularly at the initiation of an immune response, can have a major impact on outcome. Cytokines bind to specific receptors on target cells and activate downstream intracellular signalling pathways, ultimately leading to changes in gene transcription and cellular function. Two important signalling pathways are illustrated in Figure 4.3. The nuclear factor kappa B (NFκB) pathway is activated by tumour necrosis factor (TNF), by other members of the TNF superfamily such as receptor activator of nuclear kappa B ligand Neutrophils Neutrophils, also known as polymorphonuclear leucocytes, are derived from the bone marrow and circulate freely in the blood. They are short-lived cells with a half-life of 6 hours, and are produced at the rate of 1011 cells daily. Their functions are to kill microorganisms, to facilitate rapid transit of cells through tissues, and to amplify the immune response non-specifically. These functions are mediated by enzymes contained in granules, which also provide an intracellular milieu for the killing and degradation of microorganisms. Two main types of granule are recognised: primary or azurophil granules, and the more numerous secondary or specific granules. Primary granules contain myeloperoxidase and other enzymes important for intracellular killing and digestion of ingested microbes. Secondary granules are smaller and contain lysozyme, collagenase and lactoferrin, which can be released into the extracellular space. Enzyme

production is increased in response to infection, which is reflected by more intense granule staining on microscopy, known as ‘toxic granulation’. Changes in damaged or infected cells trigger local production of inflammatory molecules and cytokines. These cytokines stimulate the production and maturation of neutrophils in the bone marrow, and their release into the circulation. Neutrophils are recruited to specific sites of infection by chemotactic agents, such as interleukin 8 (IL-8), and by activation of local endothelium. Up-regulation of cellular adhesion molecules on neutrophils and the endothelium also facilitates neutrophil migration. The transit of neutrophils through the blood stream is responsible for the rise in neutrophil count that occurs in early infection. Once present within infected tissue, activated neutrophils seek out and engulf invading microorganisms. These are initially enclosed within membrane-bound vesicles, which fuse with cytoplasmic granules to form the phagolysosome. Within this protected compartment, killing of the organism occurs through a combination of oxidative and non-oxidative killing. Oxidative killing, also known as the respiratory burst, is mediated by the nicotinamide adenine dinucleotide phosphate (NADPH)–oxidase enzyme complex, which converts oxygen into reactive oxygen species such as hydrogen peroxide and superoxide that are lethal to microorganisms. The myeloperoxidase enzyme within neutrophils produces hypochlorous acid, which is a powerful oxidant and antimicrobial agent. Non-oxidative (oxygen-independent) killing occurs through the release of bactericidal enzymes into the phagolysosome. Each enzyme has a distinct antimicrobial spectrum, providing broad coverage against bacteria and fungi. An additional, recently identified form of neutrophil-mediated killing is neutrophil extracellular trap (NET) formation. Activated neutrophils can release chromatin with granule proteins such as elastase to form an extracellular matrix that binds to microbial proteins. This can immobilise or kill microorganisms without requiring phagocytosis. The process of phagocytosis and NET formation (NETosis) depletes neutrophil glycogen reserves and is followed by neutrophil death. As the cells die, their contents are released and lysosomal enzymes degrade collagen and other components of the interstitium, causing liquefaction of closely adjacent tissue. The accumulation of dead and dying neutrophils results in the formation of pus, which, if extensive, may lead to abscess formation. Monocytes and macrophages Monocytes are the precursors of tissue macrophages. They are produced in the bone marrow and enter the circulation, where they 4.1 Functions of macrophages Amplification of the inflammatory response • Stimulate the acute phase response (through production of IL-1 and IL-6) • Activate vascular endothelium (IL-1, TNF-α) • Stimulate neutrophil maturation and chemotaxis (IL-1, IL-8) • Stimulate monocyte chemotaxis Killing of microorganisms • Phagocytosis • Microbial killing through oxidative and non-oxidative mechanisms Clearance, resolution and repair • Scavenging of necrotic and apoptotic cells • Clearance of toxins and other inorganic debris • Tissue remodelling (elastase, collagenase, matrix proteins) • Down-regulation of inflammatory cytokines • Wound healing and scar formation (IL-1, platelet-derived growth factor, fibroblast growth factor) Link between innate and adaptive immune systems • Activate T cells by presenting antigen in a recognisable form • T cell-derived cytokines increase phagocytosis and microbicidal activity of macrophages in a positive feedback loop (IL = interleukin; TNF = tumour necrosis factor)

IKK, which in turn leads to phosphorylation of the inhibitor of nuclear factor kappa B protein (IκB), causing it to be degraded, and allowing NFκB to translocate to the nucleus and activate gene transcription. The Janus kinase/signal transducers and activators of transcription (JAK-STAT) pathway is involved in (RANKL; p. 985), and by the Toll-like receptors and NOD-like receptors (see

Fig. 4.2). In the case of TNF superfamily members, receptor binding causes the inhibitor of kappa B kinase (IKK) complex of three proteins to be recruited to the receptor by binding TNF receptor-associated proteins (TRAF). This activates Fig. 4.3 Cytokines signalling pathways and the immune response. Cytokines regulate the immune response through binding to specific receptors that activate a variety of intracellular signalling pathways, two of which are shown. Members of the tumour necrosis factor (TNF) superfamily and the Toll-like receptors and NOD-like receptors (Fig. 4.2) signal through the nuclear factor kappa B (NFκB) pathway. Several other cytokines, including interleukin-2 (IL-2), IL-6 and interferons, employ the Janus kinase/ signal transducer and activator of transcription (JAK-STAT) pathway to regulate cellular function (see text for more details). (IκB = inhibitor of kappa B; IKK = I kappa B kinase; P = phosphorylation of the signalling protein; TRAF = tumour necrosis factor receptorassociated factor) JAK JAK inhibitor Response genes Response genes JAK Cytokines Cytokine receptor P P P P P STAT STAT P P STAT STAT DNA TNF TNF receptor TRAF IκB P IκB NFκB NFκB IKKκ IKKα IKKβ IFN-γ IL-6 IL-2 4.2 Important cytokines in the regulation of the immune response Cytokine Source Actions Biologic therapies Interferon-alpha (IFN-α) T cells and macrophages Antiviral activity Activates NK cells, CD8+ T cells and macrophages Recombinant IFN-α used in hepatitis C and some malignancies Interferon-gamma (IFN-γ) T cells and NK cells Increases antimicrobial activity of macrophages Regulates cytokine production by T cells and macrophages Used in chronic granulomatous disease Tumour necrosis factor alpha (TNF-α) Macrophages, NK cells and others, including T cells Pro-inflammatory Increases expression of other cytokines and adhesion molecules Causes apoptosis of some target cells Directly cytotoxic TNF-α inhibitors used in rheumatoid arthritis, inflammatory bowel disease, psoriasis and many other inflammatory conditions Interleukin-1 (IL-1) Macrophages and neutrophils Stimulates neutrophil recruitment, fever, and T-cell and macrophage activation as part of the inflammatory response IL-1 inhibitors used in systemic juvenile rheumatoid arthritis, periodic fever syndromes and acute gout Interleukin-2 (IL-2) CD4+ T cells Stimulates proliferation and differentiation of antigenspecific T lymphocytes Interleukin-4 (IL-4) CD4+ T cells Stimulates maturation of B and T cells, and production of IgE antibody Antibodies to IL-4 receptor used in severe atopic dermatitis Interleukin-6 (IL-6) Monocytes and macrophages Stimulates neutrophil recruitment, fever, and T-cell and macrophage activation as part of the inflammatory response, stimulates maturation of B cells into plasma cells Antibodies to IL-6 receptor used in rheumatoid arthritis Interleukin-12 (IL-12) Monocytes and macrophages Stimulates IFN-γ and TNF-α release by T cells Activates NK cells Antibody to p40 subunit of IL-12 used in psoriasis and psoriatic arthritis Interleukin-17 (IL-17) Th17 cells (T helper), NK cells, NK-T cells Pro-inflammatory cytokine Involved in mucosal immunity and control of extracellular pathogens, synergy with IL-1 and TNF Antibody to IL-17 used in psoriasis, psoriatic arthritis and ankylosing spondylitis Interleukin-22 (IL-22) Th17 cells Induction of epithelial cell proliferation and antimicrobial proteins in keratinocytes (IgE = immunoglobulin E; NK = natural killer)

 important in the defence against encapsulated bacteria such as Neisseria spp. and Haemophilus influenzae. Complement fragments generated by activation of the cascade can also act as opsonins, rendering microorganisms more susceptible to phagocytosis by macrophages and neutrophils (see Fig. 4.2). In addition, they are chemotactic agents, promoting leucocyte trafficking to sites of inflammation. Some fragments act as anaphylotoxins, binding to complement receptors on mast cells and triggering release of histamine, which increases vascular permeability. The products of complement activation also help to target immune complexes to antigen-presenting cells, providing a link between the innate and the acquired immune systems.

Finally, activated complement products dissolve the immune complexes that triggered the cascade, minimising bystander damage to surrounding tissues. A monoclonal antibody directed against the central complement molecule C5, eculizumab, has been developed for therapeutic use in paroxysmal nocturnal haemoglobinuria and atypical haemolytic uraemic syndromes (p. 408). Invasive infection, including meningococcal sepsis, has been reported with eculizumab therapy, highlighting the importance of the complement system in preventing such infections. Mast cells and basophils Mast cells and basophils are bone marrow-derived cells that play a central role in allergic disorders. Mast cells reside predominantly in tissues exposed to the external environment, such as the skin and gut, while basophils circulate in peripheral blood and are recruited into tissues in response to inflammation. Both contain large cytoplasmic granules that enclose vasoactive substances such as histamine (see Fig. 4.14). Mast cells and Fig. 4.4 The complement pathway. The classical pathway is activated by binding of antigen–antibody complexes to C1 but is blocked by C1 inhibitor (C1inh), whereas mannose-binding lectins, which are macromolecules that bind to various microorganisms, activate the pathway by binding C4. Bacteria can directly activate the pathway through C3, which plays a pivotal role in complement activation through all three pathways. Smooth muscle contraction Activation of cells Vascular permeability Lysis of bacteria Membrane attack complex (MAC) Opsonisation of bacteria Direct activation Lectin pathway Mannosebinding lectin Classical pathway Antibody–antigen complexes Alternate pathway C4 C2 C3 C3a C1inh C5a C5 C3b C5b C6 C7 C8 C1 C9 transducing signals downstream of many cytokine receptors, including those for IL-2, IL-6 and interferon-gamma (IFN-γ). On receptor binding, JAK proteins are recruited to the intracellular portion of the receptor and are phosphorylated. These in turn phosphorylate STAT proteins, which translocate to the nucleus and activate gene transcription, altering cellular function. The function and disease associations of several important cytokines are shown in Box 4.2. Cytokine inhibitors are now routinely used in the treatment of autoimmune diseases, most of which are monoclonal antibodies to cytokines or their receptors. In addition, small-molecule inhibitors have been developed that inhibit the intracellular signalling pathways used by cytokines. These include the Janus kinase inhibitors tofacitinib and baracitinib, which are used in rheumatoid arthritis (p. 1026), and the tyrosine kinase inhibitor imatinib, which is used in chronic myeloid leukaemia (p. 959). Integrins Integrins are transmembrane proteins that play important roles in cell–cell and cell–matrix interactions. They mediate attachment of the cell to the extracellular matrix, signal transduction and cell migration. Their role in autoimmune disease has been extensively studied. Targeted therapy with a recombinant humanised anti-α4 integrin antibody, natalizumab, is an effective treatment for multiple sclerosis, which works by preventing immune cells from traversing the vascular endothelium and entering the central nervous system (p. 1109). Complement The complement system comprises a group of more than 20 tightly regulated, functionally linked proteins that act to promote inflammation and eliminate invading pathogens. Complement proteins are produced in the liver and are present in inactive form in the circulation. When the complement system is activated, it sets in motion a rapidly amplified biological cascade analogous to the coagulation cascade (p. 918). There are three mechanisms by which the complement cascade can be activated (Fig. 4.4): • The alternate pathway is triggered directly by binding of C3 to bacterial cell-wall components, such as lipopolysaccharide of Gram-negative bacteria and teichoic acid of Gram-positive bacteria. • The classical pathway is initiated when two or more IgM or IgG antibody molecules bind to antigen. The associated conformational change exposes binding sites on the antibodies for the first protein in the classical pathway, C1, which is a multiheaded molecule that can bind up to six antibody molecules. Once two or more ‘heads’ of a C1 molecule are bound to antibody, the classical cascade is triggered. An important

inhibitor of the classical pathway is C1 inhibitor (C1inh), as illustrated in Figure 4.4. • The lectin pathway is activated by the direct binding of mannose-binding lectin to microbial cell surface carbohydrates. This mimics the binding of C1 to immune complexes and directly stimulates the classical pathway, bypassing the need for immune complex formation. Activation of complement by any of these pathways results in activation of C3. This in turn activates the final common pathway, in which the complement proteins C5–C9 assemble to form the membrane attack complex (MAC). This can puncture the cell wall, leading to osmotic lysis of target cells. This step is particularly

Lymphoid organs The primary lymphoid organs are involved in lymphocyte development. They include the bone marrow, where T and B lymphocytes differentiate from haematopoietic stem cells (p. 914) and where B lymphocytes also mature, and the thymus, the site of T-cell maturation (see Fig. 4.1). After maturation, lymphocytes migrate to the secondary lymphoid organs. These include the spleen, lymph nodes and mucosa-associated lymphoid tissue. These trap and concentrate foreign substances and are the major sites of interaction between naïve lymphocytes and microorganisms. The thymus The thymus is a bi-lobed structure in the anterior mediastinum, and is organised into cortical and medullary areas. The cortex is densely populated with immature T cells, which migrate to the medulla to undergo selection and maturation. The thymus is most active in the fetal and neonatal period, and involutes after puberty. Failure of thymic development is associated with profound T-cell immune deficiency (p. 79) but surgical removal of the thymus in childhood (usually during major cardiac surgery) is not associated with significant immune dysfunction. The spleen The spleen is the largest of the secondary lymphoid organs. It is highly effective at filtering blood and is an important site of phagocytosis of senescent erythrocytes, bacteria, immune complexes and other debris, and of antibody synthesis. It is important for defence against encapsulated bacteria, and asplenic individuals are at risk of overwhelming Streptococcus pneumoniae and H. influenzae infection (see Box 4.5). Lymph nodes These are positioned to maximise exposure to lymph draining from sites of external contact, and are highly organised (Fig. 4.1) • The cortex contains primary lymphoid follicles, which are the site of B-lymphocyte interactions. When B cells encounter antigen, they undergo intense proliferation, forming germinal centres. • The paracortex is rich in T lymphocytes and dendritic cells. • The medulla is the major site of antibody-secreting plasma cells. • Within the medulla there are many sinuses, which contain large numbers of macrophages. Mucosa-associated lymphoid tissue Mucosa-associated lymphoid tissue (MALT) consists of diffusely distributed lymphoid cells and follicles present along mucosal surfaces. It has a similar function to the more organised, encapsulated lymph nodes. They include the tonsils, adenoids and Peyer’s patches in the small intestine. Lymphatics Lymphoid tissue is connected by a network of lymphatics, with three major functions: it provides access to lymph nodes, returns interstitial fluid to the venous system, and transports fat from the small intestine to the blood stream (see Fig. 14.13, p. 372). The lymphatics begin as blind-ending capillaries, which come together to form lymphatic ducts, entering and leaving regional lymph nodes as afferent and efferent ducts, respectively. They eventually coalesce and drain into the thoracic duct and left subclavian vein. Lymphatics may be either deep or superficial, and follow the distribution of major blood vessels. basophils express IgE receptors on their cell surface, which bind IgE antibody. On encounter with specific antigen, the cell is triggered to release histamine and other mediators present within the granules and to synthesise additional mediators, including

leukotrienes, prostaglandins and cytokines. An inflammatory cascade is initiated that increases local blood flow and vascular permeability, stimulates smooth muscle contraction, and increases secretion at mucosal surfaces. Natural killer cells Natural killer (NK) cells are large granular lymphocytes that play a major role in defence against tumours and viruses. They exhibit features of both the adaptive and the innate immune systems in that they are morphologically similar to lymphocytes and recognise similar ligands, but they are not antigen-specific and cannot generate immunological memory. NK cells express a variety of cell surface receptors, some of which are stimulatory and others inhibitory. The effects of inhibitory receptors normally predominate. These recognise human leucocyte antigen (HLA) molecules that are expressed on normal nucleated cells, preventing NK cell-mediated attack, whereas the stimulatory receptors recognise molecules that are expressed primarily when cells are damaged. This allows NK cells to remain tolerant to healthy cells but not to damaged ones. When cells become infected by viruses or undergo malignant change, expression of HLA class I molecules on the cell surface can be down-regulated. This is an important mechanism by which these cells then evade adaptive T-lymphocyte responses. In this circumstance, however, NK cell defences becomes important, as down-regulation of HLA class I abrogates the inhibitory signals that normally prevent NK activation. The net result is NK attack on the abnormal target cell. NK cells can also be activated by binding of antigen–antibody complexes to surface receptors. This physically links the NK cell to its target in a manner analogous to opsonisation and is known as antibody-dependent cellular cytotoxicity (ADCC). Activated NK cells can kill their targets in various ways. They secrete pore-forming proteins such as perforin into the membrane of the target cell, and proteolytic enzymes called granzymes into the target cell, which cause apoptosis. In addition, NK cells produce a variety of cytokines such as TNF-α and IFN-γ, which have direct antiviral and anti-tumour effects. The adaptive immune system If the innate immune system fails to provide effective protection against an invading pathogen, the adaptive immune system is mobilised (see Fig. 4.1). This has three key characteristics: • It has exquisite specificity and can discriminate between very small differences in molecular structure. • It is highly adaptive and can respond to an almost unlimited number of molecules. • It possesses immunological memory, and changes consequent to initial activation by an antigen allow a more effective immune response on subsequent encounters. There are two major arms of the adaptive immune response. Humoral immunity involves the production of antibodies by B lymphocytes, and cellular immunity involves the activation of T lymphocytes, which synthesise and release cytokines that affect other cells, as well as directly killing target cells. These interact closely with each other and with the components of the innate immune system to maximise effectiveness of the immune response.

68 • CLINICAL IMMUNOLOGY Humoral immunity Humoral immunity is mediated by B lymphocytes, which differentiate from haematopoietic stem cells in the bone marrow. Their major functions are to produce antibody and interact with T cells, but they are also involved in antigen presentation. Mature B lymphocytes can be found in the bone marrow, lymphoid tissue, spleen and, to a lesser extent, the blood stream. They Fig. 4.5 B-cell activation. Activation of B cells is initiated through binding of an antigen with the immunoglobulin receptor on the cell surface. For activation to proceed, an interaction with T-helper cells is also required, providing additional signals through binding of CD40 ligand (CD40L) to CD40; an interaction between the T-cell receptor (TCR) and processed antigenic peptides presented by human leucocyte antigen (HLA) molecules on the B-cell surface; and cytokines released by the T-helper cells. Fully activated B cells undergo clonal expansion with differentiation towards plasma cells that produce antibody. Following activation, memory cells are generated that allow rapid antibody responses when the same antigen is

encountered on a second occasion. (CD = cluster of differentiation; IL = interleukin) T-helper cell Immunoglobulin receptor Antigen CD40L CD40 TCR HLA B-cell activation B cell Clonal expansion Plasma cells Antibodies Memory B cells IL-4 IL-5 Fig. 4.6 The structure of an immunoglobulin (antibody) molecule. The variable region is responsible for antigen binding, whereas the constant region can interact with immunoglobulin receptors expressed on immune cells. Variable region (Fab) Constant region (Fc) Light chain Heavy chain 4.3 Classes and properties of antibody Antibody Concentration in adult serum Complement activation* Opsonisation Presence in external secretions Other properties IgG 6.0–16.0 g/L IgG1 +++ IgG2 + IgG3 +++ IgG1 ++ IgG3 ++ ++ Four subclasses: IgG1, IgG2, IgG3, IgG4 Distributed equally between blood and extracellular fluid, and transported across placenta IgG2 is particularly important in defence against polysaccharides antigens IgA 1.5–4.0 g/L – – ++++ Two subclasses: IgA1, IgA2 Highly effective at neutralising toxins Particularly important at mucosal surfaces IgM 0.5–2.0 g/L ++++ – + Highly effective at agglutinating pathogens IgE 0.003–0.04 g/L – – – Majority of IgE is bound to mast cells, basophils and eosinophils Important in allergic disease and defence against parasite infection IgD Not detected – – – Function in B-cell development *Activation of the classical pathway, also called ‘complement fixation’. express a unique immunoglobulin receptor on their cell surface, the B-cell receptor, which binds to soluble antigen targets (Fig. 4.5). Encounters with antigen usually occur within lymph nodes. If provided with appropriate cytokines and other signals from nearby T lymphocytes, antigen-specific B cells respond by rapidly proliferating in a process known as clonal expansion (Fig. 4.5). This is accompanied by a highly complex series of genetic rearrangements known as somatic hypermutation, which generates B-cell populations that express receptors with greater affinity for antigen than the original. These cells differentiate into either long-lived memory cells, which reside in the lymph nodes, or plasma cells, which produce antibody. Memory cells allow production of a more rapid and more effective response on subsequent exposure to that pathogen. Immunoglobulins Immunoglobulins (Ig) play a central role in humoral immunity. They are soluble proteins produced by plasma cells and are made up of two heavy and two light chains (Fig. 4.6). The heavy chain determines the antibody class or isotype, such as IgG, IgA, IgM, IgE or IgD. Subclasses of IgG and IgA also occur. The antigen is recognised by the antigen-binding regions (Fab) of both heavy and light chains, while the consequences of antibody binding are determined by the constant region of the heavy chain (Fc) (Box 4.3). Antibodies have several functions.

and production of antibody is decreased to 2–3 days, the amount of antibody produced is increased, and the response is dominated by IgG antibodies of high affinity. Furthermore, in contrast to the initial antibody response, secondary antibody responses do not require additional input from T lymphocytes. This allows the rapid generation of highly specific responses on re-exposure to a pathogen and is an important mechanism in vaccine efficacy. Cellular immunity Cellular immunity is mediated by T lymphocytes, which play important roles in defence against viruses, fungi and intracellular bacteria. They also play an important immunoregulatory role, by orchestrating and regulating the responses of other components of the immune system. T-lymphocyte precursors differentiate from haematopoietic stem cells in the bone marrow and are exported to the thymus when they are still immature (see Fig. 4.1). Individual T cells express a unique receptor that is highly Fig. 4.7 T-cell activation. Activation of T cells is initiated when an antigenic peptide bound to a human leucocyte antigen (HLA) molecule on antigenpresenting cells interacts with the T-cell receptor expressed by T lymphocytes. Additional signals are required for T-

cell activation, however. These include binding of the co-stimulatory molecules CD80 and CD86 with CD28 on the T cell, and interleukin 2 (IL-2), which is produced in an autocrine manner by T cells that are undergoing activation. Other molecules are present that can inhibit T-cell activation, however, including cytotoxic T-lymphocyte-associated protein 4 (CTLA4), which competes with CD28 for binding to CD80 and CD86; and PD1, which, by binding PDL1, is also inhibitory. Following activation, T cells proliferate and, depending on their subtype, have various functions with distinct patterns of cytokine production, as indicated. Memory cells are also generated that can mount a rapid immune response on encountering the same antigen. (CD = cluster of differentiation; CD40L = CD40 ligand; IFN-γ = interferon-gamma; IL = interleukin; PD1 = programmed cell death 1; PDL1 = programmed death ligand 1; TGF-β = transforming growth factor beta; TNF-α = tumour necrosis factor alpha) CD80 CD86 CD4+ T cells Th2 cells Cytokines IL-4, IL-5, IL-10, IL-13 Function B-cell activation Eosinophil activation Th1 cells Cytokines TNF-α, IFN-γ, IL-2 Function Pro-inflammatory Th17 cells Cytokines IL-17 Function Mucosal immunity Pro-inflammatory Regulatory T cells Memory T cells Memory T cells Cytokines Fas ligand TNF-α, IFN-γ Function Direct cell killing CD8+ T cells Cytokines IL-10, TGF-β Function Anti-inflammatory PDL1 HLA T-cell receptor T-cell activation Antigenic peptide PD1 CTLA4 CD28 Antigen-presenting cell IL-2 They facilitate phagocytosis by acting as opsonins (see Fig. 4.2) and facilitate cell killing by cytotoxic cells, particularly NK cells by antibody-dependent cellular cytotoxicity. Binding of antibodies to antigen can trigger activation of the classical complement pathway (see Fig. 4.4). In addition, antibodies can directly neutralise the biological activity of their antigen target. This is a particularly important feature of IgA antibodies, which act predominantly at mucosal surfaces. The humoral immune response is characterised by immunological memory, in which the antibody response to successive exposures to an antigen is qualitatively and quantitatively improved from the first exposure. When a previously unstimulated or ‘naïve’ B lymphocyte is activated by antigen, the first antibody to be produced is IgM, which appears in the serum after 5–10 days. Depending on additional stimuli provided by T lymphocytes, other antibody classes (IgG, IgA and IgE) are produced 1–2 weeks later. If the memory B cell is subsequently re-exposed to the same antigen, the lag time between exposure

 molecules that down-regulate T-cell activity. One such inhibitory molecule, CTLA4, has been harnessed therapeutically in the form of abatacept, which is a fusion protein comprised of the Fc fragment of immunoglobulin linked to CTLA4. This is used to inhibit T-cell activation in rheumatoid arthritis and solid organ transplantation. The inflammatory response Inflammation is the response of tissues to injury or infection, and is necessary for normal repair and healing. This section focuses on the general principles of the inflammatory response and its multisystem manifestations. The role of inflammation in specific diseases is discussed in many other chapters of this book. Acute inflammation Acute inflammation is the result of rapid and complex interplay between the cells and soluble molecules of the innate immune system. The classical external signs include heat, redness, pain and swelling (Fig. 4.8). The inflammatory process is initiated by local tissue injury or infection. Damaged epithelial cells produce cytokines and antimicrobial peptides, causing early infiltration of phagocytic cells. Production of leukotrienes, prostaglandins, histamine, kinins, anaphylotoxins and inducible nitric oxide synthase also occurs within inflamed tissue. These mediators cause vasodilatation and increased vascular permeability, causing trafficking of fluid and cells into the affected tissue. In addition, pro-inflammatory cytokines, such as IL-1, TNF-α and IL-6 produced at the site of injury, are released systemically and act on the hypothalamus to cause fever, and on the liver to stimulate production of acute phase proteins. The acute phase response The acute phase response refers to the production of a variety

of proteins by the liver in response to inflammatory stimuli. These proteins have a wide range of activities. Circulating levels of C-reactive protein (CRP) and serum amyloid A may be increased 1000-fold, contributing to host defence and stimulating repair and regeneration. Fibrinogen plays an essential role in wound healing, and α1-antitrypsin and α1-antichymotrypsin control the pro-inflammatory cascade by neutralising the enzymes produced by activated neutrophils, preventing widespread tissue destruction. In addition, antioxidants such as haptoglobin and manganese superoxide dismutase scavenge for oxygen free radicals, while increased levels of iron-binding proteins such as ferritin and lactoferrin decrease the iron available for uptake by bacteria (p. 941). Immunoglobulins are not acute phase proteins but are often increased in chronic inflammation. Septic shock Septic shock is the clinical manifestation of overwhelming inflammation (p. 196). It is characterised by excessive production of pro-inflammatory cytokines by macrophages, causing hypotension, hypovolaemia and tissue oedema. In addition, uncontrolled neutrophil activation causes release of proteases and oxygen free radicals within blood vessels, damaging the vascular endothelium and further increasing capillary permeability. Direct activation of the coagulation pathway combines with endothelial cell disruption to form clots within the damaged vessels. The specific for a single antigen. Within the thymus T cells undergo a process of stringent selection to ensure that autoreactive cells are destroyed. Mature T lymphocytes leave the thymus and expand to populate other organs of the immune system. It has been estimated that an individual possesses 107–109 T-cell clones, each with a unique T-cell receptor, ensuring at least partial coverage for any antigen encountered. Unlike B cells, T cells cannot recognise intact protein antigens in their native form. Instead, the protein must be broken down into component peptides by antigen-presenting cells for presentation to T lymphocytes in association with HLA molecules on the antigen-presenting cell surface. This process is known as antigen processing and presentation, and it is the complex of peptide and HLA together that is recognised by individual T cells (Fig. 4.7). The structure of HLA molecules varies widely between individuals. Since each HLA molecule has the capacity to present a subtly different peptide repertoire to T lymphocytes, this ensures enormous diversity in recognition of antigens by the T-cell population. All nucleated cells have the capacity to process and present antigens, but cells with specialised antigenpresenting functions include dendritic cells, macrophages and B lymphocytes. These carry additional co-stimulatory molecules, such as CD80 and CD86, providing the necessary ‘second signal’ for full T-cell activation. T lymphocytes can be divided into two subgroups on the basis of function and recognition of HLA molecules. These are designated CD4+ and CD8+ T cells, according to the ‘cluster of differentiation’ (CD) antigen number of key proteins expressed on their cell surface. CD8 + T lymphocytes These cells recognise antigenic peptides in association with HLA class I molecules (HLA-A, HLA-B, HLA-C). They kill infected cells directly through the production of pore-forming molecules such as perforin and release of digesting enzymes triggering apoptosis of the target cell, and are particularly important in defence against viral infection. CD4 + T lymphocytes These cells recognise peptides presented on HLA class II molecules (HLA-DR, HLA-DP and HLA-DQ) and have mainly immunoregulatory functions. They produce cytokines and provide co-stimulatory signals that support the activation of CD8+ T lymphocytes and assist the production of mature antibody by B cells. In addition, their close interaction with phagocytes determines cytokine production by both cell types. CD4+ lymphocytes can be further subdivided into subsets on the basis of the cytokines they produce: • Th1 (T helper) cells typically produce IL-2, IFN-γ and TNF-α, and support the development of delayed-type hypersensitivity responses (p. 83). • Th2 cells typically produce IL-4, IL-5, IL-10 and IL-13, and promote allergic responses (p. 84). • T-regulatory cells (T regs) are a further subset of specialised CD4+ lymphocytes that are important in actively suppressing

activation of other cells and preventing autoimmune disease. • Th17 cells are pro-inflammatory cells defined by their production of IL-17. They are related to regulatory T cells, and play a role in immune defence at mucosal surfaces T-cell activation is regulated by a balance between co-stimulatory molecules, the second signal required for activation, and inhibitory

Chronic inflammation In most instances, the development of an active immune response results in clearance and control of the inflammatory stimulus and resolution of tissue damage. Failure of this process may result in chronic inflammation, with significant associated bystander damage, known as hypersensitivity responses. Persistence of microorganisms can result in ongoing accumulation of neutrophils, macrophages and activated T lymphocytes within the lesion. If this is associated with local deposition of fibrous tissue, a granuloma may form. Granulomas are characteristic of tuberculosis and leprosy (Hansen's disease), in which the microorganism is protected by a robust cell wall that shields it from killing, despite phagocytosis. Laboratory features of inflammation Inflammation is associated with changes in many laboratory investigations. Leucocytosis is common, and reflects the transit of activated neutrophils and monocytes to the site of infection. The platelet count may also be increased. The most widely used laboratory measure of acute inflammation is CRP. Circulating clinical consequences include cardiovascular collapse, acute respiratory distress syndrome, disseminated intravascular coagulation, multi-organ failure and often death. Septic shock most frequently results from infection with Gram-negative bacteria, because lipopolysaccharide produced by these organisms is particularly effective at activating the inflammatory cascade. Early recognition and appropriate early intervention can improve patient outcome (p. 196). Resolution of inflammation Resolution of an inflammatory response is crucial for normal healing. This involves active down-modulation of inflammatory stimuli and repair of bystander damage to local tissues. Extravasated neutrophils undergo apoptosis and are phagocytosed by macrophages, along with the remains of microorganisms. Macrophages also synthesise collagenase and elastase, which break down local connective tissue and aid in the removal of debris. Normal tissue homeostasis is also associated with reversion of parenchymal cells to a non-inflammatory phenotype. Macrophage-derived cytokines, including transforming growth factor-beta (TGF-β) and platelet-derived growth factor, stimulate fibroblasts and promote the synthesis of new collagen, while angiogenic factors stimulate new vessel formation. Fig. 4.8 Clinical features of acute inflammation. In this example, the response is to a penetrating injury and infection of the foot. Hypothalamus: Change in temperature set point Fever Sweating Neuro-endocrine and autonomic stress responses Flushing ↑Respiratory rate ↑Heart rate, flow murmur Adrenal release of glucocorticoids and catecholamines Release of insulin from pancreas Bone marrow: ↑Production and mobilisation of neutrophils Vasodilatation ↑Local vascular permeability Neutrophils + Macrophages Inflammatory mediators and cytokines Tissue damage Bacteria Local infection Skin rupture Phagocytosis Cytokine production Vasodilatation ↑Local vascular permeability ↑Leucocyte influx Headache Delirium Anorexia Low blood pressure Liver: ↑Synthesis of acute phase proteins Enlarged draining lymph nodes Ascending lymphangitis Local cellulitis Pain Redness Swelling Nail

 by the composition of plasma proteins and the morphology of circulating erythrocytes. These factors govern the propensity of red cells to aggregate, the major determinant of the ESR. Erythrocytes are inherently negatively charged, which prevents them from

clumping together in the blood stream. Since plasma proteins are positively charged, an increase in plasma protein concentrations neutralises the negative charge of erythrocytes, overcoming their inherent repulsive forces and causing them aggregate, resulting in rouleaux formation. Rouleaux have a higher mass-to-surface area ratio than single red cells, and therefore sediment faster. The most common reason for an increased ESR is an acute phase response, which causes an increase in the concentration of acute phase proteins, including CRP. However, other conditions that do not affect acute phase proteins may alter the composition and concentration of other plasma protein (Box 4.4). For example, immunoglobulins comprise a significant proportion of plasma proteins but do not participate in the acute phase response. Thus any condition that causes an increase in serum immunoglobulins will increase the ESR without a corresponding increase in CRP. In addition, abnormal red cell morphology can make rouleaux formation impossible. For these reasons, an inappropriately low ESR occurs in spherocytosis and sickle-cell anaemia. Plasma viscosity Plasma viscosity is another surrogate measure of plasma protein concentration. Like the ESR, it is affected by the concentration of large plasma proteins, including fibrinogen and immunoglobulins. It is not affected by properties of erythrocytes and is generally considered to be more reliable than the ESR as a marker of inflammation. levels of many other acute phase reactants, including fibrinogen, ferritin and complement components, are also increased in response to acute inflammation, while albumin levels are reduced. Chronic inflammation is frequently associated with a normocytic normochromic anaemia (p. 943). C-reactive protein C-reactive protein (CRP) is an acute phase reactant synthesised by the liver, which opsonises invading pathogens. Circulating concentrations of CRP increase within 6 hours of the start of an inflammatory stimulus. Serum concentrations of CRP provide a direct biomarker of acute inflammation and, because the serum half-life of CRP is 18 hours, levels fall promptly once the inflammatory stimulus is removed. Sequential measurements are useful in monitoring disease activity (Box 4.4). For reasons that remain unclear, some diseases are associated with only minor elevations of CRP despite unequivocal evidence of active inflammation. These include systemic lupus erythematosus (SLE), systemic sclerosis, ulcerative colitis and leukaemia. An important practical point is that if the CRP is raised in these conditions, it suggests intercurrent infection rather than disease activity. Since the CRP is a more sensitive early indicator of the acute phase response, it is generally used in preference to the erythrocyte sedimentation rate (ESR). If both ESR and CRP are used, any discrepancy should be resolved by assessing the individual determinants of the ESR, which are discussed below. Erythrocyte sedimentation rate The ESR is an indirect measure of inflammation. It measures how fast erythrocytes fall through plasma, which is determined 4.4 Conditions commonly associated with abnormal C-reactive protein (CRP) and/or erythrocyte sedimentation rate (ESR) Condition Consequence Effect on CRP1 Effect on ESR2 Acute bacterial, fungal or viral infection Stimulates acute phase response Increased (range 50–150 mg/L; in severe infections may be

> 300 mg/L) Increased Necrotising bacterial infection Stimulates profound acute inflammatory response Greatly increased (may be 300 mg/L) Increased Chronic bacterial or fungal infection Localised abscess, bacterial endocarditis or tuberculosis Stimulates acute and chronic inflammatory response with polyclonal increase in immunoglobulins, as well as increased acute phase proteins Increased (range 50–150 mg/L) Increased disproportionately to CRP Acute inflammatory diseases Crohn’s disease, polymyalgia rheumatica,

> inflammatory arthritis Stimulates acute phase response Increased (range 50–150 mg/L) Increased Systemic lupus erythematosus, Sjögren’s syndrome, ulcerative colitis Chronic inflammatory response Normal Increased Multiple myeloma Monoclonal increase in serum immunoglobulin without acute inflammation Normal Increased Pregnancy, old age, end-stage renal disease Increased fibrinogen Normal Moderately increased 1Reference range < 10 mg/L. 2Reference range: adult males < 10 mm/hr, adult females < 20 mm/hr.

4.5 Immune deficiencies and common patterns of infection Phagocyte deficiency Complement deficiency Antibody deficiency T-lymphocyte deficiency Bacteria Staphylococcus aureus Pseudomonas aeruginosa Serratia marcescens Burkholderia cenocepacia Nocardia Mycobacterium tuberculosis Atypical mycobacteria Neisseria meningitidis Neisseria gonorrhoeae Haemophilus influenzae Streptococcus pneumoniae Haemophilus influenzae Streptococcus pneumoniae Staphylococcus aureus Mycobacterium tuberculosis Atypical mycobacteria Fungi Candida spp. Aspergillus spp. – – Candida spp. Aspergillus spp. Pneumocystis jirovecii Viruses – – Cytomegalovirus (CMV) Enteroviruses Epstein–Barr virus (EBV) Herpes zoster virus Human papillomavirus Human herpesvirus 8 Protozoa – Giardia lamblia Toxoplasma gondii Cryptosporidia Presenting problems in immune disorders Recurrent infections Infections can occur in otherwise healthy individuals but recurrent infection raises suspicion of an immune deficiency. Depending on the component of the immune system affected, the infections may involve bacteria, viruses, fungi or protozoa, as summarised in Box 4.5. T-cell deficiencies can involve pathogens from all groups. Aetiology Infections secondary to immune deficiency occur because of defects in the number or function of phagocytes, B cells, T cells or complement, as described later in this chapter. Clinical assessment Clinical features that may indicate immune deficiency are listed in Box 4.6. Frequent or severe infections, or ones caused by unusual organisms or at unusual sites are typical of immune deficiency. Investigations Initial investigations should include full blood count and white cell differential, CRP, renal and liver function tests, urine dipstick, serum immunoglobulins with protein electrophoresis, and HIV testing. Additional microbiological tests, virology and imaging are required to identify the causal organism and localise the site of infection, as outlined in Box 4.7. If primary immune deficiency is suspected on the basis of initial investigations, more specialised tests should be considered, as summarised in Box 4.8. Management If an immune deficiency is suspected but has not yet been formally characterised, patients should not receive live vaccines because of the risk of vaccine-induced disease. Further management depends on the underlying cause and details are provided later. 4.6 Warning signs of primary immune deficiency* In children In adults ≥ 4 new ear infections within 1 year ≥ 2 new ear infections within 1 year ≥ 2 serious sinus infections within 1 year ≥ 2 new sinus infections within 1 year, in the absence of allergy ≥ 2 months on antibiotics with little effect Recurrent viral infections ≥ 2 pneumonias within 1 year ≥ 1 pneumonia per year for more than 1 year Failure of an infant to gain weight or grow normally Chronic diarrhoea with weight loss Recurrent deep skin or organ abscesses Recurrent deep skin or organ abscesses Persistent thrush in mouth or elsewhere on skin after infancy Persistent thrush or fungal infection on skin or elsewhere Need for intravenous antibiotics to clear infections Recurrent need for intravenous antibiotics to clear infections ≥ 2 deep-seated infections such as sepsis, meningitis or

cellulitis Infection with atypical mycobacteria A family history of primary immune deficiency A family history of primary immune deficiency *The presence of two or more of the above features may indicate the presence of an underlying primary immunodeficiency. © Jeffrey Modell Foundation.

74 • CLINICAL IMMUNOLOGY 4.7 Initial investigations in suspected immune deficiency Test Value Comment Full blood count Full white cell differential May define pathway for further investigation Acute phase reactants Help determine presence of active infection Serum immunoglobulins Detection of antibody deficiency Serum protein electrophoresis Detection of paraprotein May be the cause of immune paresis; paraprotein should be excluded prior to diagnosis of primary antibody deficiency Serum free light chains/Bence Jones proteins Detection of paraprotein Human immunodeficiency virus (HIV) test To exclude HIV as cause of secondary immune deficiency Imaging according to history and examination findings Detection of active infection/end-organ damage May support treatment decisions, e.g. if there is evidence of bronchiectasis 4.8 Specialist investigations in suspected immune deficiency Test Value Comment Complement (C3/C4/CH50/AP50) Investigation of recurrent pyogenic bacterial infection Inherited complement deficiency likely to give low/ absent results on functional assays Test vaccination Determination of functional humoral immune response Helpful in patients with borderline low or normal immunoglobulins but confirmed recurrent infection Neutrophil function Investigation of recurrent invasive bacterial and fungal infection, especially with catalase-positive organisms Respiratory burst low/absent in chronic granulomatous disease Investigation of leucocyte adhesion deficiency Leucocytosis with absent CD11a, b, c expression Lymphocyte immunophenotyping (by flow cytometry) Determination of specific lymphocyte subsets, T cell, B cell, NK cell May define specific primary immune deficiency, e.g. absent B cells in X-linked agammaglobulinaemia Lymphocyte proliferation Determination of lymphocyte proliferation in response to mitogenic stimulation Poor responses seen in certain T-cell immune deficiencies Cytokine production To determine T-cell immune function in response to antigen stimulation; limited availability, not routine Can be helpful, for example, in investigation of atypical mycobacterial infection Genetic testing Under specialist supervision when specific primary immune deficiency suspected May confirm genetic cause, with implications for family members and future antenatal testing (NK = natural killer) Intermittent fever Intermittent fever has a wide differential diagnosis, including recurrent infection, malignancy and certain rheumatic disorders, such as Still’s disease, vasculitis and SLE (pp. 1040 and 1034), but a familial fever syndrome is a potential cause. Aetiology Familial fever syndromes are genetic disorders caused by mutations in genes responsible for regulating the inflammatory response. The symptoms are caused by activation of intracellular signalling pathways involved in the regulation of inflammation, with over-production of pro-inflammatory cytokines such as IL-1. Clinical assessment A full clinical history and physical examination should be performed, paying attention to the patient’s ethnic background and any family history of a similar disorder. If this assessment shows no evidence of underlying infection, malignancy or a rheumatic disorder and there is a positive family history and early age at onset, then the likelihood of a familial fever syndrome is increased. Investigations Blood should be taken for a full blood count, measurement of ESR and CRP, and assessment of renal and liver function. Serum ferritin should be checked, as very high levels support the diagnosis of Still’s disease. Blood and urine cultures should also be performed, along with an autoimmune screen that includes measurement of antinuclear antibodies and consideration of antineutrophil cytoplasmic antibodies to check for evidence of SLE or vasculitis, respectively. Imaging may be required to exclude occult infection. If these investigations provide

no evidence of infection or another cause, then genetic analysis should be considered to confirm the diagnosis of a familial fever syndrome (p. 81). Negative genetic testing does not, however, entirely exclude a periodic fever syndrome. Management Symptomatic management with non-steroidal anti-inflammatory drugs (NSAIDs) should be initiated, pending the results of investigations. If the response to NSAIDs is inadequate, glucocorticoids can be tried, provided that infection has been excluded. If a familial fever syndrome is confirmed, then definitive therapy should be initiated, depending on the underlying diagnosis (p. 81).

leading to hypotension, and bronchoconstriction, as summarised in Box 4.9. It can be difficult to distinguish IgE-mediated anaphylaxis clinically from non-specific degranulation of mast cells on exposure to drugs, chemicals or other triggers where IgE is not involved, previously known as anaphylactoid reactions. Common triggers are shown in Box 4.10. Clinical assessment The clinical features of anaphylaxis and ‘anaphylactoid’ reactions are indistinguishable and are summarised in Figure 4.9. Several other conditions can mimic anaphylaxis and these are listed in Box 4.11. It is important to assess the severity of the reaction, and the time between allergen exposure and onset of symptoms provides Anaphylaxis Anaphylaxis is a potentially life-threatening, systemic allergic reaction characterised by circulatory collapse, bronchospasm, laryngeal stridor, often associated with angioedema, and urticaria. The risk of death is increased in patients with pre-existing asthma, particularly if this is poorly controlled, and in situations where treatment with adrenaline (epinephrine) is delayed. Aetiology Anaphylaxis occurs when an allergen binds to and cross-links membrane-bound IgE on mast cells in a susceptible individual, causing release of histamine, tryptase and other vasoactive mediators from mast cells. These mediators have a variety of effects, including vasodilatation, increased capillary permeability 4.10 Common causes of systemic allergic reactions Anaphylaxis: IgE-mediated mast cell degranulation Foods • Peanuts • Tree nuts • Fish and shellfish • Milk • Eggs • Soy products Insect stings • Bee venom • Wasp venom Chemicals, drugs and other foreign proteins • Intravenous anaesthetic agents (suxamethonium) • Penicillin and other antibiotics • Latex Anaphylactoid: non-lgE-mediated mast cell degranulation Drugs • Aspirin and non-steroidal anti-inflammatory drugs (NSAIDs) • Opiates • Radiocontrast media Physical • Exercise • Cold Idiopathic • No cause is identified in 20% of patients with anaphylaxis 4.9 Clinical features of mast cell degranulation Mediator Biological effects Pre-formed and stored within granules Histamine Vasodilatation, chemotaxis, bronchoconstriction, increased capillary permeability and increased mucus secretion Tryptase Bronchoconstriction, activates complement C3 Eosinophil chemotactic factor Eosinophil chemotaxis Neutrophil chemotactic factor Neutrophil chemotaxis Synthesised on activation of mast cells Leukotrienes Increase vascular permeability, chemotaxis, mucus secretion and smooth muscle contraction Prostaglandins Bronchoconstriction, platelet aggregation and vasodilatation Thromboxanes Bronchoconstriction Platelet-activating factor Bronchoconstriction, chemotaxis of eosinophils and neutrophils Fig. 4.9 Clinical manifestations of anaphylaxis. In this example, the response is to an insect sting containing venom to which the patient is allergic. This causes release of histamine and other vasoactive mediators, which cause the characteristic features of anaphylaxis that are illustrated. Itching of palms, soles of feet and genitalia Feeling of impending doom, loss of consciousness Conjunctival injection Flushing Sweating Hypotension Urticaria Wheeze, bronchoconstriction Angioedema of lips and mucous membrane Abdominal pain Diarrhoea Cardiac arrhythmias Laryngeal obstruction Stridor Wasp sting

76 • CLINICAL IMMUNOLOGY Management The principles of management of the acute event are summarised in Box 4.12. Individuals who have recovered from an anaphylactic event should be referred for specialist assessment. The aim is to identify the trigger factor, to educate the patient regarding avoidance and management of subsequent episodes, and to establish whether specific treatment, such as immunotherapy, is indicated. If the trigger factor cannot be identified or avoided, recurrence is common. Patients who have previously experienced an anaphylactic event should be prescribed self-injectable adrenaline (epinephrine) and they and their families or carers should be instructed in its use (Box 4.13). The use of a MedicAlert (or similar) bracelet will increase the likelihood of the injector being administered in an emergency. Allergy in adolescence requires additional consideration and management, as set out in Box 4.14. a guide. Enquiry should be made about potential triggers. If none is immediately obvious, a detailed history of the previous 24 hours may be helpful. The most common triggers of anaphylaxis are foods, latex, insect venom and drugs (see Box 4.10). A history of previous local allergic responses to the offending agent is common. The route of allergen exposure may influence the principal clinical features of a reaction; for example, if an allergen is inhaled, the major symptom is frequently wheezing. Features of anaphylaxis may overlap with the direct toxic effects of drugs and venoms (Chs 7 and 8). Potentiating factors, such as exercise or alcohol, can lower the threshold for an anaphylactic event. It is important to identify precipitating factors so that appropriate avoidance measures may be taken in the longer term. Investigations Measurement of serum mast cell tryptase concentrations is useful to confirm the diagnosis but cannot distinguish between anaphylaxis and non-IgE-mediated anaphylactoid reactions. Specific IgE tests may be useful in confirming hypersensitivity and may be preferable to skin-prick tests when investigating patients with a history of anaphylaxis. 4.14 Allergy in adolescence • Resolution of childhood allergy: most children affected by allergy to milk, egg, soybean or wheat will grow out of their food allergies by adolescence but allergies to peanuts, tree nuts, fish and shellfish are frequently life-long. • Risk-taking behaviour and fatal anaphylaxis: serious allergy is increasingly common in adolescents and this is the highest risk group for fatal, food-induced anaphylaxis. This is associated with increased risk-taking behaviour, and food-allergic teenagers are more likely than adults to eat unsafe foods, deny reaction symptoms and delay emergency treatment. • Emotional impact of food allergies: some adolescents may neglect to carry a prescribed adrenalin autoinjector because of the associated nuisance and/or stigma. Surveys of food-allergic teens reveal that many take risks because they feel socially isolated by their allergy. 4.13 How to prescribe self-injectable adrenaline (epinephrine) Prescription (normally initiated by an immunologist or allergist) • Specify the brand of autoinjector, as they have different triggering mechanisms • Prescribe two devices Indications • Anaphylaxis to allergens that are difficult to avoid: Insect venom Foods • Idiopathic anaphylactic reactions • History of severe localised reactions with high risk of future anaphylaxis: Reaction to trace allergen Likely repeated exposure to allergen • History of severe localised reactions with high risk of adverse outcome: Poorly controlled asthma Lack of access to emergency care Patient and family education • Know when and how to use the device • Carry the device at all times • Seek medical assistance immediately after use • Wear an alert bracelet or necklace • Include the school in education for young patients (see ‘Further information’) Other considerations • Caution with β-blockers in anaphylactic patients as they may increase the severity of an anaphylactic reaction and reduce the response to adrenaline (epinephrine) 4.12 Emergency management of anaphylaxis Treatment Comment Prevent further contact with allergen Prevents ongoing mast cell activation Ensure airway patency Prevents hypoxia Administer adrenaline (epinephrine) promptly: 0.3–1.0 mL 1 : 1000 solution IM in adults Repeat at 5–10-min intervals if initial response is inadequate Intramuscular route important

because of peripheral vasoconstriction Acts within minutes Increases blood pressure Reverses bronchospasm Administer antihistamines: Chlorphenamine 10 mg IM or slow IV injection Blocks effect of histamine on target cells Administer glucocorticoids: Hydrocortisone 200 mg IV Reduces cytokine release Prevents rebound symptoms in severe cases Provide supportive treatment: Nebulised β2-agonists IV fluids Oxygen Reverses bronchospasm Restores plasma volume Reverses hypoxia (IM = intramuscular; IV = intravenous) 4.11 Differential diagnosis of anaphylaxis Causes of hypotension • Vasovagal syncope • Cardiac arrhythmia • Cardiogenic shock Causes of respiratory distress • Status asthmaticus • Pulmonary embolus Causes of laryngeal obstruction • C1 inhibitor deficiency • Idiopathic angioedema Causes of generalised flushing • Systemic mastocytosis • Carcinoid syndrome • Phaeochromocytoma

antifungal agents. The most important examples are illustrated in Figure 4.10 and discussed below. Chronic granulomatous disease This is caused by mutations in genes that encode NADPH oxidase enzymes, which results in failure of oxidative killing. The defect leads to susceptibility to catalase-positive organisms such as Staphylococcus aureus, Burkholderia cenocepacia and Aspergillus. Intracellular killing of mycobacteria in macrophages is also impaired. Infections most commonly involve the lungs, lymph nodes, soft tissues, bone, skin and urinary tract, and are characterised histologically by granuloma formation. Most cases are X-linked (p. 48). Leucocyte adhesion deficiencies These very rare disorders of phagocyte migration occur because of failure to express adhesion molecules on the surface of leucocytes, resulting in their inability to exit the blood stream. The most common cause is loss-of-function mutations affecting the ITGB2 gene, which encodes the integrin β-2 chain, a component of the adhesion molecule LFA1. They are characterised by recurrent bacterial infections but sites of infection lack evidence of neutrophil infiltration, such as pus formation. Peripheral blood neutrophil counts may be very high during acute infection because of the failure of mobilised neutrophils to exit blood vessels. Specialised tests show reduced or absent expression of adhesion molecules on neutrophils. Immune deficiency The consequences of immune deficiency include recurrent infection, autoimmunity as a result of immune dysregulation, and increased susceptibility to malignancy, especially malignancy driven by viral infections such as Epstein–Barr virus. Immune deficiency may arise through intrinsic defects in immune function but is much more commonly due to secondary causes, including infection, drug therapy, malignancy and ageing. This section gives an overview of primary immune deficiencies. More than a hundred such deficiencies have been described, most of which are genetically determined and present in childhood or adolescence. The presentation of immune deficiency depends on the component of the immune system that is defective (see Box 4.5). There is considerable overlap and redundancy in the immune network, however, and some diseases do not fall easily into this classification. Primary phagocyte deficiencies Primary phagocyte deficiencies typically present with recurrent bacterial and fungal infections, which may involve unusual sites. Affected patients require aggressive management of infections, including intravenous antibiotics and surgical drainage of abscesses, and long-term prophylaxis with antibacterial and Fig. 4.10 Normal phagocyte function and mechanisms of primary phagocyte deficiency. Under normal circumstances, neutrophils traverse the endothelium to enter tissues by the cell surface molecule lymphocyte function-associated antigen 1 (LFA1), which binds to intercellular adhesion molecule 1 (ICAM1) on endothelium. In order for macrophages to engulf and kill microorganisms, they need to be activated by cytokines and also require nicotinamide adenine dinucleotide phosphate (NADPH)

oxidase to generate free radicals. Primary phagocyte deficiencies can occur as the result of leucocytes being unable to traverse endothelium due to defects in LFA1, because of mutations in cytokines or their receptors, or because of defects in NADPH oxidase. (IFN-γ = interferon-gamma; IL = interleukin) Neutrophils traverse endothelium through binding of LFA1 to ICAM1 Normal Primary phagocyte deficiency Leucocyte adhesion deficiency Neutrophils cannot traverse endothelium due to defects in ITGB2, a component of LFA1 Chronic granulomatous disease Cytokines activate macrophages Destruction of microorganisms through NADPH oxidase-mediated killing Cytokine defects LFA1 IL-23 IL-12 IFN-γ Phagocytes cannot be activated due to defects in cytokines or their receptors Microorganisms cannot be destroyed in lysosomes due to NADPH oxidase deficiency IL-23 IL-12 IFN-γ IL-23 IL-12 IFN-γ ICAM1

78 • CLINICAL IMMUNOLOGY Management Patients with complement deficiencies should be vaccinated with meningococcal, pneumococcal and H. influenzae B vaccines to boost their adaptive immune responses. Lifelong prophylactic penicillin to prevent meningococcal infection is recommended, as is early access to acute medical assessment in the event of infection. Patients should also carry a MedicAlert or similar. At-risk family members should be screened for complement deficiencies with functional complement assays. The management of C1 esterase deficiency is discussed elsewhere. Primary antibody deficiencies Primary antibody deficiencies occur as the result of abnormalities in B-cell function, as summarised in Figure 4.11. They are characterised by recurrent bacterial infections, particularly of the respiratory and gastrointestinal tract. The most common causative organisms are encapsulated bacteria such as Streptococcus pneumoniae and H. influenzae. These disorders usually present in infancy, when the protective benefit of placental transfer of maternal immunoglobulin has waned. The most important causes are discussed in more detail below. X-linked agammaglobulinaemia This rare X-linked disorder (p. 48) is caused by mutations in the BTK gene, which encodes Bruton tyrosine kinase, a signalling protein that is required for B-cell development. Affected males present with severe bacterial infections during infancy. There is a marked reduction in B-cell numbers and immunoglobulin levels are low or undetectable. Management is with immunoglobulin replacement therapy and antibiotics to treat infections. Selective IgA deficiency This is the most common primary antibody deficiency, affecting 1 : 600 northern Europeans. Although IgA deficiency is usually asymptomatic with no clinical sequelae, about 30% of individuals experience recurrent mild respiratory and gastrointestinal infections. The diagnosis can be confirmed by measurement of IgA levels, which are low or undetectable (< 0.05 g/L). In some Defects in cytokines and cytokine receptors Mutations of the genes encoding cytokines such as IFN-γ, IL-12, IL-23 or their receptors result in failure of intracellular killing by macrophages, and affected individuals are particularly susceptible to mycobacterial infections. Complement pathway deficiencies Loss-of-function mutations have been identified in almost all the complement pathway proteins (see Fig. 4.4). While most complement deficiencies are rare, mannose-binding lectin deficiency is common and affects about 5% of the northern European population, many of whom are asymptomatic (see below). Clinical features Patients with deficiency in complement proteins can present in different ways. In some cases, the presenting feature is recurrent infection with encapsulated bacteria, particularly Neisseria spp., reflecting the importance of the membrane attack complex in defence against these organisms. However, genetic deficiencies of the classical complement pathway (C1, C2 and C4) also present with an increased risk of autoimmune disease, particularly SLE (p. 1034). Individuals with mannose-binding lectin deficiency have an increased incidence of bacterial infections if subjected to an additional cause of immune compromise, such as premature birth or

chemotherapy. The significance of this condition has been debated, however, since population studies have shown no overall increase in infectious disease or mortality in patients with this disorder. Deficiency of the regulatory protein Cl inhibitor is not associated with recurrent infection but causes recurrent angioedema (p. 87). Investigations Screening for complement deficiencies usually involves specialised functional tests of complement-mediated haemolysis. These are known as the CH50 (classical haemolytic pathway 50) and AP50 (alternative pathway 50) tests. If abnormal, haemolytic tests are followed by measurement of individual complement components. Fig. 4.11 B lymphocytes and primary antibody deficiencies (green boxes). (Ig = immunoglobulin) Failure of lymphocyte precursors: Severe combined immune deficiency Stem cells Lymphoid progenitors Bone marrow Failure of production of IgG antibodies: Common variable immune deficiency Specific antibody deficiency IgM-producing B cells Failure of B-cell maturation: X-linked agammaglobulinaemia Immature B cells IgG IgE IgA Plasma cells Failure of IgA production: Selective IgA deficiency

IgG level just prior to an infusion) within the normal range. This has been shown to minimise progression of end-organ damage and improve clinical outcome. Treatment may be self-administered and is life-long. Benefits of immunisation are limited because of the defect in IgG antibody production, and as with all primary immune deficiencies, live vaccines should be avoided. Primary T-lymphocyte deficiencies These are a group of diseases characterised by recurrent viral, protozoal and fungal infections (see Box 4.5). Many T-cell deficiencies are also associated with defective antibody production because of the importance of T cells in providing help for B cells. These disorders generally present in childhood. Several causes of T-cell deficiency are recognised. These are summarised in Figure 4.12 and discussed in more detail below. DiGeorge syndrome This results from failure of development of the third and fourth pharyngeal pouches, and is usually caused by a deletion of chromosome 22q11. The immune deficiency is accounted for by failure of thymic development; however, the immune deficiency can be very heterogeneous. Affected patients tend to have very low numbers of circulating T cells despite normal development in the bone marrow. It is associated with multiple developmental anomalies, including congenital heart disease, hypoparathyroidism, tracheo-oesophageal fistulae, cleft lip and palate. Bare lymphocyte syndromes These rare disorders are caused by mutations in a variety of genes that regulate expression of HLA molecules or their transport to the cell surface. If HLA class I molecules are affected, CD8+ lymphocytes fail to develop normally, while absent expression of HLA class II molecules affects CD4+ lymphocyte maturation. In addition to recurrent infections, failure to express HLA class I is associated with systemic vasculitis caused by uncontrolled activation of NK cells. Severe combined immune deficiency Severe combined immune deficiency (SCID) results from mutations in a number of genes that regulate lymphocyte development, with failure of T-cell maturation, with or without accompanying B- and NK-cell maturation. The most common cause is X-linked SCID, resulting from loss-of-function mutations in the interleukin-2 receptor gamma (IL2RG) gene. The gene product is a component of several interleukin receptors, including those for IL-2, IL-7 and IL-15, which are absolutely required for T-cell and NK development. This results in T-cell-negative, NK-cell-negative, patients, there is a compensatory increase in serum IgG levels. Specific treatment is generally not required. Common variable immune deficiency Common variable immune deficiency (CVID) is characterised by low serum IgG levels and failure to make antibody responses to exogenous pathogens. It is a heterogeneous adult-onset primary immune

deficiency of unknown cause. The presentation is with recurrent infections, and bronchiectasis is a recognised complication. Paradoxically, antibody-mediated autoimmune diseases, such as idiopathic thrombocytopenic purpura and autoimmune haemolytic anaemia, are common in CVID. It is also associated with an increased risk of malignancy, particularly lymphoproliferative disease. Functional IgG antibody deficiency This is a poorly characterised condition resulting in defective antibody responses to polysaccharide antigens. Some patients are also deficient in the antibody subclasses IgG2 and IgG4, and this condition was previously called IgG subclass deficiency. There is overlap between specific antibody deficiency, IgA deficiency and CVID, and some patients may progress to a more global antibody deficiency over time. Investigations Serum immunoglobulins (Box 4.15) should be measured in conjunction with protein and urine electrophoresis to exclude secondary causes of hypogammaglobulinaemia, and B- and T-lymphocyte subsets should be measured. Specific antibody responses to known pathogens should be assessed by measuring IgG antibodies against tetanus, H. influenzae and S. pneumoniae (most patients will have been exposed to these antigens through infection or immunisation). If specific antibody levels are low, immunisation with the appropriate killed vaccine should be followed by repeat antibody measurement 6–8 weeks later; failure to mount a response indicates a significant defect in antibody production. These functional tests have generally superseded IgG subclass quantitation. Management Patients with antibody deficiencies generally require aggressive treatment of infections and prophylactic antibiotics may be indicated. An exception is deficiency of IgA, which usually does not require treatment. The mainstay of treatment in most patients with antibody deficiency is immunoglobulin replacement therapy. This is derived from plasma from hundreds of donors and contains IgG antibodies to a wide variety of common organisms. Replacement immunoglobulin may be administered either intravenously or subcutaneously, with the aim of maintaining trough IgG levels (the 4.15 Investigation of primary antibody deficiencies Serum immunoglobulin (Ig) concentrations Circulating lymphocyte numbers IgM IgG IgA IgE B cells T cells Test immunisation Selective IgA deficiency Normal Often elevated Absent Normal Normal Normal Not applicable* Common variable immune deficiency Normal or low Low Low or absent Low or absent Variable Variable No antibody response Specific antibody deficiency Normal Normal Normal Normal Normal Normal No antibody response to polysaccharide antigens *Test immunisation is not usually performed in IgA deficiency but some patients may have impaired responses.

 include lymphadenopathy, splenomegaly and a variety of other autoimmune diseases. Susceptibility to infection is increased because of the neutropenia. Secondary immune deficiencies Secondary immune deficiencies are much more common than primary immune deficiencies and occur when the immune system is compromised by external factors (Box 4.16). Common causes include infections, such as HIV and measles, and cytotoxic B-cell-positive SCID. Another cause is deficiency of the enzyme adenosine deaminase (ADA), which causes lymphocyte death due to accumulation of toxic purine metabolites intracellularly, resulting in T-cell-negative, B-cell-negative and NK-cell-negative SCID. The absence of an effective adaptive immune response causes recurrent bacterial, fungal and viral infections soon after birth. Bone marrow transplantation (BMT; p. 936) is the treatment option of first choice. Gene therapy has been approved for treatment of ADA deficiency when there is no suitable donor for BMT and is under investigation for a number of other causes of SCID. Investigations The principal tests for T-lymphocyte deficiencies are a total lymphocyte count and quantitation of individual lymphocyte subpopulations. Serum immunoglobulins should also be measured. Second-line, functional tests of T-cell activation and proliferation may be indicated. Patients in whom T-lymphocyte deficiencies

are suspected should be tested for HIV infection (p. 310). Management Patients with T-cell deficiencies should be considered for antiPneumocystis and antifungal prophylaxis, and require aggressive management of infections when they occur. Immunoglobulin replacement is indicated for associated defective antibody production. Stem cell transplantation (p. 936) or gene therapy may be appropriate in some disorders. Where a family history is known and antenatal testing confirms a specific defect, stem cell therapy prior to recurrent invasive infection can improve outcome. Autoimmune lymphoproliferative syndrome This rare disorder is caused by failure of normal lymphocyte apoptosis, most commonly due to mutations in the FAS gene, which encodes Fas, a signalling protein that regulates programmed cell death in lymphocytes. This results in massive accumulation of autoreactive T cells, which cause autoimmune-mediated anaemia, thrombocytopenia and neutropenia. Other features Fig. 4.12 T-lymphocyte function and dysfunction (green boxes). (HLA = human leucocyte antigen) Failure of lymphocyte precursors: Severe combined immune deficiency Stem cells Lymphoid progenitors Bone marrow Failure of expression of HLA molecules: Bare lymphocyte syndromes Failure of thymic development: DiGeorge syndrome Proliferation and maturation of thymocytes Export of mature T lymphocytes to periphery T-lymphocyte activation and effector function Apoptotic cell death Failure of apoptosis: Autoimmune lymphoproliferative syndromes Thymus Failure of cytokine production: Cytokine deficiencies 4.16 Causes of secondary immune deficiency Physiological • Ageing • Prematurity • Pregnancy Infection • HIV infection • Measles • Mycobacterial infection Iatrogenic • Immunosuppressive therapy • Antineoplastic agents • Glucocorticoids • Stem cell transplantation • Radiation injury • Antiepileptic agents Malignancy • B-cell malignancies including leukaemia, lymphoma and myeloma • Solid tumours • Thymoma Biochemical and nutritional disorders • Malnutrition • Renal insufficiency/dialysis • Diabetes mellitus • Specific mineral deficiencies (iron, zinc) Other conditions • Burns • Asplenia/hyposplenism

attacks. Standard anti-inflammatory drugs, including colchicine and glucocorticoids, are ineffective in suppressing the attacks but IL-1 inhibitors, such as anakinra, and TNF inhibitors, such as etanercept, may improve symptoms and can induce complete remission in some patients. TNF receptor-associated periodic syndrome TNF receptor-associated periodic syndrome (TRAPS) also known as Hibernian fever, is an autosomal dominant syndrome caused by mutations in the TNFRSF1A gene. The presentation is with recurrent attacks of fever, arthralgia, myalgia, serositis and rashes. Attacks may be prolonged for 1 week or more. During a typical attack, laboratory findings include neutrophilia, increased CRP and elevated IgA levels. The diagnosis can be confirmed by low serum levels of the soluble type 1 TNF receptor and by mutation screening of the TNFRSF1A gene. As in FMF, the major complication is amyloidosis, and regular screening for proteinuria is advised. Acute episodes respond to systemic glucocorticoids. Therapy with IL-1 inhibitors, such as anakinra, can be effective in preventing attacks. Amyloidosis Amyloidosis is the name given to a group of acquired and hereditary disorders characterised by the extracellular deposition of insoluble proteins. Pathophysiology Amyloidosis is caused by deposits consisting of fibrils of the specific protein involved, linked to glycosaminoglycans, proteoglycans and serum amyloid P. Protein accumulation may be localised or systemic, and the clinical manifestations depend on the organ(s) affected. Amyloid diseases are classified by the aetiology and type of protein deposited (Box 4.18). Clinical features The clinical presentation may be with nephrotic syndrome (p. 395), cardiomyopathy (p. 538) or peripheral neuropathy (p. 1138). Amyloidosis

should always be considered as a potential diagnosis in patients with these disorders when the cause is unclear. Investigations The diagnosis is established by biopsy, which may be of an affected organ, rectum or subcutaneous fat. The pathognomonic histological feature is apple-green birefringence of amyloid deposits when stained with Congo red dye and viewed under polarised light. Immunohistochemical staining can identify the type of amyloid fibril present. Quantitative scintigraphy with radiolabelled serum amyloid P is a valuable tool in determining the overall load and distribution of amyloid deposits. Management The aims of treatment are to support the function of affected organs and, in acquired amyloidosis, to prevent further amyloid deposition through treatment of the primary cause. When the latter is possible, regression of existing amyloid deposits may occur. Autoimmune disease Autoimmunity can be defined as the presence of immune responses against self-tissue. This may be a harmless phenomenon, identified and immunosuppressive drugs, particularly those used in the management of transplantation, autoimmunity and cancer. Physiological immune deficiency occurs at the extremes of life; the decline of the immune response in the elderly is known as immune senescence (Box 4.17). Management of secondary immune deficiency is described in the relevant chapters on infectious diseases (Ch. 11), HIV (Ch. 12), haematological disorders (Ch. 23) and oncology (Ch. 33). Periodic fever syndromes These rare disorders are characterised by recurrent episodes of fever and organ inflammation, associated with an elevated acute phase response (p. 74). Familial Mediterranean fever Familial Mediterranean fever (FMF) is the most common of the familial periodic fevers, predominantly affecting Mediterranean people, including Arabs, Turks, Sephardic Jews and Armenians. It results from mutations of the MEFV gene, which encodes a protein called pyrin that regulates neutrophil-mediated inflammation by indirectly suppressing the production of IL-1. FMF is characterised by recurrent painful attacks of fever associated with peritonitis, pleuritis and arthritis, which last for a few hours to 4 days and are associated with markedly increased CRP levels. Symptoms resolve completely between episodes. Most individuals have their first attack before the age of 20. The major complication of FMF is AA amyloidosis (see below). Colchicine significantly reduces the number of febrile episodes in 90% of patients but is ineffective during acute attacks. Mevalonic aciduria (mevalonate kinase deficiency) Mevalonate kinase deficiency, previously known as hyper-IgD syndrome, is an autosomal recessive disorder that causes recurrent attacks of fever, abdominal pain, diarrhoea, lymphadenopathy, arthralgia, skin lesions and aphthous ulceration. Most patients are from Western Europe, particularly the Netherlands and northern France. It is caused by loss-of-function mutations in the gene encoding mevalonate kinase, which is involved in the metabolism of cholesterol. It remains unclear why this causes an inflammatory periodic fever. Serum IgD and IgA levels may be persistently elevated, and CRP levels are increased during acute

4.17 Immune senescence • T-cell responses: decline, with reduced delayed-type hypersensitivity responses. • Antibody production: decreased for many exogenous antigens. Although autoantibodies are frequently detected, autoimmune disease is less common. • Response to vaccination: reduced; 30% of healthy older people may not develop protective immunity after influenza vaccination. • Allergic disorders and transplant rejection: less common. • Susceptibility to infection: increased; community-acquired pneumonia by threefold and urinary tract infection by 20-fold. Latent infections, including tuberculosis and herpes zoster, may be reactivated. • Manifestations of inflammation: may be absent, with lack of pyrexia or leucocytosis. • Secondary immune deficiency: common.

 those determining cytokine activity, co-stimulation (the expression of second signals required for full T-cell activation; see Fig. 4.7) and cell death. Many of the same

gene variants underlie multiple autoimmune disorders, reflecting their common pathogenesis (Box 4.19). Even though some of these associations are the strongest that have been identified in complex genetic diseases, only by the presence of low-titre autoantibodies or autoreactive T cells. However, if these responses cause significant organ damage, autoimmune diseases occur. These are a major cause of chronic morbidity and disability, affecting up to 1 in 30 adults at some point during life. Pathophysiology Autoimmune diseases result from the failure of immune tolerance, the process by which the immune system recognises and accepts self-tissue. Central immune tolerance occurs during lymphocyte development, when T and B lymphocytes that recognise self-antigens are eliminated before they develop into fully immunocompetent cells. This process is most active in fetal life but continues throughout life as immature lymphocytes are generated. Some autoreactive cells inevitably evade deletion and escape into the circulation, however, and are controlled through peripheral tolerance mechanisms. Peripheral immune tolerance mechanisms include the suppression of autoreactive cells by regulatory T cells; the generation of functional hyporesponsiveness (anergy) in lymphocytes that encounter antigen in the absence of the co-stimulatory signals that accompany inflammation; and cell death by apoptosis. Autoimmune diseases develop when selfreactive lymphocytes escape from these tolerance mechanisms. Multiple genetic and environmental factors contribute to the development of autoimmune disease. Autoimmune diseases are much more common in women than in men, for reasons that remain unclear. Many are associated with genetic variations in the HLA loci, reflecting the importance of HLA genes in shaping lymphocyte responses. Other important susceptibility genes include 4.19 Association of specific gene polymorphisms with autoimmune diseases Gene Function Diseases HLA complex Key determinants of antigen presentation to T cells Most autoimmune diseases PTPN22 Regulation of T- and B-cell receptor signalling Rheumatoid arthritis, type 1 diabetes, systemic lupus erythematosus CTLA4 Important co-stimulatory molecule that transmits inhibitory signals to T cells Rheumatoid arthritis, type 1 diabetes IL23R Cytokine-mediated control of T cells Inflammatory bowel disease, psoriasis, ankylosing spondylitis TNFRSF1A Control of tumour necrosis factor network Multiple sclerosis ATG5 Autophagy Systemic lupus erythematosus 4.18 Causes of amyloidosis Disorder Pathological basis Predisposing conditions Other features Acquired systemic amyloidosis Reactive (AA) amyloidosis (p. 81) Increased production of serum amyloid A as part of prolonged or recurrent acute inflammatory response Chronic infection (tuberculosis, bronchiectasis, chronic abscess, osteomyelitis) Chronic inflammatory diseases (untreated rheumatoid arthritis, familial Mediterranean fever) 90% of patients present with non-selective proteinuria or nephrotic syndrome Light chain amyloidosis (AL) Increased production of monoclonal light chain Monoclonal gammopathies, including myeloma, benign gammopathies and plasmacytoma Restrictive cardiomyopathy, peripheral and autonomic neuropathy, carpal tunnel syndrome, proteinuria, spontaneous purpura, amyloid nodules and plaques Macroglossia occurs rarely but is pathognomonic Prognosis is poor Dialysis-associated (Aβ2M) amyloidosis Accumulation of circulating β2-microglobulin due to failure of renal catabolism in kidney failure Renal dialysis Carpal tunnel syndrome, chronic arthropathy and pathological fractures secondary to amyloid bone cyst formation Manifestations occur 5–10 years after the start of dialysis Senile systemic amyloidosis Normal transthyretin protein deposited in tissues Age > 70 years Feature of normal ageing (affects

> 90% of 90-year-olds) Usually asymptomatic Hereditary systemic amyloidosis 20 forms of hereditary systemic amyloidosis Production of protein with an abnormal structure that predisposes to amyloid fibril formation. Most commonly due to mutations in transthyretin gene Autosomal dominant inheritance Peripheral and autonomic neuropathy, cardiomyopathy Renal involvement unusual 10% of gene carriers are asymptomatic throughout life

Investigations Autoantibodies Many autoantibodies have been identified and are used in the diagnosis and monitoring of autoimmune diseases, as discussed elsewhere in this book. Antibodies can be quantified either by titre (the maximum dilution of the serum at which the antibody can be detected) or by concentration in standardised units using an enzyme-linked immunosorbent assay (ELISA) in which the antigen is used to coat microtitre plates to which the patient's serum is added (Fig. 4.13A). Qualitative tests are also employed for antinuclear antibodies in which the pattern of nuclear staining is recorded (Fig. 4.13B). they have very limited predictive value and are generally not useful in determining management of individual patients. Several environmental factors may be associated with autoimmunity in genetically predisposed individuals, including infection, cigarette smoking and hormone levels. The most widely studied of these is infection, as occurs in acute rheumatic fever following streptococcal infection or reactive arthritis following bacterial infection. Several mechanisms have been invoked to explain the autoimmunity that occurs after an infectious trigger. These include crossreactivity between proteins expressed by the pathogen and the host (molecular mimicry), such as Guillain–Barré syndrome and Campylobacter infection (p. 1140); release of sequestered antigens from tissues that are damaged during infections that are not usually visible to the immune system; and production of inflammatory cytokines that overwhelm the normal control mechanisms that prevent bystander damage. Occasionally, autoimmune disease may be an adverse effect of drug treatment. For example, metabolic products of the anaesthetic agent halothane can bind to liver enzymes, resulting in a structurally novel protein that is recognised as a foreign antigen by the immune system. This can provoke the development of autoantibodies and activated T cells, which can cause hepatic necrosis. Clinical features The clinical presentation of autoimmune disease is highly variable. Autoimmune diseases can be classified by organ involvement or by the predominant mechanism responsible for tissue damage. The Gell and Coombs classification of hypersensitivity is the most widely used, and distinguishes four types of immune response that result in tissue damage (Box 4.20). • Type I hypersensitivity is relevant in allergy but is not associated with autoimmune disease. • Type II hypersensitivity causes injury to a single tissue or organ and is mediated by specific autoantibodies. • Type III hypersensitivity results from deposition of immune complexes, which initiates activation of the classical complement cascade, as well as recruitment and activation of phagocytes and CD4+ lymphocytes. The site of immune complex deposition is determined by the relative amount of antibody, size of the immune complexes, nature of the antigen and local haemodynamics. Generalised deposition of immune complexes gives rise to systemic diseases such as SLE. • Type IV hypersensitivity is mediated by activated T cells and macrophages, which together cause tissue damage. 4.20 Gell and Coombs classification of hypersensitivity diseases Type Mechanism Example of disease in response to exogenous agent Example of autoimmune

disease Type I Immediate hypersensitivity IgE-mediated mast cell degranulation Allergic disease None described Type II Antibody-mediated Binding of cytotoxic IgG or IgM antibodies to antigens on cell surface causes cell killing ABO blood transfusion reaction Hyperacute transplant rejection Autoimmune haemolytic anaemia Idiopathic thrombocytopenic purpura Goodpasture's disease Type III Immune complex-mediated IgG or IgM antibodies bind soluble antigen to form immune complexes that trigger classical complement pathway activation Serum sickness Farmer's lung Systemic lupus erythematosus Cryoglobulinaemia Type IV Delayed type Activated T cells, and phagocytes Acute cellular transplant rejection Nickel hypersensitivity Type 1 diabetes Hashimoto's thyroiditis Fig. 4.13 Autoantibody testing. A Measurement of antibody levels by enzyme-linked immunosorbent assay (ELISA). The antigen of interest is used to coat microtitre plates to which patient serum is added. If autoantibodies are present, these bind to the target antigen on the microtitre plate. The amount of bound antibody is quantitated by adding a secondary antibody linked to an enzyme that converts a colourless substrate to a coloured one, which can be detected by a plate reader. B Qualitative analysis of autoantibodies by patterns of nuclear staining. In this assay, patient serum is added to cultured cells and a secondary antibody is added with a fluorescent label to detect any bound antibody. If antinuclear antibodies are present, they are detected as bright green staining. Different antinuclear antibody patterns may be seen in different types of connective tissue disease (Ch. 24). B (Nucleolar and Homogenous), Courtesy of Juliet Dunphy, Biomedical Scientist, Royal United Hospital Bath, previously of Bath Institute of Rheumatic Diseases, UK; (Speckled), Courtesy of Mr Richard Brown, Clinical Scientist in Immunology, Southwest Pathology Services, UK. Antibodies bind to target Target antigen Wash Detection of bound antibody Quantitate on plate reader Target antigen Wash Nucleolar Homogenous Speckled A B

 They comprise a range of disorders from mild to life-threatening and affect many organs. Atopy is the tendency to produce an exaggerated IgE immune response to otherwise harmless environmental substances, while an allergic disease can be defined as the clinical manifestation of this inappropriate IgE immune response. Pathophysiology The immune system does not normally respond to the many environmental substances to which it is exposed on a daily basis. In allergic individuals, however, an initial exposure to a normally harmless exogenous substance (known as an allergen) triggers the production of specific IgE antibodies by activated B cells. These bind to high-affinity IgE receptors on the surface of mast cells, a step that is not itself associated with clinical sequelae. However, re-exposure to the allergen binds to and cross-links membrane-bound IgE, which activates the mast cells, releasing a variety of vasoactive mediators (the early phase response; Fig. 4.14 and see Box 4.9). This type I hypersensitivity reaction forms the basis of an allergic reaction, which can range from sneezing and rhinorrhoea to anaphylaxis (Box 4.22). In some individuals, the early phase response is followed by persistent activation of mast cells, manifest by ongoing swelling and local inflammation. This is known as the late phase reaction and is mediated by mast cell metabolites, basophils, eosinophils and macrophages. Long-standing or recurrent allergic inflammation may give rise to a chronic inflammatory response characterised by a complex infiltrate of macrophages, eosinophils and T lymphocytes, in addition to mast cells and basophils. Once this has been established, inhibition of mast cell mediators with antihistamines is clinically ineffective in isolation. Mast cell activation may also be non-specifically triggered through other signals, such as neuropeptides, anaphylotoxins and bacterial peptides. The increasing incidence of allergic diseases is largely unexplained but one widely held theory is the 'hygiene hypothesis'. This proposes that infections in early life are critically important in

maturation of the immune response and bias the immune system against the development of allergies; the high prevalence Complement Measurement of complement components can be useful in the evaluation of immune complex-mediated diseases. Classical complement pathway activation leads to a decrease in circulating C4 levels and is often also associated with decreased C3 levels. Serial measurement of C3 and C4 is a useful surrogate measure of disease activity in conditions such as SLE. Cryoglobulins Cryoglobulins are antibodies directed against other immunoglobulins, forming immune complexes that precipitate in the cold. They can lead to type III hypersensitivity reactions, with typical clinical manifestations including purpuric rash, often of the lower extremities, arthralgia and peripheral neuropathy. Cryoglobulins are classified into three types, depending on the properties of the immunoglobulin involved (Box 4.21). Testing for cryoglobulins requires the transport of a serum specimen to the laboratory at 37°C. Cryoglobulins should not be confused with cold agglutinins; the latter are autoantibodies specifically directed against the I/i antigen on the surface of red cells, which can cause intravascular haemolysis in the cold (p. 950). Management The management of autoimmune disease depends on the organ system involved and further details are provided elsewhere in this book. In general, treatment of autoimmune diseases involves the use of glucocorticoids and immunosuppressive agents, which are increasingly used in combination with biologic agents targeting disease-specific cytokines and their receptors. Not all conditions require immune suppression, however. For example, the management of coeliac disease involves dietary gluten withdrawal, while autoimmune hypothyroidism requires appropriate thyroxine supplementation. Allergy Allergic diseases are a common and increasing cause of illness, affecting between 15% and 20% of the population at some time. 4.21 Classification of cryoglobulins Type I Type II Type III Immunoglobulin (Ig) isotype and specificity Isolated monoclonal IgM paraprotein with no particular specificity Immune complexes formed by monoclonal IgM paraprotein directed towards constant region of IgG Immune complexes formed by polyclonal IgM or IgG directed towards constant region of IgG Prevalence 25% 25% 50% Disease association Lymphoproliferative disease, especially Waldenström macroglobulinaemia (p. 966) Infection, particularly hepatitis C; lymphoproliferative disease Infection, particularly hepatitis C; autoimmune disease, including rheumatoid arthritis and systemic lupus erythematosus Symptoms Hyperviscosity: Raynaud’s phenomenon Acrocyanosis Retinal vessel occlusion Arterial and venous thrombosis Small-vessel vasculitis: Purpuric rash Arthralgia Neuropathy Cutaneous ulceration, hepatosplenomegaly, glomerulonephritis, Raynaud’s phenomenon Small-vessel vasculitis: Purpuric rash, arthralgia Cutaneous ulceration Hepatosplenomegaly, glomerulonephritis Raynaud’s phenomenon Protein electrophoresis Monoclonal IgM paraprotein Monoclonal IgM paraprotein No monoclonal paraprotein Rheumatoid factor Negative Strongly positive Strongly positive Complement Usually normal Decreased C4 Decreased C4 Serum viscosity Raised Normal Normal

of insect venom frequently stimulates the production of IgE antibodies, and thus may be followed by allergic reactions to single stings. Allergic IgE-mediated reactions vary from mild to life-threatening. Antigen-specific immunotherapy (desensitisation; see below) with bee or wasp venom can reduce the incidence of recurrent anaphylaxis from 50–60% to approximately 10% but requires up to 5 years of treatment. Peanut allergy Peanut allergy is the most common food-related allergy. More than 50% of patients present before the age of 3 years and some individuals react to their first known exposure to peanuts, thought to result from sensitisation to arachis oil in topical

creams. Peanuts are ubiquitous in the Western diet, and every year up to 25% of peanut-allergic individuals experience a reaction as a result of inadvertent exposure. Birch oral allergy syndrome This syndrome is characterised by the combination of birch pollen hay fever and local oral symptoms, including itch and angioedema, after contact with certain raw fruits, raw vegetables and nuts. Cooked fruits and vegetables are tolerated without difficulty. It is due to shared or cross-reactive allergens that are destroyed by cooking or digestion, and can be confirmed by skin prick testing using fresh fruit. Severe allergic reactions are unusual. Diagnosis When assessing a patient with a complaint of allergy, it is important to identify what the patient means by the term, as up to 20% of the UK population describe themselves as having a food allergy; in fact, less than 1% have true allergy, as defined by an IgE-mediated hypersensitivity reaction confirmed on double-blind challenge. The nature of the symptoms should be established and specific triggers identified, along with the predictability of a reaction, and the time lag between exposure to a potential allergen and onset of symptoms. An allergic reaction usually occurs within minutes of exposure and provokes predictable, reproducible symptoms such as angioedema, urticaria and wheezing. Specific enquiry should be made about other allergic symptoms, past and present, and about a family history of allergic disease. Potential allergens in the home and workplace should be identified. A detailed drug history should always be taken, including details of adherence to medication, possible adverse effects and the use of over-the-counter or complementary therapies. of allergic disease is the penalty for the decreased exposure to infection that has resulted from improvements in sanitation and health care. Genetic factors also contribute strongly to the development of allergic diseases. A positive family history is common in patients with allergy, and genetic association studies have identified a wide variety of predisposing variants in genes controlling innate immune responses, cytokine production, IgE levels and the ability of the epithelial barrier to protect against environmental agents. The expression of a genetic predisposition is complex; it is governed by environmental factors, such as pollutants and cigarette smoke, and the incidence of bacterial and viral infection. Clinical features Common presentations of allergic disease are shown in Box 4.22. Those that affect the respiratory system and skin are discussed in more detail in Chapters 17 and 29, respectively. Here we focus on general principles of the approach to the allergic patient, some specific allergies and anaphylaxis. Insect venom allergy Local non-IgE-mediated reactions to insect stings are common and may cause extensive swelling around the site lasting up to 7 days. These usually do not require specific treatment. Toxic reactions to venom after multiple (50–100) simultaneous stings may mimic anaphylaxis. In addition, exposure to large amounts Fig. 4.14 Type I (immediate) hypersensitivity response. A After an encounter with allergen, B cells produce immunoglobulin E (IgE) antibody against the allergen. B Specific IgE antibodies bind to circulating mast cells via high-affinity IgE cell surface receptors. C On re-encounter with allergen, the allergen binds to the IgE antibody-coated mast cells. This cross-linking of the IgE triggers mast cell activation with release of vasoactive mediators (see Box 4.9). B B B T T B B B B Allergen T and B cells IgE antibody IgE receptor Mast cell Histamine, tryptase and vasoactive peptides A B C 4.22 Clinical manifestations of allergy Dermatological • Urticaria • Atopic eczema if chronic • Allergic contact eczema • Angioedema Respiratory • Asthma • Atopic rhinitis Ophthalmological • Allergic conjunctivitis Gastrointestinal • Food allergy Other • Anaphylaxis • Drug allergy • Allergy to insect venom

 causes, including parasitic and helminth infections (pp. 299 and 288), lymphoma (p. 961), drug reactions and eosinophilic granulomatosis with polyangiitis (previously known as Churg–Strauss vasculitis; p. 1043). Normal total IgE levels do not exclude allergic

disease. Eosinophilia Peripheral blood eosinophilia is common in atopic individuals but lacks specificity. Eosinophilia of more than 20% or an absolute eosinophil count over 1.5 × 109/L should initiate a search for a non-atopic cause, such as eosinophilic granulomatosis with polyangiitis or parasitic infection (p. 928). Management Several approaches can be deployed in the management of allergic individuals, as discussed below. Avoidance of the allergen This is indicated in all cases and should be rigorously attempted, with the advice of specialist dietitians and occupational physicians if necessary. Antihistamines Antihistamines are useful in the management of allergy as they inhibit the effects of histamine on tissue H1 receptors. Long-acting, non-sedating preparations are particularly useful for prophylaxis. Glucocorticoids These are highly effective in allergic disease, and if used topically, adverse effects can be minimised. Sodium cromoglicate Sodium cromoglicate stabilises the mast cell membrane, inhibiting release of vasoactive mediators. It is effective as a prophylactic agent in asthma and allergic rhinitis but has no role in management of acute attacks. It is poorly absorbed and therefore ineffective in the management of food allergies. Antigen-specific immunotherapy This involves the sequential administration of increasing doses of allergen extract over a prolonged period of time. The mechanism of action is not fully understood but it is highly effective in the prevention of insect venom anaphylaxis and of allergic rhinitis secondary to grass pollen. The traditional route of administration is by subcutaneous injection, which carries a risk of anaphylaxis and should be performed only in specialised centres. Sublingual immunotherapy is also increasingly used. Clinical studies to date do not support the use of allergen immunotherapy for food hypersensitivity, although this is an area of active investigation. Omalizumab Omalizumab is a monoclonal antibody directed against IgE; it inhibits the binding of IgE to mast cells and basophils. It is licensed for treatment of refractory chronic spontaneous urticaria and also for severe persistent allergic asthma that has failed to respond to standard therapy (p. 572). The dose and frequency are determined by baseline IgE (measured before the start of treatment) and body weight. It is under investigation for allergic rhinitis but not yet approved for this indication. Adrenaline (epinephrine) Adrenaline given by injection in the form of a pre-loaded selfinjectable device can be life-saving in the acute management of anaphylaxis (see Box 4.12). Investigations Skin-prick tests Skin-prick testing is a key investigation in the assessment of patients suspected of having allergy. A droplet of diluted standardised allergen is placed on the forearm and the skin is superficially punctured through the droplet with a sterile lancet. Positive and negative control material must be included in the assessment. After 15 minutes, a positive response is indicated by a local weal and flare response 2 mm or more larger than the negative control. A major advantage of skin-prick testing is that patient can clearly see the results, which may be useful in gaining adherence to avoidance measures. Disadvantages include the remote risk of a severe allergic reaction, so resuscitation facilities should be available. Results are unreliable in patients with extensive skin disease. Antihistamines inhibit the magnitude of the response and should be discontinued for at least 3 days before testing; low-dose glucocorticoids do not influence test results. A number of other prescribed medicines can also lead to false-negative results, including amitriptyline and risperidone. Specific IgE tests An alternative to skin-prick testing is the quantitation of IgE directed against the suspected allergen. The sensitivity and specificity of specific IgE tests (previously known as radioallergosorbent tests, RAST) are lower than those of skin-prick tests. However, IgE tests may be very useful if skin testing is inappropriate, such as in patients taking antihistamines or those with severe skin disease or dermatographism. They can also be used to test for cross-reactivity – for example, with multiple insect venoms, where component-resolved diagnostics, using recombinant allergens, is increasingly used rather than crude allergen extract. Specific IgE tests can also be used post-mortem to identify allergens

responsible for lethal anaphylaxis. Supervised exposure to allergen Tests involving supervised exposure to an allergen (allergen challenge) are usually performed in specialist centres on carefully selected patients, and include bronchial provocation testing, nasal challenge, and food or drug challenge. These may be particularly useful in the investigation of occupational asthma or food allergy. Patients can be considered for challenge testing when skin tests and/or IgE tests are negative, as they can be helpful in ruling out allergic disease. Mast cell tryptase Measurement of serum mast cell tryptase is extremely useful in investigating a possible anaphylactic event. Ideally, measurements should be made at the time of the reaction following appropriate resuscitation, and 3 hours and 24 hours later. The basis of the test is the fact that circulating levels of mast cell degranulation products rise dramatically to peak 1–2 hours after a systemic allergic reaction. Tryptase is the most stable of these and is easily measured in serum. Serum total IgE Serum total IgE measurements are not routinely indicated in the investigation of allergic disease, other than to aid in the interpretation of specific IgE results, as false-positive specific IgEs are common in patients with atopy, who often have a high total IgE level. Although atopy is the most common cause of an elevated total IgE in developed countries, there are many other

diagnosis. If no obvious trigger can be identified, measurement of complement C4 is useful in differentiating hereditary and acquired angioedema from other causes. If C4 levels are low, further investigations should be initiated to look for evidence of C1 inhibitor deficiency. Management Management depends on the underlying cause. Angioedema associated with allergen exposure generally responds to antihistamines and glucocorticoids. Following acute management of angioedema secondary to drug therapy, drug withdrawal Angioedema Angioedema is an episodic, localised, non-pitting swelling of submucous or subcutaneous tissues. Pathophysiology The causes of angioedema are summarised in Box 4.23. It may be a manifestation of allergy or non-allergic degranulation of mast cells in response to drugs and toxins. In these conditions the main cause is mast cell degranulation with release of histamine and other vasoactive mediators. In hereditary angioedema, the cause is C1 inhibitor deficiency, which causes increased local release of bradykinin. Angiotensin-converting enzyme (ACE) inhibitor-induced angioedema also occurs as the result of increased bradykinin levels due to inhibition of its breakdown. Clinical features Angioedema is characterised by soft-tissue swelling that most frequently affects the face (Fig. 4.15) but can also affect the extremities and genitalia. Involvement of the larynx or tongue may cause life-threatening respiratory tract obstruction, and oedema of the intestinal mucosa may cause abdominal pain and distension. Investigations Differentiating the mechanism of angioedema is important in determining the most appropriate treatment. A clinical history of allergy or drug exposure can give clues to the underlying 4.23 Types of angioedema Allergic reaction to specific trigger Idiopathic angioedema Hereditary angioedema ACE-inhibitor associated angioedema Pathogenesis IgE-mediated degradation of mast cells Non-IgE-mediated degranulation of mast cells C1 inhibitor deficiency, with resulting increased local bradykinin concentration Inhibition of breakdown of bradykinin Key mediator Histamine Histamine Bradykinin Bradykinin Prevalence Common Common Rare autosomal dominant disorder 0.1–0.2% of patients treated with ACE inhibitors Clinical features Usually associated with urticaria History of other allergies common Follows exposure to specific allergen, in food, animal dander or insect venom Usually associated with urticaria May be triggered by physical stimuli such as heat, pressure or exercise Dermatographism common Occasionally associated with underlying infection or thyroid disease

Not associated with urticaria or other features of allergy Does not cause anaphylaxis May cause life-threatening respiratory tract obstruction Can cause severe abdominal pain Not associated with urticaria Does not cause anaphylaxis Usually affects the head and neck, and may cause life-threatening respiratory tract obstruction Can occur years after the start of treatment Investigations Specific IgE tests or skin-prick tests Specific IgE tests and skin-prick tests often negative Hypothyroidism should be excluded Complement C4 (invariably low in acute attacks) C1 inhibitor levels No specific investigations Treatment Allergen avoidance Antihistamines Antihistamines are mainstay of treatment and prophylaxis Unresponsive to antihistamines Anabolic steroids C1 inhibitor concentrate or icatibant for acute attacks ACE inhibitor should be discontinued ARBs should be avoided if possible unless there is a strong indication Associated drug reactions Specific drug allergies NSAIDs Opioids, radiocontrast media ACE inhibitors, ARBs (ACE = angiotensin-converting enzyme; ARBs = angiotensin II receptor blockers; NSAIDs = non-steroidal anti-inflammatory drugs) Fig. 4.15 Angioedema. This young man has hereditary angioedema. A Normal appearance. B During an acute attack. From Helbert M. Flesh and bones of immunology. Edinburgh: Churchill Livingstone, Elsevier Ltd; 2006. B A

88 • CLINICAL IMMUNOLOGY should prevent further attacks, although ACE inhibitor-induced angioedema can continue for a limited period post drug withdrawal. Management of angioedema associated with C1 inhibitor deficiency is discussed below. Hereditary angioedema Hereditary angioedema (HAE), also known as inherited C1 inhibitor deficiency, is an autosomal dominant disorder caused by decreased production or activity of C1 inhibitor protein. This complement regulatory protein inhibits spontaneous activation of the classical complement pathway (see Fig. 4.4). It also acts as an inhibitor of the kinin cascade, activation of which increases local bradykinin levels, giving rise to local pain and swelling. Clinical features The angioedema in HAE may be spontaneous or triggered by local trauma or infection. Multiple parts of the body may be involved, especially the face, extremities, upper airway and gastrointestinal tract. Oedema of the intestinal wall causes severe abdominal pain and many patients with undiagnosed HAE undergo exploratory laparotomy. The most important complication is laryngeal obstruction, often associated with minor dental procedures, which can be fatal. Episodes of angioedema are self-limiting and usually resolve within 48 hours. Patients with HAE generally present in adolescence but may go undiagnosed for many years. A family history can be identified in 80% of cases. HAE is not associated with allergic diseases and is specifically not associated with urticaria. Investigations Acute episodes are accompanied by low C4 levels; a low C4 during an episode of angioedema should therefore trigger further investigation. The diagnosis can be confirmed by measurement of C1 inhibitor levels and function. Management Severe acute attacks should be treated with purified C1 inhibitor concentrate or the bradykinin receptor antagonist icatibant. Anabolic steroids, such as danazol, can be used to prevent attacks and act by increasing endogenous production of complement proteins. Tranexamic acid can be helpful as prophylaxis in some patients. Patients can be taught to self-administer therapy and should be advised to carry a MedicAlert or similar. Acquired C1 inhibitor deficiency This rare disorder is clinically indistinguishable from HAE but presents in late adulthood. It is associated with autoimmune and lymphoproliferative diseases. Most cases are due to the 4.24 Immunological diseases in pregnancy Allergic disease • Maternal dietary restrictions during pregnancy or lactation: current evidence does not support these for prevention of allergic disease. • Breastfeeding for at least 4 months: prevents or delays the occurrence of atopic dermatitis, cow’s milk allergy and wheezing in early childhood, as compared with feeding formula milk containing intact cow’s milk protein. Autoimmune disease • Suppressed T-cell-mediated immune responses in

pregnancy: may suddenly reactivate post-partum. Some autoimmune diseases may improve during pregnancy but flare immediately after delivery. Systemic lupus erythematosus (SLE) is an exception, however, as it is prone to exacerbation in pregnancy or the puerperium. • Passive transfer of maternal antibodies: can mediate autoimmune disease in the fetus and newborn, including SLE, Graves' disease and myasthenia gravis. • Antiphospholipid syndrome (p. 977): an important cause of fetal loss, intrauterine growth restriction and pre-eclampsia. • HIV in pregnancy: see p. 326. 4.25 Classification of transplant rejection Type Time Pathological findings Mechanism Treatment Hyperacute rejection Minutes to hours Thrombosis, necrosis Pre-formed antibody to donor antigens results in complement activation (type II hypersensitivity) None – irreversible graft loss Acute cellular rejection 5–30 days Cellular infiltration CD4+ and CD8+ T cells (type IV hypersensitivity) Increase immunosuppression Acute vascular rejection 5–30 days Vasculitis Antibody and complement activation Increase immunosuppression Chronic allograft failure

> 30 days Fibrosis, scarring Immune and non-immune mechanisms Minimise drug toxicity, control hypertension and hyperlipidaemia development of autoantibodies to C1 inhibitor, but the condition can also be caused by autoantibodies that activate C1. Treatment of the underlying disorder may induce remission of angioedema. As with HAE, a low C4 is seen during acute episodes. Transplantation and graft rejection Transplantation provides the opportunity for definitive treatment of end-stage organ disease. The major complications are graft rejection, drug toxicity and infection consequent to immunosuppression. Transplant survival continues to improve, as a result of the introduction of less toxic immunosuppressive agents and increased understanding of the processes of transplant rejection. Stem cell transplantation and its complications are discussed on page 936. Transplant rejection Solid organ transplantation inevitably stimulates an aggressive immune response by the recipient, unless the transplant is between monozygotic twins. The type and severity of the rejection response is determined by the genetic disparity between the donor and recipient, the immune status of the host and the nature of the tissue transplanted (Box 4.25). The most important genetic determinant

Investigations Pre-transplantation testing HLA typing determines an individual's HLA polymorphisms and facilitates donor–recipient matching. Potential transplant recipients are also screened for the presence of anti-HLA antibodies. The recipient is excluded from receiving a transplant that carries these alleles. Donor–recipient cross-matching is a functional assay that directly tests whether serum from a recipient (which potentially contains anti-donor antibodies) is able to bind and/or kill donor lymphocytes. It is specific to a prospective donor–recipient pair and is done immediately prior to transplantation. A positive cross-match is a contraindication to transplantation because of the risk of hyperacute rejection. Post-transplant biopsy: C4d staining C4d is a fragment of the complement protein C4 (see Fig. 4.4). Deposition of C4d in graft capillaries indicates local activation of the classical complement pathway and provides evidence of antibody-mediated damage. This is useful in the early diagnosis of vascular rejection. Complications of

transplant immunosuppression Transplant recipients require indefinite treatment with immunosuppressive agents. In general, two or more immunosuppressive drugs are used in synergistic combination in order to minimise adverse effects (Box 4.26). The major complications of long-term immunosuppression are infection and malignancy. The risk of some opportunistic infections may be minimised through the use of prophylactic medication, such as ganciclovir for cytomegalovirus prophylaxis and trimethoprim–sulfamethoxazole for Pneumocystis prophylaxis. Immunisation with killed vaccines is appropriate, although the immune response may be curtailed. Live vaccines should not be given. is the difference between donor and recipient HLA proteins (p. 67). The extensive polymorphism of these proteins means that donor HLA antigens are almost invariably recognised as foreign by the recipient immune system, unless an active attempt has been made to minimise incompatibility. • Hyperacute rejection results in rapid and irreversible destruction of the graft (Box 4.25). It is mediated by pre-existing recipient antibodies against donor HLA antigens, which arise as a result of previous exposure through transplantation, blood transfusion or pregnancy. It is very rarely seen in clinical practice, as the use of screening for anti-HLA antibodies and pre-transplant cross-matching ensures the prior identification of recipient–donor incompatibility. • Acute cellular rejection is the most common form of graft rejection. It is mediated by activated T lymphocytes and results in deterioration in graft function. If allowed to progress, it may cause fever, pain and tenderness over the graft. It is usually amenable to increased immunosuppressive therapy. • Acute vascular rejection is mediated by antibody formed de novo after transplantation. It is more curtailed than the hyperacute response because of the use of intercurrent immunosuppression but it is also associated with reduced graft survival. Aggressive immunosuppressive therapy is indicated and physical removal of antibody through plasmapheresis may be indicated in severe causes. Not all post-transplant anti-donor antibodies cause graft damage; their consequences are determined by specificity and ability to trigger other immune components, such as the complement cascade. • Chronic allograft failure, also known as chronic rejection, is a major cause of graft loss. It is associated with proliferation of transplant vascular smooth muscle, interstitial fibrosis and scarring. The pathogenesis is poorly understood but contributing factors include immunological damage caused by subacute rejection, hypertension, hyperlipidaemia and chronic drug toxicity. 4.26 Immunosuppressive drugs used in transplantation Drug Mechanism of action Major adverse effects Anti-proliferative agents Azathioprine, mycophenolate mofetil Inhibit lymphocyte proliferation by blocking DNA synthesis May be directly cytotoxic at high doses Increased susceptibility to infection Leucopenia Hepatotoxicity Calcineurin inhibitors Ciclosporin, tacrolimus Inhibit T-cell signalling; prevent lymphocyte activation; block cytokine transcription Increased susceptibility to infection Hypertension Nephrotoxicity Diabetogenic (especially tacrolimus) Gingival hypertrophy, hirsutism (ciclosporin) Glucocorticoids Decrease phagocytosis and release of proteolytic enzymes; decrease lymphocyte activation and proliferation; decrease cytokine production; decrease antibody production Increased susceptibility to infection Multiple other complications (p. 670) Anti-thymocyte globulin (ATG) Antibodies to cell surface proteins deplete or block T cells Profound non-specific immunosuppression Increased susceptibility to infection Basiliximab Monoclonal antibody directed against CD25 (IL-2Rα chain), expressed on activated T cells Increased susceptibility to infection Gastrointestinal side-effects Belatacept Selectively inhibits T-cell activation through blockade of CTLA4 Increased susceptibility to infection and malignancy Gastrointestinal side-effects Hypertension Anaemia/leucopenia

 arise. The ability of the immune system to kill cancer cells effectively is influenced by tumour immunogenicity and specificity. Many cancer antigens are poorly expressed and specific antigens can mutate, either spontaneously or in response to treatment, which can result in evasion of immune responses. In addition, the inhibitory pathways that are used to maintain self-tolerance and limit collateral tissue damage during antimicrobial immune responses can be co-opted by cancerous cells to evade immune destruction. Recognition and understanding of these immune checkpoint pathways has led to the development of a number of new treatments for cancers that are otherwise refractory to treatment. For example, antibodies to CTLA4, a co-stimulatory molecule normally involved in down-regulation of immune responses, have been licensed for refractory melanoma, and antibodies to PD1 (programmed cell death protein 1) are used in melanoma, non-small-cell lung cancer and renal cell carcinoma. Potential risks include the development of autoimmunity, reflecting the importance of these pathways in the control of self-tolerance. Further information allergy.org.au An Australasian site providing information on allergy, asthma and immune diseases. allergyuk.org UK site for patients and health-care professionals. anaphylaxis.org.uk Provides information and support for patients with severe allergies. info4pi.org A US site managed by the non-profit Jeffrey Modell Foundation, which provides extensive information about primary immune deficiencies. niaid.nih.gov National Institute of Allergy and Infectious Diseases: provides useful information on a variety of allergic diseases, immune deficiency syndromes and autoimmune diseases. The increased risk of malignancy arises because T-cell suppression results in failure to control viral infections associated with malignant transformation. Virus-associated tumours include lymphoma (associated with Epstein–Barr virus), Kaposi’s sarcoma (associated with human herpesvirus 8) and skin tumours (associated with human papillomavirus). Immunosuppression is also linked with a small increase in the incidence of common cancers not associated with viral infection (such as lung, breast and colon cancer), reflecting the importance of T cells in anticancer surveillance. Organ donation The major problem in transplantation is the shortage of organ donors. Cadaveric organ donors are usually previously healthy individuals who experience brainstem death (p. 211), frequently as a result of road traffic accidents or cerebrovascular events. Even if organs were obtained from all potential cadaveric donors, though, their numbers would be insufficient to meet current needs. An alternative is the use of living donors. Altruistic living donation, usually from close relatives, is widely used in renal transplantation. Living organ donation is inevitably associated with some risk to the donor and it is highly regulated to ensure appropriate appreciation of the risks involved. Because of concerns about coercion and exploitation, non-altruistic organ donation (the sale of organs) is illegal in most countries. Tumour immunology Surveillance by the immune system is critically important in monitoring and removing damaged and mutated cells as they

05-5 Population health and epidemiology

5 Population health and epidemiology

Population health and epidemiology H Campbell DA McAllister Global burden of disease and underlying risk factors 92 Life expectancy 92 Global causes of death and disability 92 Risk factors underlying disease 93 Social determinants of health 93 The hierarchy of systems – from molecules to ecologies 93 The life course 93 Preventive medicine 93 Principles of screening 94 Epidemiology 95 Understanding causes and effect 95 Health data/informatics 97

92 • POPULATION HEALTH AND EPIDEMIOLOGY the population to older ages, and this is placing an increasing burden on health systems. For a few conditions (e.g. HIV/AIDS, diabetes mellitus and chronic kidney disease), age-standardised death rates continue to rise. Within this overall pattern, significant regional variations exist: for example, communicable, maternal, neonatal and nutritional causes still account for about two-thirds of premature mortality in sub-Saharan Africa. GBD also provides estimates of disability from disease (Box 5.2). This has raised awareness of the importance of conditions like depression, low back and neck pain, and asthma, which account for a relatively large disease burden but relatively few deaths. This, in turn, has resulted in greater health policy priority being given to these conditions. Since the policy focus in national health systems is increasingly on keeping people healthy rather than only on reducing premature deaths, it is important to have measures of these health outcomes. It is also essential to recognise that, although these estimates represent the best overall picture of burden of disease, they are based on imperfect data. Nevertheless, the quality of data underlying the estimates and the modelling processes are The UK Faculty of Public Health defines public health as ‘the science and art of promoting and protecting health and well-being, preventing ill-health and prolonging life through the organised efforts of society’. This definition recognises that there is a collective responsibility for the health of the population that requires partnerships between government, health services and others to promote and protect health and prevent disease. Population health has been defined as ‘the health outcomes of a group of individuals, including the distribution of such outcomes within the group’. Medical doctors can play a role in all these efforts to improve health both through their clinical work and through their support of broader actions to improve public health. Global burden of disease and underlying risk factors The Global Burden of Disease (GBD) exercise was initiated by the World Bank in 1992, with first estimates appearing in 1993. Regular updated

figures have been published since, together with projections of future disease burden. The aim was to produce reliable and internally consistent estimates of disease burden for all diseases and injuries, and to assess their physiological, behavioural and social risk factors, so that this information could be made available to health workers, researchers and policy-makers. The GBD exercise adopted the metric ‘disability adjusted life year’ (DALY) to describe population health. This combines information about premature mortality in a population (measured as Years of Life Lost from an ‘expected’ life expectancy) and years of life lived with disability (Years of Life lived with Disability, which is weighted by a severity factor). The International Classification of Disease (ICD) rules, which assign one cause to each death, are followed. All estimates are presented by age and sex groups and by regions of the world. Many countries now also report their own national burden of disease data. Life expectancy Global life expectancy at birth increased from 61.7 years in 1980 to 71.8 years in 2015, an increase of 0.29 years per calendar year. This change is due to a substantial fall in child mortality (mainly caused by common infections), partly offset by rises in mortality from adult conditions such as diabetes and chronic kidney disease. Some areas have not shown these increases in life expectancy in men, often due to war and interpersonal violence. Global causes of death and disability Box 5.1 shows a ranked list of the major causes of global premature deaths in 2015. Communicable, maternal, neonatal and nutritional causes accounted for about one-quarter of deaths worldwide, down from about one-third in 1990. In contrast, deaths from non-communicable diseases are increasing in importance and now account for about two-thirds of all deaths globally, including about 13 million from ischaemic heart disease and stroke, and about 8 million from cancer. The age-standardised death rates for most diseases globally are falling. However, despite this, the numbers of deaths from many diseases are rising due to global population growth and the change in age structure of 5.2 Global disability: top 15 ranked causes, 20151,2

1. Lower back and neck pain (1)
2. Sense organ diseases (3)
3. Depressive disorders (4)
4. Iron deficiency anaemia (2)
5. Skin diseases (5)
6. Diabetes (9)
7. Migraine (6)
8. Other musculoskeletal conditions3 (7)
9. Anxiety disorders (8)
10. Oral disorders (11)
11. Asthma (10)
12. Schizophrenia (13)
13. Osteoarthritis (19)
14. Chronic obstructive pulmonary disease (14)
15. Falls (12) 1By Years of Life lived with Disability (YLD). 2Rank in 1990 is shown in brackets. 3Not otherwise classified as specific conditions such as osteoarthritis. 5.1 Global premature mortality: top 15 ranked causes, 20151,2
16. Ischaemic heart disease (4)
17. Cerebrovascular disease (5)
18. Lower respiratory infections (1)
19. Neonatal preterm birth complications (2)

20. Diarrhoeal diseases (3)
21. Neonatal encephalopathy (6)
22. HIV/AIDS (29)
23. Road injuries (10)
24. Malaria (7)
25. Chronic obstructive pulmonary disease (12)
26. Congenital anomalies (9)
27. Tuberculosis (11)
28. Lung cancer3 (20)
29. Self-harm (16)
30. Diabetes (> 30) 1By Years of Life Lost (YLL). 2Rank in 1990 is shown in brackets. 3‘All cancers combined’ would rank in the top three causes.

to higher risk of hypertension and type 2 diabetes in young adults, and of cardiovascular disease in middle age. It has been suggested that under-nutrition during middle to late gestation permanently ‘programs’ cardiovascular and metabolic responses. This ‘life course’ perspective highlights the cumulative effect on health of exposures to illness, adverse environmental conditions and behaviours that damage health. Preventive medicine The complexity of interactions between physical, social and economic determinants of health means successful prevention is often difficult. Moreover, the life-course perspective illustrates that it may be necessary to intervene early in life or even before birth, to prevent important disease later. Successful prevention is likely to require many interventions across the life course and at several levels in the hierarchy of systems. The examples below illustrate this. improving over time and provide an increasingly robust basis for evidence-based health planning and priority setting. Risk factors underlying disease Box 5.3 shows a ranked list of the main risk factors that underlay GBD in 2015 and how this ranking has changed in recent years. Social determinants of health Health emerges from a highly complex interaction between a person’s genetic background and environmental factors (aspects of the physical, biological (microbes), built and social environments, and also distant influences such as the global ecosystem; Fig. 5.1). The hierarchy of systems – from molecules to ecologies Influences on health exist at many levels and extend beyond the individual to include the family, community, population and ecology. Box 5.4 shows an example of this for determinants of coronary heart disease and demonstrates the importance of considering not only the disease process in a patient but also its context. Health care is not the only determinant – and is usually not the major determinant – of health status in the population. The concept of ‘global health’ recognises the global dimension of health problems, whether these be emerging or pandemic infections or global economic influences on health. The life course The determinants of health operate over the whole lifespan. Values and behaviours acquired during childhood and adolescence have a profound influence on educational outcomes, job prospects and risk of disease. These can have a strong influence, for example, on whether a young person takes up damaging behaviour like smoking, risky sexual activity and drug misuse. Influences on health can operate even before birth. Low birth weight can lead Fig. 5.1 Hierarchy of systems that influence population health. Adapted from an original model by Whitehead M, Dahlgren G. What can be done about inequalities in health? Lancet 1991; 338:1059–1063. Macro-economy, politics, culture, global forces Other neighbourhoods, other regions People Age, sex and hereditary factors

L i f e s t y l e

W o r k ,

p l a y D i e t ,

p h y s i c a l

a c t i v i t y

S o c i a l c a p i t a l

C o m m u n i t y

N e t w o r k s W e a lt h

c r e a ti o n

L o c a l e c o n o m y

M a r k e t s B ui ld in g s,

pl a c e s

E n vi r o n m e n t

A ir , w at e r, l a n d Cl i m at e c h a n g e

Gl o b al e c o s y st e m

Bi o di v er si ty 5.3 Global risk factors: top 15 ranked causes, 20151–3

1. High blood pressure (3)
2. Smoking/second-hand smoke exposure (5)
3. High fasting blood glucose (10)
4. High body mass index (13)
5. Childhood underweight (1)
6. Ambient particulate matter pollution (6)
7. High total cholesterol (12)
8. Household air pollution (4)
9. Alcohol use (11)
10. High sodium intake (14)
11. Low wholegrain intake (15)
12. Unsafe sex (20)
13. Low fruit intake (16)
14. Unsafe water (2)

15. Low glomerular filtration rate (21) 1By percentage of burden of disease they cause. 2Rank in 1990 is shown in brackets. 3All dietary risk factors and physical inactivity combined accounted for 10% of global burden of disease. Low physical activity was ranked 21, iron deficiency 16 and suboptimal breastfeeding 22 in 2015. 5.4 ‘Hierarchy of systems’ applied to ischaemic heart disease Level in the hierarchy Example of effect Molecular ApoB mutation causing hypercholesterolaemia Cellular Macrophage foam cells accumulate in vessel wall Tissue Atheroma and thrombosis of coronary artery Organ Ischaemia and infarction of myocardium System Cardiac failure Person Limited exercise capacity, impact on employment Family Passive smoking, diet Community Shops and leisure opportunities Population Prevalence of obesity Society Policies on smoking, screening for risk factors Ecology Agriculture influencing fat content in diet

 of obesity, therefore, we not only need to help those who are already obese but also develop strategies that impact on the whole population and reverse the obesogenic environment. Poverty and affluence The adverse health and social consequences of poverty are well documented: high birth rates, high death rates and short life expectancy. Typically, with industrialisation, the pattern changes: low birth rates, low death rates and longer life expectancy. Instead of infections, chronic conditions such as heart disease dominate in an older population. Adverse health consequences of excessive affluence are also becoming apparent. Despite experiencing sustained economic growth for the last 50 years, people in many industrialised countries are not growing any happier and the litany of socioeconomic problems – crime, congestion, inequality – persists. Many countries are now experiencing a ‘double burden’. They have large populations still living in poverty who are suffering from problems such as diarrhoea and malnutrition, alongside affluent populations (often in cities) who suffer from chronic illness such as diabetes and heart disease. Atmospheric pollution Emissions from industry, power plants and motor vehicles of sulphur oxides, nitrogen oxides, respirable particles and metals are severely polluting cities and towns in Asia, Africa, Latin America and Eastern Europe. Burning of fossil and biomass fuels, with production of short-lived carbon pollutants (SLCPs – methane, ozone, black carbon and hydrofluorocarbons), contributes to increased death rates from respiratory and cardiovascular disease in vulnerable adults, such as those with established respiratory disease and the elderly, while children experience an increase in bronchitic symptoms. Developing countries also suffer high rates of respiratory disease as a result of indoor pollution caused mainly by heating and cooking using solid biomass fuels. Climate change and global warming Climate change is arguably the world’s most important environmental health issue. A combination of habitat destruction and increased production of carbon dioxide and SLCPs, caused primarily by human activity, seems to be the main cause. The temperature of the globe is rising, and if current trends continue, warming by 4°C is predicted by 2050. The climate is being affected, putting millions of people at risk of rising sea levels, flooding, droughts and failed crops These have already claimed millions of lives during the past 20 years and have adversely affected the lives of many more. The economic costs of property damage and the impact on agriculture, food supplies and prosperity have also been substantial. Global warming will also include changes in the geographical range of some vector-borne infectious diseases. Currently, politicians cannot agree an effective framework of actions to tackle the problem, but reducing emissions of CO2 and SLCPs is essential. Principles of screening Screening is the application of a test to a large number of asymptomatic people with the aim of reducing morbidity or mortality from a disease. The World Health Organisation (WHO) Alcohol Alcohol use is an increasingly important risk factor underlying GBD (see Box 5.3). Reasons

for increasing rates of alcohol-related harm vary by place and time but include the falling price of alcohol (in real terms), increased availability and cultural change fostering higher levels of consumption. Public, professional and governmental concern has now led to a minimum price being charged for a unit of alcohol, tightening of licensing regulations and curtailment of some promotional activity in many countries. However, even more aggressive public health measures will be needed to reverse the levels of harm in the population. The approach for individual patients suffering adverse effects of alcohol is described elsewhere (e.g. pp. 1184 and 880). Smoking Smoking is one of the top three risk factors underlying GBD (see Box 5.3). It is responsible for a substantial majority of cases of chronic obstructive pulmonary disease (COPD) and lung cancer (pp. 573 and 598), and most smokers die either from these or from ischaemic heart disease. Smoking also causes cancers of the upper respiratory and gastrointestinal tracts, pancreas, bladder and kidney, and increases risks of peripheral vascular disease, stroke and peptic ulceration. Maternal smoking is an important cause of fetal growth retardation. Moreover, there is evidence that passive ('second-hand') smoking has adverse effects on cardiovascular and respiratory health. The decline in smoking in many high-income countries has been achieved not only by warning people of the health risks but also by increasing taxation of tobacco, banning advertising, legislating against smoking in public places and giving support for smoking cessation to maintain this decline. However, smoking rates remain high in many poorer areas and are increasing among young women. In many developing countries, tobacco companies have found new markets and rates are rising. A complex hierarchy of systems interacts to cause smokers to initiate and maintain their habit. At the molecular and cellular levels, nicotine acts on the nervous system to create dependence and maintain the smoking habit. There are also strong influences at the personal and social level, such as young female smokers being motivated to 'stay thin' or 'look cool' and peer pressure. Other important influences include cigarette advertising, with the advertising budget of the tobacco industry being much greater than that of health services. Strategies to help individuals stop smoking (such as nicotine replacement therapy, anti-smoking advice and behavioural support) are cost-effective and form an important part of the overall strategy. Obesity Obesity is an increasingly important risk factor underlying GBD (see Box 5.3). The weight distribution of almost the whole population is shifting upwards: the slim are becoming less slim while the fat are getting fatter (p. 698). In the UK, this translates into a 1 kg increase in weight per adult per year (on average over the adult population). The current obesity epidemic cannot be explained simply by individual behaviour and poor choice but also requires an understanding of the obesogenic environment that encourages people to eat more and exercise less. This includes the availability of cheap and heavily marketed energy-rich foods, the increase in labour-saving devices (e.g. lifts and remote controls) and the rise in passive transport (cars as opposed to walking, cycling, or walking to public transport hubs). To combat the health impact

has identified a set of ('Wilson and Jungner') criteria to guide health systems in deciding when it is appropriate to implement screening programmes. The essential criteria are: • Is the disease an important public health problem? • Is there a suitable screening test available? • Is there a recognisable latent or early stage? • Is there effective treatment for the disease at this stage that improves prognosis? A suitable screening test is one that is cheap, acceptable, easy to perform and safe, and gives a valid result in terms of sensitivity and specificity (p. 4). Screening programmes should always be evaluated in trials so that robust evidence is provided in favour of their adoption.

These evaluations are prone to several biases – self-selection bias, lead-time bias and length bias – and these need to be accounted for in the analysis. Examples of large-scale programmes in the UK include breast, colorectal and cervical cancer national screening programmes and a number of screening tests carried out in pregnancy and in the newborn, such as the: • diabetic eye screening programme • fetal anomaly screening programme • infectious diseases in pregnancy screening programme • newborn and infant physical examination screening programme • newborn blood spot screening programme • newborn hearing screening programme • sickle-cell and thalassaemia screening programme. These are illustrated in Figure 5.2. Problems with screening include: • over-diagnosis (of a disease that would not have come to attention on its own or would not have led to death) • false reassurance • diversion of resources from investments that could control the disease more cost-effectively. An example of these problems is the use of prostate-specific antigen (PSA) testing as a screening test for the diagnosis of prostate cancer (p. 438). Epidemiology Epidemiologists study disease in free-living humans, seeking to describe patterns of health and disease and to understand how different exposures cause or prevent disease (Box 5.5). Chronic diseases and risk factors (e.g. smoking, obesity etc.) are often described in terms of their prevalence. A prevalence is simply a proportion: e.g. the prevalence of diabetes in people aged 80 and older in developed countries is around 10%. Events such as deaths, hospitalisations and first occurrences of a disease are described using incidence rates: e.g. if there are 100 new cases of a disease in a single year in a population of 1000, the incidence rate is 105 per 1000 person-years, not 100, because of the effect of ‘person-time’. Person-time is the sum of the total ‘exposed’ time for the population and in this example is 950 person-years. The reason person-time is less than 1000 is that 100 people experienced the event. These 100 people are assumed to have had an event, on average, halfway through the time period, removing 100 × 0.5 person-years from the exposure time (as it is not possible to have a first occurrence of a disease twice). Hence, the incidence per 1000 person-years is 105, not 100. A similar measure is the cumulative incidence or risk, which is the number of new cases as a proportion of the total people at risk at the beginning of the exposure time. If, in the example above, the same 1000 people were observed for a year (i.e. with no one joining or leaving the group), then the 1-year risk is 10% (100/1000). The time period should always be specified. These rates and proportions are used to describe how diseases (and risk factors) vary according to time, person and place. Temporal variation may occur seasonally (e.g. malaria occurs in the wet season but not the dry) or as longer-term ‘secular’ trends (e.g. malaria may re-emerge due to drug resistance). Person comparisons include age, sex, socioeconomic status, employment, and lifestyle characteristics. Place comparisons include the local environment (e.g. urban versus rural) and international comparisons. Understanding causes and effect Epidemiological research complements that based on animal, cell and tissue models, the findings of which do not always translate to humans. For example, only a minority of drug discoveries from laboratory research are effective when tested in people. However, differentiating causes from mere non-causal associations is a considerable challenge for epidemiology. This is because while laboratory researchers can directly manipulate conditions to isolate and understand causes, such approaches are impossible in free-living populations. Epidemiologists have developed a different approach, based around a number of study designs (Box 5.6). Of these, the clinical trial is closest to the laboratory experiment. An early example of a clinical trial is shown in Figure 5.3, along with ‘effect measures’, which are used to quantify the difference in rates and risks. In clinical trials, patients are usually allocated randomly to treatments so that, on average, groups are similar, apart from the intervention of interest. Nevertheless, for any particular trial, especially a small trial, the laws of probability mean that differences can and do occur by chance. Poorly

designed or executed trials can also limit comparability between groups. Allocation may not be truly random (e.g. because of inadequate concealment of the randomisation sequence), and there may 5.5 Calculation of risk using descriptive epidemiology Prevalence • The ratio of the number of people with a longer-term disease or condition, at a specified time, to the number of people in the population Incidence • The number of events (new cases or episodes) occurring in the population at risk during a defined period of time Attributable risk • The difference between the risk (or incidence) of disease in exposed and non-exposed populations Attributable fraction • The ratio of the attributable risk to the incidence Relative risk • The ratio of the risk (or incidence) in the exposed population to the risk (or incidence) in the non-exposed population

96 • POPULATION HEALTH AND EPIDEMIOLOGY Fig. 5.2 UK NHS Pregnancy and Newborn Screening Programmes: optimum times for testing. (GA1 = glutaric aciduria type 1; HCU = homocystinuria; IVA = isovaleric acidaemia; MCADD = medium-chain acyl-CoA dehydrogenase deficiency; MSUD = maple syrup urine disease; PKU = phenylketonuria; T13, 18, 21 = trisomy 13, 18 and 21) Based on Version 8.1, March 2016, Gateway ref: 2014696, Public Health England. Blood for sickle cell and thalassaemia Commence folic acid Pre-conception Pregnancy Newborn Blood for haemoglobin, group, Rhesus and antibodies as early as possible, or as soon as a woman arrives for care, including labour Blood for syphilis, hepatitis B and HIV as early as possible, or at any stage of the pregnancy, including labour Reoffer screening for infectious diseases if initially declined Hepatitis B vaccination ± immunoglobulin within 24 hours Repeat haemoglobin and antibodies

+1 +2 +3 +4 +5 +6 Birth Key Week Blood for T21, T18 and T13 (combined test) Blood for T21 (quadruple test) Newborn physical examination by 72 hours Newborn blood spot screens (ideally on day 5) for sickle cell disease (SCD), cystic fibrosis (CF), congenital hypothyroidism (CHT) and inherited metabolic diseases (PKU, MCADD, MSUD, IVA, GA1 and HCU). NB: babies who missed the screen can be tested up to 1 year (except CF offered up to 8 weeks) Early pregnancy scan to support T21, T18 and T13 screening Women with type 1 or type 2 diabetes are offered diabetic eye (DE) screening annually. In pregnancy women with type 1 or type 2 diabetes are offered a DE screen when they first present for care Give screening information as soon as possible Give and discuss newborn screening information Follow-up DE screen for women with type 1 or 2 diabetes found to have diabetic retinopathy Further DE screen for women with type 1 or 2 diabetes Detailed ultrasound scan for structural abnormalities, including T18 and T13 Newborn hearing screen Infant physical examination at 6–8 weeks T21, T18, T13 and fetal anomaly ultrasound Sickle cell and thalassaemia Newborn and infant physical examination Newborn blood spot Infectious diseases in pregnancy Diabetic eye Newborn hearing

Health data/informatics • 97

or more often practical, considerations. Epidemiologists therefore seek to minimise bias and confounding by good study analysis and design. They subsequently make causal inferences by balancing the probability that an observed association has been caused by chance, bias and/or confounding against the alternative probability that the relationship is causal. This weighing-up requires an understanding of the frequency and importance of different sources of bias and confounding, as well as the scientific rationale of the putative causal relationship. It was this approach, collectively and over a number of years, that settled the fact that smoking causes lung cancer and, subsequently, heart disease. Health data/informatics As patients pass through health

and social care systems, data are recorded concerning their family background, lifestyle and disease states, which is of potential interest to healthcare organisations seeking to deliver services, policy-makers concerned with improving health, scientific researchers trying to understand health, and also pharmaceutical and other commercial organisations aiming to identify markets. There is a long tradition of maintaining health information systems. In most countries, registration of births and deaths is required by law, and in the majority, the cause of death is also recorded (Fig. 5.4). There are many challenges in ensuring such data are useful, especially for comparisons across time and place: • A system of standard terminologies is needed, such as the WHO International Classification of Diseases (ICD-10), which provides a list of diagnostic codes attempting to cover every diagnostic entity. • These terms must be understood to refer to the same, or at least similar, diseases in different places. • Access to diagnostic skill and facilities is required. • Standard protocols for assigning clinical diagnoses to ICD-10 codes are needed • Robust quality control processes are needed to maintain some level of data completeness and accuracy. Many countries employ similar systems for hospitalisations, to allow recovery of health-care utilisation costs or to manage and plan services. Similar data are rarely collected for communitybased health care, nor are detailed data on health-care processes generally included in national data systems. Consequently, there has been considerable interest in using data from information technology systems used to deliver care, such as electronic be systematic differences (biases) in the way people allocated to different groups are treated or studied. Such biases also occur in observational epidemiological study designs, such as cohort, case–control and cross-sectional studies (Box 5.6). These designs are also much more subject to the problem of confounding than are randomised trials. Confounding is where the relationship between an exposure and outcome of interest is confused by the presence of some other causal factor. For example, coffee consumption may be associated with lung cancer because smoking is more common among coffee-drinkers. Here, smoking is said to confound the association between coffee and lung cancer. Despite these limitations, for most causes of diseases, randomised controlled trials are not feasible because of ethical, Fig. 5.3 An example of a clinical trial: streptomycin versus bed rest in tuberculosis. Both prevalences and risks are, in fact, proportions, and are therefore frequently expressed as odds. The reasons for doing so are beyond the scope of this text. Enrolled 107 patients with tuberculosis Effect measures Risk ratio (relative risk, RR) Odds ratio (OR) Absolute risk reduction (ARR) Relative risk reduction (RRR) Number needed to treat to prevent one death (NNT= 1/ARR) Random allocation Streptomycin 55 patients Bed rest 52 patients Follow-up and count deaths Events

Risk 7.3% Odds 0.068 Events

Risk 28.8% Odds 0.224 0.25 0.30 21.6% 74.8% 4.6 5.6 Epidemiological study designs Design Description Example Clinical trial Enrols a sample from a population and compares outcomes after randomly allocating patients to an intervention Medical Research Council (MRC) Streptomycin Trial – demonstrated effectiveness of streptomycin in tuberculosis Cohort Enrols a sample from a population and compares outcomes according to exposures Framingham Study – identified risk factors for cardiovascular disease Case–control Enrols cases with an outcome of interest and controls without that outcome and compares exposures between the groups Doll R, Hill AB. Smoking and carcinoma of the lung. British Medical Journal 1950 – demonstrated that smoking caused lung cancer Cross-sectional Enrols a cross-section (sample) of people from the population of interest; obtains data on exposures and outcomes World Health Organisation Demographic and Health Survey – captures risk factor data in a uniform way across many countries

98 • POPULATION HEALTH AND EPIDEMIOLOGY Further information Books and journal articles GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016; 388:1545–1602. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016; 388:1459–1544. GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016; 388:1659–1724. Kindig D, Stoddart G. What is population health? Am J Public Health 2003; 93:380–383. Websites fph.org.uk UK Faculty of Public Health: What is public health? gov.uk UK Government: population screening programmes. patient records, drug-dispensing databases, radiological software and clinical laboratory information systems. Data from such systems are, of course, much less structured than those obtained from vital registrations. Moreover, the completeness of such data depends greatly on local patterns of health-care utilisation, as well as how clinicians and others use information technology systems within different settings. As such, deriving useful, unbiased information from such data is a considerable challenge. Much of the discipline of health informatics is concerned with addressing this challenge. One approach has been to develop comprehensive standard classification systems such as SNOMED-CT, ‘a standardised, multilingual vocabulary of terms relating to the care of the individual’, which has been designed for electronic health-care records. An alternative has been to use statistical methods such as natural language processing to derive information automatically from free text (such as culling diagnoses from radiological reports), or to employ ‘machine learning’, in which software algorithms are applied to data in order to derive useful insights. Such approaches are suited to large, messy data where the costs of systematisation would be prohibitive. It is likely that such innovations will, over the coming years, provide useful information to complement that obtained from more traditional health information systems. Fig. 5.4 Completed death certificate. International Classification of Diseases 10 (ICD-10) codes are appended in red. WHO ICD-10, vol. 2; 1990. Available at https://commons.m.wikimedia.org/wiki/File:International_form_of_medical_certificate_of_cause_of_death.png. Cause of death Approximate interval between onset and death I Disease or condition directly leading to death* INTERNATIONAL FORM OF MEDICAL CERTIFICATE OF CAUSE OF DEATH due to (or as a consequence of) due to (or as a consequence of) due to (or as a consequence of) (a) (b) (c) (d) I21.9 E78.0 J47 *This does not mean the mode of dying, e.g. heart failure, respiratory failure. It means the disease, injury, or complication that caused death. Antecedent causes Morbid conditions, if any, giving rise to the above cause, stating the underlying condition last II Other significant conditions contributing to the death, but not related to the disease or condition causing it

06-6 Principles of infectious disease

6 Principles of infectious disease

100 • PRINCIPLES OF INFECTIOUS DISEASE tackling infection in resource-poor countries. Microorganisms are continually mutating and evolving; the emergence of new infectious agents and antimicrobial-resistant microorganisms is therefore inevitable. This chapter describes the biological and epidemiological principles of infectious diseases and the general approach to their prevention, diagnosis and treatment. Specific infectious diseases are described in Chapters 11–13 and many of the organ-based chapters. Infectious agents The concept of an infectious agent was established by Robert Koch in the 19th century (Box 6.1). Although fulfilment of ‘Koch’s postulates’ became the standard for the definition of an infectious agent, they do not apply to uncultivable organisms (e.g. Mycobacterium leprae, Tropheryma whipplei) or members of the normal human flora (e.g. Escherichia coli, Candida spp.). The following groups of infectious agents are now recognised. Viruses Viruses are incapable of independent replication. Instead, they subvert host cellular processes to ensure synthesis of their nucleic acids and proteins. Viruses’ genetic material (the genome) consists of single- or double-stranded DNA or RNA. Retroviruses transcribe their RNA into DNA in the host cell by reverse transcription. An antigenically unique protein coat (capsid) encloses the genome, and together these form the nucleocapsid. In many viruses, the nucleocapsid is packaged within a lipid envelope. Enveloped viruses are less able to survive in the environment and are spread by respiratory, sexual or blood-borne routes, including arthropod-based transmission. Non-enveloped viruses survive better in the environment and are predominantly transmitted by faecal–oral or, less often, respiratory routes. A generic virus life cycle is shown in Figure 6.2. A virus that infects a bacterium is a bacteriophage (phage). Prokaryotes: bacteria

(including mycobacteria and actinomycetes) Prokaryotic cells are capable of synthesising their own proteins and nucleic acids, and are able to reproduce autonomously, although they lack a nucleus. The bacterial cell membrane is bounded by a peptidoglycan cell wall, which is thick (20–80 nm) in Gram-positive organisms and thin (5–10 nm) in Gram-negative ones. The Gram-negative cell wall is surrounded by an outer membrane containing lipopolysaccharide. Genetic information is contained within a chromosome but bacteria may also contain rings of extra-chromosomal DNA, known as plasmids, which can be transferred between organisms, without cells having to divide. Bacteria may be embedded in a polysaccharide capsule, 'Infection' in its strict sense describes the situation where microorganisms or other infectious agents become established in the host organism's cells or tissues, replicate, cause harm and induce a host response. If a microorganism survives and replicates on a mucosal surface without causing harm or illness, the host is said to be 'colonised' by that organism. If a microorganism survives and lies dormant after invading host cells or tissues, infection is said to be 'latent'. When the infectious agent, or the host response to it, is sufficient to cause illness or harm, then the process is termed an 'infectious disease'. Most pathogens (infectious agents that can cause disease) are microorganisms but some are multicellular organisms. The manifestations of disease may aid pathogen dissemination (e.g. diarrhoea). The term 'infection' is often used interchangeably with 'infectious disease' but not all infections are 'infectious', i.e. transmissible from person to person. Infectious diseases transmitted between hosts are called communicable diseases, whereas those caused by organisms that are already colonising the host are described as endogenous. The distinction is blurred in some situations, including health care-associated infections such as meticillinresistant Staphylococcus aureus (MRSA) or Clostridium difficile infection (CDI), in which colonisation precedes infection but the colonising bacteria may have been recently transmitted between patients. The chain of infection (Fig. 6.1) describes six essential elements for communicable disease transmission. Despite dramatic advances in hygiene, immunisation and antimicrobial therapy, infectious agents still cause a massive burden of disease worldwide. Key challenges remain in Fig. 6.1 Chain of infection. The infectious agent is the organism that causes the disease. The reservoir is the place where the population of an infectious agent is maintained. The portal of exit is the point from which the infectious agent leaves the reservoir. Transmission is the process by which the infectious agent is transferred from the reservoir to the human host, either directly or via a vector or fomite. The portal of entry is the body site that is first accessed by the infectious agent. Finally, in order for disease to ensue, the person to whom the infectious agent is transmitted must be a susceptible host. Susceptible host Exit Entry Transmission Infectious agent Reservoir 6.1 Definition of an infectious agent – Koch's postulates

1. The same organism must be present in every case of the disease
2. The organism must be isolated from the diseased host and grown in pure culture
3. The isolate must cause the disease, when inoculated into a healthy, susceptible animal
4. The organism must be re-isolated from the inoculated, diseased animal

Eukaryotes: fungi, protozoa and helminths Eukaryotic cells contain membrane-bound organelles, including nuclei, mitochondria and Golgi apparatus. Pathogenic eukaryotes are unicellular (e.g. fungi, protozoa) or complex multicellular organisms (e.g. nematodes, trematodes and cestodes, p. 288). and motile bacteria are equipped with flagella. Although many prokaryotes are capable of

independent existence, some (e.g. Chlamydia trachomatis, Coxiella burnetii) are obligate intracellular organisms. Bacteria that can grow in artificial culture media are classified and identified using a range of characteristics (Box 6.2); examples are shown in Figures 6.3 and 6.4. Fig. 6.2 A generic virus life cycle. Life cycle components common to most viruses are host cell attachment and penetration, virus uncoating, nucleic acid and protein synthesis, virus assembly and release. Virus release is achieved either by budding, as illustrated, or by lysis of the cell membrane. Life cycles vary between viruses. Host cell 2 Penetration Receptor-mediated endocytosis or, in some enveloped viruses, membrane fusion (shown here)

Uncoating Nucleic acid is liberated from the phagosome (if endocytosed) and/or capsid by complex enzymatic and/or receptor-mediated processes Interaction between host receptor molecule and virus ligand (determines host-specificity of the virus) Adsorption

Lipid envelope Capsid Nucleic acid Virus Assembly 5 Assembly of virus components is mediated by host and/or viral enzymes Release 6 Complete virus particles are released by budding of host cell membrane (shown here) or disintegration of host cell

Synthesis Nucleic acid and protein synthesis is mediated by host and/or viral enzymes. This takes place in nucleus or cytoplasm, depending on the specific virus Gram stain reaction (see Fig. 6.3) • Gram-positive (thick peptidoglycan layer), Gram-negative (thin peptidoglycan) or unstainable Microscopic morphology • Cocci (round cells) or bacilli (elongated cells) • Presence or absence of capsule Cell association • Association in clusters, chains or pairs Colonial characteristics • Colony size, shape or colour • Effect on culture media (e.g. β-haemolysis of blood agar in haemolytic streptococci; see Fig. 6.4) Atmospheric requirements • Strictly aerobic (requires O2), strictly anaerobic (requires absence of O2), facultatively aerobic (grows with or without O2) or microaerophilic (requires reduced O2) Biochemical reactions • Expression of enzymes (oxidase, catalase, coagulase) • Ability to ferment or hydrolyse various biochemical substrates Motility • Motile or non-motile Antibiotic susceptibility • Identifies organisms with invariable susceptibility (e.g. to optochin in Streptococcus pneumoniae or metronidazole in obligate anaerobes) Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) • A rapid technique that identifies bacteria and some fungi from their specific molecular composition Sequencing bacterial 16s ribosomal RNA gene • A highly specific test for identification of organisms in pure culture and in samples from normally sterile sites Whole-genome sequencing • Although not yet in routine use, whole-genome sequencing (WGS) offers the potential to provide rapid and simultaneous identification, sensitivity testing and typing of organisms from pure culture and/or directly from clinical samples. As such, WGS is likely to replace many of the technologies described above over the next few years (p. 58) 6.2 How bacteria are identified

102 • PRINCIPLES OF INFECTIOUS DISEASE Fig. 6.3 Flow chart for bacterial identification, including Gram film appearances on light microscopy (×100). (MALDI-TOF-MS = matrix-assisted laser desorption/ionisation time-of-flight mass spectroscopy) Gram-positive bacilli Colony morphology, growth characteristics (e.g. growth in anaerobic atmosphere), Gram stain appearance, MALDI-TOF-MS identification Colony morphology, growth characteristics, oxidase reaction, sugar fermentation/MALDI-TOF-MS identification Colony morphology, growth characteristics, lactose fermentation, oxidase reaction, MALDI-TOF-MS identification Colony morphology (e.g. haemolysis), Gram stain appearance, agglutination reactions, coagulase test, catalase Examples Actinomycetes

Arcanobacterium haemolyticum Bacillus spp. Corynebacterium diphtheriae Lactobacillus spp. Listeria monocytogenes Nocardia spp. Clostridium spp. or Examples Neisseria meningitidis Neisseria gonorrhoeae Moraxella catarrhalis Examples Escherichia coli Klebsiella pneumoniae Proteus spp. Enterobacter spp. Serratia spp. Salmonella spp. Shigella spp. Yersinia spp. Vibrio spp. Pseudomonas aeruginosa Gram-positive cocci–clusters Examples Staphylococcus aureus Coagulase-negative staphylococci Gram-negative cocci Gram stain Gram-negative bacilli Gram-positive cocci Gram-positive cocci–chains Examples Oral streptococci Streptococcus pneumoniae (often pairs) Beta-haemolytic streptococci Enterococci (short chains) Fig. 6.4 Beta-haemolytic streptococci (A) and alpha-haemolytic streptococci (B) spread on each half of a blood agar plate (backlit). This image is half life size. ×0.5. Beta-haemolysis renders the agar transparent around the colonies (A) and alpha-haemolysis imparts a green tinge to the agar (B). B A Fungi exist as either moulds (filamentous fungi) or yeasts. Dimorphic fungi exist in either form, depending on environmental conditions (see Fig. 11.59, p. 300). The fungal plasma membrane differs from the human cell membrane in that it contains the sterol, ergosterol. Fungi have a cell wall made up of polysaccharides, chitin and mannoproteins. In most fungi, the main structural component of the cell wall is β-1,3-D-glucan, a glucose polymer. These differences from mammalian cells are important because they offer useful therapeutic targets. Protozoa and helminths are often referred to as parasites. Many parasites have complex multi-stage life cycles, which involve animal and/or plant hosts in addition to humans. Prions Although prions are transmissible and have some of the characteristics of infectious agents, they are not microorganisms and are not diagnosed in microbiology laboratories. Prions are covered on page 250. Normal microbial flora The human body is colonised by large numbers of microorganisms (collectively termed the human microbiota). These colonising

Fig. 6.5 Human non-sterile sites and normal flora in health. Pharynx Haemophilus spp. Moraxella catarrhalis Neisseria spp. (including N. meningitidis) Staph. aureus Strep. pneumoniae Strep. pyogenes (group A) Oral streptococci (α-haemolytic) Oral cavity Oral streptococci (α-haemolytic) Anaerobic Gram-positive bacilli (including Actinomyces spp.) Anaerobic Gram-negative bacilli Prevotella spp. Fusobacterium spp. Candida spp. Small bowel Distally, progressively increasing numbers of large bowel bacteria Candida spp. Large bowel Enterobacteriaceae Escherichia coli Klebsiella spp. Enterobacter spp. Proteus spp. Enterococci E. faecalis E. faecium Streptococcus anginosus group Strep. anginosus Strep. intermedius Strep. constellatus Anaerobic Gram-positive bacilli Clostridium spp. Anaerobic Gram-negative bacilli Bacteroides spp. Prevotella spp. Candida spp. Scalp As for skin Nares Staph. aureus Coagulase-negative staphylococci Skin Coagulase-negative staphylococci Staph. aureus Corynebacterium spp. Propionibacterium spp. Malassezia spp. Hands Resident: as for skin Transient: skin flora (including meticillin-resistant and other Staph. aureus), bowel flora (including Clostridium difficile, Candida spp. and Enterobacteriaceae) Vagina Lactobacillus spp. Staph. aureus Candida spp. Enterobacteriaceae Strep. agalactiae (group B) Perineum As for skin As for large bowel bacteria, also referred to as the ‘normal flora’, are able to survive and replicate on skin and mucosal surfaces. The gastrointestinal tract and the mouth are the two most heavily colonised sites in the body and their microbiota are distinct, in both composition and function. Knowledge of non-sterile body sites and their normal flora is required to inform microbiological sampling strategies and interpret culture results (Fig. 6.5). The microbiome is the total burden of microorganisms, their genes and their environmental interactions, and is now

recognised to have a profound influence over human health and disease. Maintenance of the normal flora is beneficial to health. For example, lower gastrointestinal tract bacteria synthesise and excrete vitamins (e.g. vitamins K and B12); colonisation with normal flora confers ‘colonisation resistance’ to infection with pathogenic organisms by altering the local environment (e.g. lowering pH), producing antibacterial agents (e.g. bacteriocins (small antimicrobial peptides/proteins), fatty acids and metabolic waste products), and inducing host antibodies that cross-react with pathogenic organisms. Conversely, normally sterile body sites must be kept sterile. The mucociliary escalator transports environmental material deposited in the respiratory tract to the nasopharynx. The urethral sphincter prevents flow from the non-sterile urethra to the sterile bladder. Physical barriers, including the skin, lining of the gastrointestinal tract and other mucous membranes, maintain sterility of the submucosal tissues, blood stream and peritoneal and pleural cavities, for example. The normal flora contribute to endogenous disease mainly by translocation to a sterile site but excessive growth at the ‘normal’ site (overgrowth) can also cause disease. Overgrowth is exemplified by dental caries, vaginal thrush and ‘blind loop’ syndrome (p. 808). Translocation results from spread along a surface or penetration though a colonised surface, e.g. urinary tract infection caused by perineal/enteric flora, and surgical site infections, particularly of prosthetic materials, caused by skin flora such as staphylococci. Normal flora also contribute to disease by cross-infection, in which organisms that are colonising one individual cause disease when transferred to another, more susceptible, individual. The importance of limiting perturbations of the microbiota by antimicrobial therapy is increasingly recognised. Probiotics are microbes or mixtures of microbes that are given to a patient to prevent or treat infection and are intended to restore a beneficial profile of microbiota. Although probiotics have been used in a number of settings, whether they have demonstrable clinical benefits remains a subject of debate.

104 • PRINCIPLES OF INFECTIOUS DISEASE Histoplasma capsulatum), are able to survive in intracellular environments, including after phagocytosis by macrophages. Pathogenic bacteria express different genes, depending on environmental stress (pH, iron starvation, O2 starvation etc.) and anatomical location. Genetic diversity enhances the pathogenic capacity of bacteria. Some virulence factor genes are found on plasmids or in phages and are exchanged between different strains or species. The ability to acquire genes from the gene pool of all strains of the species (the ‘bacterial supragenome’) increases diversity and the potential for pathogenicity. Viruses exploit their rapid reproduction and potential to exchange nucleic acid with host cells to enhance diversity. Once a strain acquires a particularly effective combination of virulence genes, it may become an epidemic strain, accounting for a large subset of infections in a particular region. This phenomenon accounts for influenza pandemics (see Box 6.10). The host response Innate and adaptive immune and inflammatory responses, which humans use to control the normal flora and respond to pathogens, are reviewed in Chapter 4. Pathogenesis of infectious disease The harmful manifestations of infection are determined by a combination of the virulence of the organism and the host response to infection. Despite the obvious benefits of an intact host response, an excessive response is undesirable. Cytokines and antimicrobial factors contribute to tissue injury at the site of infection, and an excessive inflammatory response may lead to hypotension and organ dysfunction (p. 196). The contribution of the immune response to disease manifestations is exemplified by the immune reconstitution inflammatory syndrome (IRIS). This is seen, for example, in human immunodeficiency virus (HIV) infection, post-transplantation neutropenia or tuberculosis (which causes suppression of T-cell function): there is a paradoxical worsening of the clinical condition as the immune dysfunction is corrected, caused by an exuberant but dysregulated

inflammatory response. The febrile response Thermoregulation is altered in infectious disease, which may cause both hyperthermia (fever) and hypothermia. Fever is mediated mainly by ‘pyrogenic cytokines’ (e.g. interleukins IL-1 and IL-6, and tumour necrosis factor alpha (TNF-α)), which are released in response to various immunological stimuli including activation of pattern recognition receptors (PRRs) by microbial pyrogens (e.g. lipopolysaccharide) and factors released by injured cells. Their ultimate effect is to induce the synthesis of prostaglandin E2, which binds to specific receptors in the preoptic nucleus of the hypothalamus (thermoregulatory centre), causing the core temperature to rise. Rigors are a clinical symptom (or sign if they are witnessed) characterised by feeling very cold (‘chills’) and uncontrollable shivering, usually followed by fever and sweating. Rigors occur when the thermoregulatory centre attempts to correct a core temperature to a higher level by stimulating skeletal muscle activity and shaking. There are data to support the hypothesis that raised body temperature interferes with the replication and/or virulence of pathogens. The mechanisms and possible protective role of infection-driven hypothermia, however, are poorly understood, and require further study. Host–pathogen interactions ‘Pathogenicity’ is the capability of an organism to cause disease and ‘virulence’ is the extent to which a pathogen is able to cause disease. Pathogens produce proteins and other factors, termed virulence factors, which contribute to disease. • Primary pathogens cause disease in a proportion of individuals to whom they are exposed, regardless of the host’s immunological status. • Opportunistic pathogens cause disease only in individuals whose host defences are compromised, e.g. by an intravascular catheter, or when the immune system is compromised, by genetic susceptibility or immunosuppressive therapy. Characteristics of successful pathogens Successful pathogens have a number of attributes. They compete with host cells and colonising flora by various methods, including sequestration of nutrients and production of bacteriocins. Motility enables pathogens to reach their site of infection, often sterile sites that colonising bacteria do not reach, such as the distal airway. Many microorganisms, including viruses, use ‘adhesins’ to attach to host cells initially. Some pathogens can invade through tissues. Many bacterial and fungal infections form ‘biofilms’. After initial adhesion to a host surface, bacteria multiply in biofilms to form complex three-dimensional structures surrounded by a matrix of host and bacterial products that afford protection to the colony and limit the effectiveness of antimicrobials. Biofilms forming on man-made medical devices such as vascular catheters or grafts can be particularly difficult to treat. Pathogens may produce toxins, microbial molecules that cause adverse effects on host cells, either at the site of infection, or remotely following carriage through the blood stream. Endotoxin is the lipid component of Gram-negative bacterial outer membrane lipopolysaccharide. It is released when bacterial cells are damaged and has generalised inflammatory effects. Exotoxins are proteins released by living bacteria, which often have specific effects on target organs (Box 6.3). Intracellular pathogens, including viruses, bacteria (e.g. Salmonella spp., Listeria monocytogenes and Mycobacterium tuberculosis), parasites (e.g. Leishmania spp.) and fungi (e.g. 6.3 Exotoxin-mediated bacterial diseases Disease Organism Antibiotic-associated diarrhoea/ pseudomembranous colitis Clostridium difficile (p. 230) Botulism Clostridium botulinum (p. 1126) Cholera Vibrio cholerae (p. 264) Diphtheria Corynebacterium diphtheriae (p. 265) Haemolytic uraemic syndrome Enterohaemorrhagic Escherichia coli (E. coli O157 and other strains) (p. 263) Necrotising pneumonia Staphylococcus aureus (p. 250) Tetanus Clostridium tetani (p. 1125) Toxic shock syndrome Staphylococcus aureus (p. 252) Streptococcus pyogenes (p. 253)

organism and its background. Examples include Gram staining of bacteria and Ziehl–Neelsen or auramine staining of acid- and alcohol-fast bacilli (AAFB) in tuberculosis (the latter requires an ultraviolet light source). In histopathological examination of tissue samples, multiple stains are used to demonstrate not only the presence of microorganisms but also features of disease pathology. • Dark field microscopy (in which light is scattered to make organisms appear bright on a dark background) is used, for example, to examine genital chancre fluid in suspected syphilis. • Electron microscopy may be used to examine stool and vesicle fluid to detect enteric and herpesviruses, respectively, but its use has largely been supplanted by nucleic acid detection (see below). • Flow cytometry can be used to analyse liquid samples (e.g. urine) for the presence of particles based on properties such as size, impedance and light scatter. This technique can detect bacteria but may misidentify other particles as bacteria too. Investigation of infection The aims of investigating a patient with suspected infection are to confirm the presence of infection, identify the specific pathogen(s) and identify its susceptibility to specific antimicrobial agents in order to optimise therapy. The presence of infection may be suggested by identifying proteins that are produced in response to pathogens as part of the innate immune and acute phase responses (p. 70). Pathogens may be detected directly (e.g. by culturing a normally sterile body site) or their presence may be inferred by identifying the host response to the organism, (‘indirect detection’, Box 6.4). Careful sampling increases the likelihood of diagnosis (Box 6.5). Culture results must be interpreted in the context of the normal flora at the sampled site (see Fig. 6.5). The extent to which a microbiological test result supports or excludes a particular diagnosis depends on its statistical performance (e.g. sensitivity, specificity, positive and negative predictive value, p. 4). Sensitivity and specificity vary according to the time between infection and testing, and positive and negative predictive values depend on the prevalence of the condition in the test population. The complexity of test interpretation is illustrated in Figure 6.8 below, which shows the ‘windows of opportunity’ afforded by various testing methods. Given this complexity, effective communication between the clinician and the microbiologist is vital to ensure accurate test interpretation. Direct detection of pathogens Some direct detection methods provide rapid results and enable detection of organisms that cannot be grown easily on artificial culture media, such as Chlamydia spp.; they can also provide information on antimicrobial sensitivity, e.g. Mycobacterium tuberculosis. Detection of whole organisms Whole organisms are detected by examination of biological fluids or tissue using a microscope. • Bright field microscopy (in which the test sample is interposed between the light source and the objective lens) uses stains to enhance visual contrast between the 6.5 How to provide samples for microbiological sampling Communicate with the laboratory • Discuss samples that require processing urgently or that may contain hazardous or unusual pathogens with laboratory staff before collection • Communication is key to optimising microbiological diagnosis. If there is doubt about any aspect of sampling, it is far better to discuss it with laboratory staff beforehand than to risk diagnostic delay by inappropriate sampling or sample handling Take samples based on a clinical diagnosis • Sampling in the absence of clinical evidence of infection is rarely appropriate (e.g. collecting urine, or sputum for culture) Use the correct container • Certain tests (e.g. nucleic acid and antigen detection tests) require proprietary sample collection equipment Follow sample collection procedures • Failure to follow sample collection instructions precisely can result in false-positive (e.g. contamination of blood culture samples) or false-negative (e.g. collection of insufficient blood for culture) results Label sample and request form correctly • Label sample containers and request forms according to local policies, with demographic identifiers, specimen type and time/date collected • Include clinical details on request forms • Identify samples carrying a high risk of infection (e.g. blood liable to contain a blood-borne virus)

with a hazard label Use appropriate packaging • Close sample containers tightly and package securely (usually in sealed plastic bags) • Attach request forms to samples but not in the same compartment (to avoid contamination, should leakage occur) Manage storage and transport • Transport samples to the microbiology laboratory quickly • If pre-transport storage is required, conditions (e.g. refrigeration, incubation, storage at room temperature) vary with sample type • Notify the receiving laboratory prior to arrival of unusual or urgent samples, to ensure timely processing 6.4 Tests used to diagnose infection Non-specific markers of inflammation/infection • e.g. White cell count in blood sample (WCC), plasma C-reactive protein (CRP), procalcitonin, serum lactate, cell counts in urine or cerebrospinal fluid (CSF), CSF protein and glucose Direct detection of organisms or organism components • Microscopy • Detection of organism components (e.g. antigen, toxin) • Nucleic acid amplification (e.g. polymerase chain reaction) Culture of organisms • ± Antimicrobial susceptibility testing Tests of the host's specific immune response • Antibody detection • Interferon-gamma release assays (IGRA)

106 • PRINCIPLES OF INFECTIOUS DISEASE even in rapid-culture systems. Certain organisms, such as Mycobacterium leprae and Tropheryma whipplei, cannot be cultivated on artificial media, and others (e.g. Chlamydia spp. and viruses) grow only in culture systems, which are slow and labour-intensive. Blood culture The terms 'bacteraemia' and 'fungaemia' describe the presence of bacteria and fungi in the blood. 'Blood-stream infection' (p. 225) is the association of bacteraemia/fungaemia with clinical evidence of infection. The presence of bacteraemia/fungaemia can be determined by inoculating a liquid culture medium with freshly drawn blood, which is then incubated in a system that monitors it constantly for growth of microorganisms (e.g. by detecting products of microbial respiration using fluorescence; Fig. 6.6). If growth is detected, organisms are identified and sensitivity testing is performed. Traditionally, identification has been achieved by Gram stain appearance and biochemical reactions. However, matrix-assisted laser desorption/ionisation time-of-flight mass spectroscopy (MALDI-TOF-MS; see Box 6.2) is being used increasingly to identify organisms. MALDI-TOF-MS produces a profile of proteins of different sizes from the target microorganism and uses databases of such profiles to identify the organism (Fig. 6.7). It is rapid and accurate. Taking multiple blood samples for culture at different times allows differentiation of transient (one or two positive samples) and persistent (majority are positive) bacteraemia. This can be clinically important in the identification of the source of infection (p. 530). Indirect detection of pathogens Tests may be used to detect the host's immune (antibody) response to a specific microorganism, and can enable the diagnosis of infection with organisms that are difficult to detect by other methods or are no longer present in the host. The term 'serology' describes tests carried out on serum and includes both antigen (direct) and antibody (indirect) detection. Antibody detection Organism-specific antibody detection is applied mainly to blood (Fig. 6.8). Results are typically expressed as titres: that is, the reciprocal of the highest dilution of the serum at which antibody is detectable (for example, detection at serum dilution of 1 : 64 gives a titre of 64). 'Seroconversion' is defined as either a change from negative to positive detection or a fourfold rise in titre between acute and convalescent serum samples. An acute sample is usually taken during the first week of disease and the convalescent sample 2–4 weeks later. Earlier diagnosis can be achieved by detection of immunoglobulin M (IgM) antibodies, which are produced early in infection (p. 68). A limitation of these tests is that antibody production requires a fully functional host immune system, so there may be false-negative results in immunocompromised patients. Also, other than in chronic infections and with IgM detection, antibody tests usually provide a retrospective diagnosis. Antibody detection methods are described

below (antigen detection methods are also described here as they share similar methodology). Enzyme-linked immunosorbent assay The principles of the enzyme-linked immunosorbent assay (ELISA, EIA) are illustrated in Figure 6.9. These assays rely on linking Detection of components of organisms Components of microorganisms detected for diagnostic purposes include nucleic acids, cell wall molecules, toxins and other antigens. Commonly used examples include Legionella pneumophila serogroup 1 antigen in urine and cryptococcal polysaccharide antigen in cerebrospinal fluid (CSF). Most antigen detection methods are based on in vitro binding of specific antigen/antibody and are described below. Other methods may be used, such as tissue culture cytotoxicity assay for C. difficile toxin. In toxin-mediated disease, detection of toxin may be of greater relevance than identification of the organism itself (e.g. stool C. difficile toxin). Nucleic acid amplification tests In a nucleic acid amplification test (NAAT), specific sequences of microbial DNA and RNA are identified using a nucleic acid primer that is amplified exponentially by enzymes to generate multiple copies of a target nucleotide sequence. The most commonly used amplification method is the polymerase chain reaction (PCR; see Fig. 3.11, p. 53). Reverse transcription (RT) PCR is used to detect RNA from RNA viruses (e.g. hepatitis C virus and HIV-1). The use of fluorescent labels in the reaction enables ‘real-time’ detection of amplified DNA; quantification is based on the principle that the time taken to reach the detection threshold is proportional to the initial number of copies of the target nucleic acid sequence. In multiplex PCR, multiple primer pairs are used to enable detection of several different organisms at once. Determination of nucleotide sequences in a target gene(s) can be used to assign microorganisms to specific strains, which may be relevant to treatment and/or prognosis (e.g. in hepatitis C infection, p. 877). Genes that are relevant to pathogenicity (such as toxin genes) or antimicrobial resistance can also be detected; for example, the mecA gene is used to screen for MRSA. NAATs are the most sensitive direct detection methods and are also relatively rapid. They are used widely in virology, where the possibility of false-positive results from colonising or contaminating organisms is remote, and are applied to blood, respiratory samples, stool and urine. In bacteriology, PCR is used to examine CSF, blood, tissue and genital samples, and multiplex PCR is being developed for use in faeces. PCR is particularly helpful for microorganisms that cannot be readily cultured, e.g. Tropheryma whipplei, and is being used increasingly in mycology and parasitology. Culture Microorganisms may be both detected and further characterised by culture from clinical samples (e.g. tissue, swabs and body fluids). • Ex vivo culture (tissue or cell culture) was widely used in the isolation of viruses but has been largely supplanted by NAAT. • In vitro culture (in artificial culture media) of bacteria and fungi is used to confirm the presence of pathogens, allow identification, test antimicrobial susceptibility and subtype the organism for epidemiological purposes. Culture has its limitations: results are not immediate, even for organisms that are easy to grow, and negative cultures rarely exclude infection. Organisms such as Mycobacterium tuberculosis are slow-growing, typically taking at least 2 weeks,

Immunofluorescence assays Indirect immunofluorescence assays (IFAs) detect antibodies by incubating a serum sample with immobilised antigen (e.g. cells known to be infected with virus on a glass slide); any virus-specific antibody present in the serum binds to antigen and is then detected using a fluorescent-labelled anti-human immunoglobulin (‘secondary’ antibody). Fluorescence is visualised using a microscope. This method can also detect organisms in clinical samples (usually tissue or centrifuged cells) using a specific antibody in place of immobilised

antigen to achieve capture. Complement fixation test In a complement fixation test (CFT), patient serum is heat-treated to inactivate complement and mixed with the test antigen. Any specific antibody in the serum will complex with the antigen. Complement is then added to the reaction. If antigen–antibody an antibody with an enzyme that generates a colour change on exposure to a chromogenic substrate. Various configurations allow detection of antigens or specific subclasses of immunoglobulin (e.g. IgG, IgM, IgA). ELISA may also be adapted to detect PCR products, using immobilised oligonucleotide hybridisation probes and various detection systems. Immunoblot (Western blot) Microbial proteins are separated according to molecular weight by polyacrylamide gel electrophoresis (PAGE) and transferred (blotted) on to a nitrocellulose membrane, which is incubated with patient serum. Binding of specific antibody is detected with an enzyme–anti-immunoglobulin conjugate similar to that used in ELISA, and specificity is confirmed by its location on the membrane. Immunoblotting is a highly specific test, which may be used to confirm the results of less specific tests such as ELISA (e.g. in Lyme disease, p. 255). Fig. 6.6 An overview of the processing of blood cultures. *In laboratories equipped with MALDI-TOF-MS (p. 106), rapid definitive organism identification may be achieved at stage 6 and/or stage 8. Department of Microbiology 1 Patient sampling Contamination minimised by aseptic technique. Maximise sensitivity by sampling correct volume 2 Sample handling Follow local instructions for safety, labelling, and numbers of samples and bottles required 3 Specimen transport Transport samples to laboratory as quickly as possible. Follow manufacturer's instructions for the blood culture system used if temporary storage is required 4 Incubation Incubate at 35–37°C for 5–7 days. Microbial growth is usually detected by constant automatic monitoring of CO2. If no growth, specimen is negative and discarded 5 Growth detection Time to positivity (TTP) is usually 12–24 hrs in significant bacteraemia, but may be shorter in overwhelming sepsis or longer with fastidious organisms (e.g. Brucella spp.) 6 Preliminary results* A Gram film of the blood culture medium is examined and results are communicated immediately to the clinician to guide antibiotic therapy 7 Incubation A small amount of the medium is incubated on a range of culture media. Preliminary susceptibility testing may be carried out 8 Culture results* Preliminary susceptibility results are communicated to the clinician 9 Definitive results Further overnight incubation is often required for definitive identification of organisms (by biochemical testing) and additional susceptibility testing; identification by MALDI-TOF MS (Fig. 6.7) is more rapid Overnight incubation required Urgent communication required 10 Reporting A final summary is released when all testing is complete. For clinical care, communication of interim results (Gram film, preliminary identification and susceptibility) is usually more important than the final report. Effective clinical–laboratory communication is vital

 For example, in the Weil–Felix test, if a patient's serum contains antibodies to rickettsial species they cause agglutination when Proteus spp. surface (O) antigens are added because the antibodies cross-react with the Proteus antigens. The test lacks sensitivity and specificity but is still used to diagnose rickettsial infection in resource-limited settings. The Widal test reaction uses a suspension of Salmonella typhi and S. paratyphi 'A' and 'B', treated to retain only 'O' and 'H' antigens. These antigens are kept to detect corresponding antibodies in serum from a patient suspected of having typhoid fever. The test is not specific but is still used in some parts of the world. • In indirect (passive) agglutination, specific antigen is attached to the surface of carrier particles, which agglutinate when incubated with patient samples that contain specific antibodies. • In reverse passive agglutination (an antigen detection test), the carrier particle is coated with antibody rather than antigen. Other tests Immunodiffusion involves

antibodies and antigen migrating through gels, with or without the assistance of electrophoresis, and forming insoluble complexes where they meet. The complexes are seen on staining as ‘precipitin bands’. Immunodiffusion is used in the diagnosis of dimorphic fungi (p. 300) and some forms of aspergillosis (p. 596). Immunochromatography is used to detect antigen. The system consists of a porous test strip (e.g. a nitrocellulose membrane), at one end of which there is target-specific antibody, complexed with coloured microparticles. Further specific antibody is immobilised in a transverse narrow line some distance along the strip. Test material (e.g. blood or urine) is added to the antibody–particle complexes, which then migrate along the strip by capillary action. If these are complexed with antigen, they will be immobilised by the specific antibody and visualised as a transverse line across the strip. If the test is negative, the antibody–particle complexes will bind to a line of immobilised anti-immunoglobulin antibody placed further along the strip, which acts as a negative control. Immunochromatographic tests are rapid and relatively cheap to perform, and are appropriate for point-of-care testing, e.g. in HIV 1 and malaria (p. 276). complexes are present, the complement will be ‘fixed’ (consumed). Sheep erythrocytes, coated with an anti-erythrocyte antibody, are added. The degree of erythrocyte lysis reflects the remaining complement and is inversely proportional to the quantity of the specific antigen–antibody complex present. Agglutination tests When antigens are present on the surface of particles (e.g. cells, latex particles or microorganisms) and cross-linked with antibodies, visible clumping (or ‘agglutination’) occurs. • In direct agglutination, patient serum is added to a suspension of organisms that express the test antigen. Fig. 6.7 The workings of matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS). Adapted from Sobin K, Hameer D, Ruparel T. Digital genotyping using molecular affinity and mass spectrometry. Nature Rev Genet 2003; 4:1001–1008. Lighter m/z Mass spectrum Detector Flight tube Laser Sample plate Voltage grid Intensity Heavier Separation region (electric field-free) Fig. 6.8 Detection of antigen, nucleic acid and antibody in infectious disease. The acute sample is usually taken during the first week of illness, and the convalescent sample 2–4 weeks later. Detection limits and duration of detectability vary between tests and diseases, although in most diseases immunoglobulin M (IgM) is detectable within the first 1–2 weeks. Acute sample Nucleic acid (NA) detection Antigen (Ag) detection IgM NA Ag IgG Limit of detection Antibody detection: IgM Antibody detection: IgG (seroconversion) Antibody detection: IgG (fourfold rise in titre) Convalescent sample Windows of diagnostic opportunity Serum levels

Antibody-independent specific immunological tests The interferon-gamma release assay (IGRA) is being used increasingly to diagnose latent tuberculosis infection (LTBI). The principle behind IGRA is discussed on page 594. IGRA cannot distinguish between latent and active tuberculosis infection and is therefore appropriate for use only in countries where the background incidence of tuberculosis is low. Antimicrobial susceptibility testing If growth of microorganisms in culture is inhibited by the addition of an antimicrobial agent, the organism is considered to be susceptible to that antimicrobial. Bacteriostatic agents cause reversible inhibition of growth and bactericidal agents cause cell death; the terms fungistatic/fungicidal are equivalent for antifungal agents, and virustatic/virucidal for antiviral agents. The lowest concentration of antimicrobial agent at which growth is inhibited is the minimum inhibitory concentration (MIC), and the lowest concentration that causes cell death is the minimum bactericidal concentration (MBC). If the MIC is less than or equal to a predetermined breakpoint threshold, the organism is considered susceptible, and if the MIC is greater than the breakpoint, it is resistant. Breakpoints are determined for each

antimicrobial agent from a combination of pharmacokinetic (p. 17) and clinical data. The relationship between in vitro antimicrobial susceptibility and clinical response is complex, as response also depends on immune status, pharmacokinetic variability (p. 17), comorbidities that may influence pharmacokinetics or pharmacodynamics, and antibiotic dosing, as well as MIC/MBC. Thus, although treating a patient according to the results of susceptibility testing increases the likelihood of recovery, it does not guarantee therapeutic success. Susceptibility testing is often carried out by disc diffusion (Fig. 6.10). Antibiotic-impregnated filter paper discs are placed on agar plates containing bacteria; antibiotic diffuses into the agar, resulting in a concentration gradient centred on the disc. Bacteria are unable to grow where the antibiotic concentration exceeds the MIC, which may therefore be inferred from the size of the zone of inhibition. The MIC is commonly measured in diagnostic laboratories using ‘diffusion strips’. Fig. 6.9 Antibody (Ab) and antigen (Ag) detection by enzyme-linked immunosorbent assay (ELISA). This can be configured in various ways. A Patient Ab binds to immobilised specific Ag and is detected by addition of anti-immunoglobulin–enzyme conjugate and chromogenic substrate. B Patient Ab binds to immobilised Ig subclass-specific Ab and is detected by addition of specific Ag, followed by antibody–enzyme conjugate and chromogenic substrate. C Patient Ab and antibody–enzyme conjugate bind to immobilised specific Ag. Magnitude of colour change reaction is inversely proportional to concentration of patient Ab. D Patient Ag binds to immobilised Ab and is detected by addition of antibody–enzyme conjugate and chromogenic substrate. In A, the conjugate Ab is specific for human immunoglobulin. In B–D, it is specific for Ag from the disease-causing organism. Antibody capture ELISA Patient Ab Ig subclass-specific Ab Ab specific to Ag from the disease-causing organism Specific Ag Chromogenic substrate Antibody–enzyme conjugate Competitive antibody detection ELISA Double antibody sandwich ELISA (for antigen detection) Antibody detection ELISA A B C D Fig. 6.10 Antimicrobial susceptibility testing by disc diffusion (panels 1-4) and minimum inhibitory concentration (MIC, panel 5).

1. The test organism is spread over the surface of an agar plate.
2. Antimicrobial-impregnated discs (A–F) are placed on the surface and the plate is incubated (e.g. overnight). 3–4. After incubation, zones of growth inhibition may be seen. The organism is considered susceptible if the diameter of the zone of inhibition exceeds a pre-determined threshold.
3. In a ‘diffusion strip’ test, the strip is impregnated with antimicrobial at a concentration gradient that decreases steadily from top to bottom. The system is designed so that the MIC value is the point at which the ellipse cuts a scale on the strip (arrow). 4, Kindly supplied by Charlotte Symes. Zone of inhibition Zone of inhibition

A B C D E F A B C D E F

 known and thought to have been controlled or eradicated, it is considered to be re-emerging. Many emerging diseases are caused by organisms that infect animals and have undergone adaptations that enable them to infect humans. This is exemplified by HIV-1, which is believed to have originated in higher primates in Africa. The geographical pattern of some recent emerging and re-emerging infections is shown in Figure 6.11. Reservoirs of infection The US Centers for Disease Control (CDC) define a reservoir of infection as any person, other living organism, environment or combination of these in which the infectious agent lives and replicates and on which the infectious agent is dependent for its survival. The

infectious agent is transmitted from this reservoir to a susceptible host. Human reservoirs Both colonised individuals and those with infection can act as reservoirs, e.g. with Staph. aureus (including MRSA), Strep. pyogenes and C. difficile. For infected humans to act as reservoirs, the infections caused must be long-lasting in at least a proportion of those affected, to enable onward transmission (e.g. tuberculosis, sexually transmitted infections). Humans are the only reservoir for some infections (e.g. measles). Animal reservoirs The World Health Organization (WHO) defines a zoonosis as ‘a disease or infection that is naturally transmissible from vertebrate animals to humans’. Infected animals may be asymptomatic. Zoonotic agents may be transmitted via any of the routes described below. Primary infection with zoonoses may be transmitted onward between humans, causing secondary disease (e.g. Q fever, brucellosis, Ebola). Environmental reservoirs Many infective pathogens are acquired from an environmental source. However, some of these are maintained in human or animal reservoirs, with the environment acting only as a conduit for infection. Epidemiology of infection The communicability of infectious disease means that, once a clinician has diagnosed an infectious disease, potential exposure of other patients must also be considered. The patient may require separation from other patients (‘isolation’), or an outbreak of disease may need to be investigated in the community (Ch. 5). The approach will be specific to the microorganism involved (Chs 11–13) but the principles are outlined below. Geographical and temporal patterns of infection Endemic disease Endemic disease has a constant presence within a given geographical area or population. The infectious agent may have a reservoir, vector or intermediate host that is geographically restricted, or may itself have restrictive environmental requirements (e.g. temperature range, humidity). The population affected may be geographically isolated or the disease may be limited to unvaccinated populations. Factors that alter geographical restriction include: • expansion of an animal reservoir (e.g. Lyme disease from reforestation) • vector escape (e.g. airport malaria) • extension of host range (e.g. schistosomiasis from dam construction) • human migration (e.g. carbapenemase-producing Klebsiella pneumoniae) • public health service breakdown (e.g. diphtheria in unvaccinated areas) • climate change (e.g. dengue virus and Rift Valley fever). Emerging and re-emerging disease An emerging infectious disease is one that has newly appeared in a population, or has been known for some time but is increasing in incidence or geographical range. If the disease was previously Fig. 6.11 Geographical locations of some infectious disease outbreaks, with examples of emerging and re-emerging diseases. (CPE = carbapenemase-producing Enterobacteriaceae; MDR-TB = multidrug-resistant tuberculosis; MERS-Co-V = Middle East respiratory syndrome coronavirus; XDR-TB = extensively drug-resistant tuberculosis) CPE Ebola virus disease Cholera Cholera Cholera Cryptococcus gattii Cryptococcus gattii Zika virus Zika virus Cyclospora Chikungunya virus Chikungunya virus XDR-TB MERS-Co-V Anthrax MDR-TB

and humans in MRSA. Fomites are inanimate objects such as door handles, water taps and ultrasound probes, which are particularly associated with health care-associated infection (HCAI). The likelihood of infection following transmission of a pathogen depends on organism factors (virulence, p. 104) and host susceptibility. The incubation period is the time between exposure and development of symptoms, and the period of infectivity is the period after exposure during which the patient is infectious to others. Knowledge of incubation periods and of periods of infectivity is important in controlling the spread of disease, although for many diseases these estimates are imprecise (Boxes 6.6 and 6.7). Deliberate release Deliberate release of pathogens with the

intention of causing disease is known as biological warfare or bioterrorism, depending on the scale and context. Deliberate release incidents have included a 750-person outbreak of Salmonella typhimurium caused by contamination of salads in 1984 (Oregon, USA) and 22 cases of anthrax (five fatal) from the mailing of finely powdered (weaponised) anthrax spores in 2001 (New Jersey, USA). Diseases with high potential for deliberate release include anthrax, plague, tularaemia, smallpox and botulism (through toxin release). Infection prevention and control Infection prevention and control (IPC) describes the measures applied to populations with the aim of breaking the chain of infection (see Fig. 6.1, p. 100). Health care-associated infection The risk of developing infection following admission to a health-care facility (health care-associated infection, HCAI) in Transmission of infection Communicable diseases may be transmitted by one or more of the following routes: • Respiratory route: inhalation. • Faecal–oral route: ingestion of material originating from faeces. • Sexually transmitted infections: direct contact between mucous membranes. • Blood-borne infections: direct inoculation of blood or body fluids. • Direct contact: very few organisms are capable of causing infection by direct contact with intact skin. Most infection by this route requires contact with damaged skin (e.g. surgical wound). • Via a vector or fomite: the vector/fomite bridges the gap between the infected host or reservoir and the uninfected host. Vectors are animate, and include mosquitoes in malaria, dengue and Zika virus infection, fleas in plague 6.7 Periods of infectivity in common childhood infectious diseases Disease Infectious period Chickenpox1 From 4 days before until 5 days after appearance of the rash (transmission before 48 hrs prior to the onset of rash is rare)4 Measles2 From 4 days before onset to 4 days after onset of the rash Mumps3 From 2–3 days before to 5 days after disease onset5 Rubella3 From 10 days before until 15 days after the onset of the rash, but most infectious during prodromal illness4 Scarlet fever1 Unknown6 Whooping cough1 Unknown6,7 1From Richardson M, Elliman D, Maguire H, et al. Pediatr Infect Dis J 2001; 20:380–388. 2Centers for Disease Control, USA; cdc.gov/measles/hcp/. 3Bennett JE, Dolin R, Blaser MJ. Mandell, Douglas and Bennett's Principles and practice of infectious diseases, 8th edn. Philadelphia: Elsevier; 2015. 4–6Exclude from contact with non-immune and immunocompromised people for 5 days from 4onset of rash, 5onset of parotitis, or 6start of antibiotic treatment. 7Exclude for 3 weeks if untreated. Durations are approximate and vary between information sources, and these recommendations may differ from local or national guidance. 6.6 Incubation periods of important infections1 Infection Incubation period Short incubation periods Anthrax, cutaneous3 9 hrs to 2 weeks Anthrax, inhalational3 2 days2 Bacillary dysentery5 1–6 days Cholera3 2 hrs to 5 days Dengue haemorrhagic fever6 3–14 days Diphtheria6 1–10 days Gonorrhoea4 2–10 days Influenza5 1–3 days Meningococcaemia3 2–10 days Norovirus1 1–3 days SARS coronavirus3 2–7 days2 Scarlet fever5 2–4 days Intermediate incubation periods Amoebiasis6 1–4 weeks Brucellosis4 5–30 days Chickenpox5 11–20 days Lassa fever3 3–21 days Malaria3 10–15 days Measles5 6–19 days Mumps5 15–24 days Poliomyelitis6 3–35 days Psittacosis4 1–4 weeks Rubella5 15–20 days Typhoid5 5–31 days Whooping cough5 5–21 days Long incubation periods Hepatitis A5 3–7 weeks Hepatitis B4 6 weeks to 6 months Leishmaniasis, cutaneous6 Weeks to months Leishmaniasis, visceral6 Months to years Leprosy (Hansen's disease)3 5–20 years Rabies4 2–8 weeks2 Trypanosoma brucei gambiense infection6 Months to years Tuberculosis5 1–12 months 1Incubation periods are approximate and may differ from local or national guidance. 2Longer incubation periods have been reported. 3WHO. 4Health Protection Agency (now Health Protection England). 5Richardson M, Elliman D, Maguire H, et al. Pediatr Infect Dis J 2001; 20:380–388. 6Centers for Disease Control, USA. (SARS = severe acute respiratory syndrome)

112 • PRINCIPLES OF INFECTIOUS DISEASE Fig. 6.12 Commonly encountered health care-associated infections (HCAIs) and the factors that predispose to them. Temporary central venous catheter infection Staphylococcus aureus (incl. MRSA) Coagulase-negative staphylococci Coliforms Candida Prosthetic joint infection Coagulase-negative staphylococci Staphylococcus aureus Streptococci Coliforms Propionibacterium acnes Surgical site infection Staphylococcus aureus Beta-haemolytic streptococci Coliforms Anaerobes Cuffed/tunnelled central venous catheter infection Coagulase-negative staphylococci Staphylococcus aureus (incl. MRSA) Coliforms Candida Pseudomonas spp. Enterococcus spp. External ventricular drain and ventriculoperitoneal shunt infection Coagulase-negative staphylococci Staphylococcus aureus Diphtheroids Pseudomonas aeruginosa Peritoneal dialysis-related peritonitis Staphylococcus aureus Coagulase-negative staphylococci Coliforms Pseudomonas spp. Breast implant infection Staphylococcus aureus Coagulase-negative staphylococci the developed world is about 10%. Many nosocomial bacterial infections are caused by organisms that are resistant to numerous antibiotics (multi-resistant bacteria), including MRSA, extendedspectrum β-lactamases (ESBLs) and carbapenemase-producing Enterobacteriaceae (CPE), and glycopeptide-resistant enterococci (GRE). Other infections of particular concern in hospitals include C. difficile (p. 264) and norovirus (p. 249). Some examples are shown in Figure 6.12. IPC measures are described in Box 6.8. The most important is maintenance of good hand hygiene (Fig. 6.13). Hand (CPE = carbapenemase-producing Enterobacteriaceae; GRE = glycopeptide-resistant enterococci; MRSA = meticillin-resistant Staphylococcus aureus) Institutions • Handling, storage and disposal of clinical waste • Containment and safe removal of spilled blood and body fluids • Cleanliness of environment and medical equipment • Specialised ventilation (e.g. laminar flow, air filtration, controlled pressure gradients) • Sterilisation and disinfection of instruments and equipment • Food hygiene • Laundry management Health-care staff • Education • Hand hygiene, including hand-washing (see Fig. 6.13) • Sharps management and disposal • Use of personal protective equipment (masks, sterile and non-sterile gloves, gowns and aprons) • Screening health workers for disease (e.g. tuberculosis, hepatitis B virus, MRSA) • Immunisation and post-exposure prophylaxis Clinical practice • Antibiotic stewardship (p. 115) • Aseptic technique • Perioperative antimicrobial prophylaxis • Screening patients for colonisation or infection (e.g. MRSA, GRE, CPE) Response to infections • Surveillance to detect alert organism (see text) outbreaks and antimicrobial resistance • Antibiotic chemoprophylaxis in infectious disease contacts, if indicated (see Box 6.18) • Isolation (see Box 6.9) • Reservoir control • Vector control 6.8 Measures used in infection prevention and control (IPC)

Fig. 6.13 Hand-washing. Good hand hygiene, whether with soap/water or alcohol handrub, includes areas that are often missed, such as fingertips, web spaces, palmar creases and the backs of hands. Adapted from the ‘How to Handwash’ URL: who.int/gpsc/5may/How_To_Handwash_Poster.pdf Wash hands only when visibly soiled! Otherwise use handrub! Duration of the entire procedure: 40–60 sec.

Wet hands with water using elbow-operated or nontouch taps (if available) Apply enough soap to cover all hand surfaces Rub hands palm to palm

Right palm over left dorsum with interlaced fingers and vice versa Palm to palm with fingers interlaced Backs of fingers to opposing palms with fingers interlaced

Rotational rubbing of left thumb clasped in right palm and vice versa Rotational rubbing, backwards and forwards with clasped fingers of right hand in left palm and vice versa Rinse hands with water

Dry thoroughly with a single-use towel If hand-operated taps have been used, use towel to turn off tap ...and your hands are clean decontamination (e.g. using alcohol gel or washing) is mandatory before and after every patient contact. Decontamination with alcohol gel is usually adequate but hand-washing (with hot water, liquid soap and complete drying) is required after any procedure that involves more than casual physical contact, or if hands are visibly soiled. In situations where the prevalence of C. difficile is high (e.g. a local outbreak), alcohol gel decontamination between patient contacts is inadequate as it does not kill C. difficile spores, and hands must be washed. Some infections necessitate additional measures to prevent cross-infection (Box 6.9). To minimise risk of infection, invasive procedures must be performed using strict aseptic technique. Airborne transmission Contact transmission Droplet transmission Precautions Negative pressure room with air exhausted externally or filtered N95 masks or personal respiratiors for staff; avoid using non-immune staff Private room preferred (otherwise, inter-patient spacing ≥ 1 m) Gloves and gown for staff in contact with patient or contaminated areas Private room preferred (otherwise, inter-patient spacing ≥ 1 m) Surgical masks for staff in close contact with patient Infections managed with these precautions Measles Tuberculosis, pulmonary or laryngeal, confirmed or suspected Enteroviral infections in young children (diapered or incontinent) Norovirus2 C. difficile infection Multidrug-resistant organisms (e.g. MRSA, ESBL, GRE, VRSA, penicillin-resistant Strep. pneumoniae)3 Parainfluenza in infants and young children Rotavirus RSV in infants, children and immunocompromised Viral conjunctivitis, acute Diphtheria, pharyngeal Haemophilus influenzae type B infection Herpes simplex virus, disseminated or severe Influenza Meningococcal infection Mumps Mycoplasma pneumoniae Parvovirus (erythrovirus) B19 (erythema infectiosum, fifth disease) Pertussis Plague, pneumonic/bubonic Rubella Strep. pyogenes (group A), pharyngeal Infections managed with multiple precautions Smallpox, monkeypox, VZV (chickenpox or disseminated disease)4 Adenovirus pneumonia SARS, viral haemorrhagic fever2 6.9 Types of isolation precaution1 1Recommendations based on 2007 CDC guideline for isolation precautions. May differ from local or national recommendations. 2Not a CDC recommendation. 3Subject to local risk assessment. 4Or in any immunocompromised patient until possibility of disseminated infection excluded. (ESBL = extended-spectrum β-lactamase; GRE = glycopeptide-resistant enterococci; MRSA = meticillin-resistant Staph. aureus; RSV = respiratory syncytial virus; SARS = severe acute respiratory syndrome; VRSA = vancomycin-resistant Staph. aureus; VZV = varicella zoster virus)

 6.10 Terminology in outbreaks of infection Term Definition Classification of related cases of infectious disease* Cluster An aggregation of cases of a disease that are closely grouped in time and place, and may or may not exceed the expected number Epidemic The occurrence of more cases of disease than expected in a given area or among a specific group of people over a particular period of time Outbreak Synonymous with epidemic. Alternatively, a localised, as opposed to generalised, epidemic Pandemic An epidemic occurring over a very wide area (several countries or continents) and usually affecting a large proportion of the population Classification of affected patients (cases) Index case The first case identified in an

outbreak Primary cases Cases acquired from a specific source of infection Secondary cases Cases acquired from primary cases Types of outbreak Common source outbreak Exposure to a common source of infection (e.g. water-cooling tower, medical staff member shedding MRSA). New primary cases will arise until the source is no longer present Point source outbreak Exposure to a single source of infection at a specific point in time (e.g. contaminated food at a party). Primary cases will develop disease synchronously Person-toperson spread Outbreak with both primary and secondary cases. May complicate point source or common source outbreak *Adapted from cdc.gov. (MRSA = meticillin-resistant Staphylococcus aureus) 6.11 Reasons for including an infectious disease on a regional/national list of reportable diseases* Reason for inclusion Examples Endemic/local disease with the potential to spread and/or cause outbreaks Influenza, Salmonella, tuberculosis Imported disease with the propensity to spread and/or cause outbreaks Typhoid, cholera (depending on local epidemiology) Evidence of a possible breakdown in health protection/public health functions Legionella, Cryptosporidium Evidence of a possible breakdown in food safety practices Botulism, verotoxigenic E. coli Evidence of a possible failure of a vaccination programme Measles, poliomyelitis, pertussis Disease with the potential to be a novel or increasing threat to human health SARS, MERS-CoV, multi-resistant bacteria Evidence of expansion of the range of a reservoir/vector Lyme disease, rabies, West Nile encephalitis Evidence of possible deliberate release Anthrax, tularaemia, plague, smallpox, botulism *Given the different geographical ranges of individual diseases and wide national variations in public health services, vaccination programmes and availability of resources, reporting regulations vary between regions, states and countries. Many diseases are reportable for more than one reason. (MERS-CoV = Middle East respiratory syndrome coronavirus; SARS = severe acute respiratory syndrome) Outbreaks of infection Descriptive terms are defined in Box 6.10. Confirmation of an infectious disease outbreak usually requires evidence from ‘typing’ that the causal organisms have identical phenotypic and/or genotypic characteristics. If this is found not to be the case, the term pseudo-outbreak is used. When an outbreak of infection is suspected, a case definition is agreed. The number of cases that meet the case definition is then assessed by case-finding, using methods ranging from administration of questionnaires to national reporting systems. Case-finding usually includes microbiological testing, at least in the early stages of an outbreak. Temporal changes in cases are noted in order to plot an outbreak curve, and demographic details are collected to identify possible sources of infection. A case–control study, in which recent activities (potential exposures) of affected ‘cases’ are compared to those of unaffected ‘controls’, may be undertaken to establish the outbreak source, and measures are taken to manage the outbreak and control its spread. Good communication between relevant personnel during and after the outbreak is important to inform practice in future outbreaks. Surveillance ensures that disease outbreaks are either prevented or identified early. In hospitals, staff are made aware of the isolation of alert organisms, which have the propensity to cause outbreaks, and alert conditions, which are likely to be caused by such organisms. Analogous systems are used nationally; many countries publish lists of organisms and diseases, which, if detected (or suspected), must be reported to public health authorities (reportable or notifiable diseases). Reasons for a disease to be classified as reportable are shown in Box 6.11. Immunisation Immunisation may be passive or active. Passive immunisation is achieved by administering antibodies targeting a specific pathogen. Antibodies are obtained from blood, so confer some of the risks associated with blood products (p. 933). The protection afforded by passive immunisation is immediate but of short duration (a few weeks or months); it is used to prevent or attenuate infection before or after exposure (Box 6.12). Vaccination Active immunisation is achieved by vaccination with whole organisms or organism components (Box

6.13). Types of vaccine Whole-cell vaccines consist of live or inactivated (killed) microorganisms. Component vaccines contain only extracted or synthesised components of microorganisms (e.g. polysaccharides or proteins). Live vaccines contain organisms with attenuated (reduced) virulence, which cause only mild symptoms but induce T-lymphocyte and humoral responses (p. 68) and are therefore more immunogenic than inactivated whole-cell vaccines. The use of live vaccines in immunocompromised individuals is not generally recommended, but they may be used by specialists following a risk/benefit assessment. Component vaccines consisting only of polysaccharides, such as the pneumococcal polysaccharide vaccine (PPV), are poor activators of T lymphocytes and produce a short-lived antibody response without long-lasting memory. Conjugation of polysaccharide to a protein, as in the Haemophilus influenzae type B (Hib) vaccine and the protein conjugate pneumococcal

vaccinated to curtail further spread. Vaccination is aimed mainly at preventing infectious disease. However, vaccination against human papillomavirus (HPV) was introduced to prevent cervical and other cancers that complicate HPV infection. Vaccination guidelines for individuals are shown in Box 6.14. Vaccination becomes successful once the number of susceptible hosts in a population falls below the level required to sustain continued transmission of the target organism (herd immunity). Naturally acquired smallpox was declared to have been eradicated worldwide in 1980 through mass vaccination. In 1988, the WHO resolved to eradicate poliomyelitis by vaccination; the number of cases worldwide has since fallen from approximately 350 000 per annum to 74 in 2015. Recommended vaccination schedules vary between countries. In addition to standard vaccination schedules, catch-up schedules are specified for individuals who join vaccination programmes later than the recommended age. Antimicrobial stewardship Antimicrobial stewardship (AMS) refers to the systems and processes applied to a population to optimise the use of antimicrobial agents. The populations referred to here may be a nation, region, hospital, or a unit within a health-care organisation (e.g. ward or clinic). AMS aims to improve patient outcomes and reduce antimicrobial resistance (AMR). IPC and AMS complement each other (Fig. 6.14). Elements of AMS include treatment guidelines, antimicrobial formularies and ward rounds by infection specialists. vaccine (PCV), activates T lymphocytes, which results in a sustained response and immunological memory. Toxoids are bacterial toxins that have been modified to reduce toxicity but maintain antigenicity. Vaccine response can be improved by co-administration with mildly pro-inflammatory adjuvants, such as aluminium hydroxide. Use of vaccines Vaccination may be applied to entire populations or to subpopulations at specific risk through travel, occupation or other activities. In ring vaccination, the population immediately surrounding a case or outbreak of infectious disease is *Active immunisation is preferred if contact is with a patient who is within 1 week of onset of jaundice. 6.12 Indications for post-exposure prophylaxis with immunoglobulins Human normal immunoglobulin (pooled immunoglobulin)* • *Hepatitis A (unvaccinated contacts*) • Measles (exposed child with heart or lung disease) Human specific immunoglobulin • Hepatitis B (sexual partners, inoculation injuries, infants born to infected mothers) • Tetanus (high-risk wounds or incomplete or unknown immunisation status) • Rabies • Chickenpox (immunosuppressed children and adults, pregnant women) 6.13 Vaccines in current clinical use Live attenuated vaccines • Measles, mumps, rubella (MMR) • Oral poliomyelitis (OPV, not used in UK) • Rotavirus • Tuberculosis (bacille Calmette–Guérin, BCG) • Typhoid (oral typhoid vaccine) • Varicella zoster virus Inactivated (killed) whole-cell vaccines • Cholera • Hepatitis A • Influenza • Poliomyelitis (inactivated polio virus, IPV) •

Rabies Component vaccines • Anthrax (adsorbed extracted antigens) • Diphtheria (adsorbed toxoid) • Hepatitis B (adsorbed recombinant hepatitis B surface antigen, HBsAg) • Haemophilus influenzae type B (conjugated capsular polysaccharide) • Human papillomavirus (recombinant capsid proteins) • Meningococcal, quadrivalent A, C, Y, W135 (conjugated capsular polysaccharide) • Meningococcal, serogroup C (conjugated capsular polysaccharide) • Pertussis (adsorbed extracted antigens) • Pneumococcal conjugate (PCV; conjugated capsular polysaccharide, 13 serotypes) • Pneumococcal polysaccharide (PPV; purified capsular polysaccharide, 23 serotypes) • Tetanus (adsorbed toxoid) • Typhoid (purified Vi capsular polysaccharide) 6.14 Guidelines for vaccination against infectious disease • The principal contraindication to inactivated vaccines is an anaphylactic reaction to a previous dose or a vaccine component • Live vaccines should not be given during an acute infection, to pregnant women or to the immunosuppressed, unless the immunosuppression is mild and the benefits outweigh the risks • If two live vaccines are required, they should be given either simultaneously in opposite arms or 4 weeks apart • Live vaccines should not be given for 3 months after an injection of human normal immunoglobulin (HNI) • HNI should not be given for 2 weeks after a live vaccine • Hay fever, asthma, eczema, sickle-cell disease, topical glucocorticoid therapy, antibiotic therapy, prematurity and chronic heart and lung diseases, including tuberculosis, are not contraindications to vaccination Fig. 6.14 The relationship between infection prevention and control (IPC) and antimicrobial stewardship (AMS). Antimicrobial stewardship Infection prevention and control Effective antimicrobial stewardship reduces health care-associated infections Effective infection control reduces the need for antimicrobials

116 • PRINCIPLES OF INFECTIOUS DISEASE • when no single agent’s spectrum covers all potential pathogens (e.g. polymicrobial infection) • when there is a need to reduce development of antimicrobial resistance in the target pathogen, as the organism would need to develop resistance to multiple agents simultaneously (e.g. antituberculous chemotherapy, p. 592; antiretroviral therapy (ART), p. 324). Antimicrobial resistance Microorganisms have evolved in the presence of naturally occurring antibiotics and have therefore developed resistance mechanisms (categorised in Fig. 6.16) to all classes of antimicrobial agent (antibiotics and their derivatives). Intrinsic resistance is an innate property of a microorganism, whereas acquired resistance arises by spontaneous mutation or horizontal transfer of genetic material from another organism (e.g. via a plasmid, p. 100). Plasmids often encode resistance to multiple antibiotics. The mecA gene encodes a penicillin-binding protein, which has a low affinity for penicillins and therefore confers resistance to β-lactam antibiotics in staphylococci. Extended-spectrum β-lactamases (ESBLs) are frequently encoded on plasmids, which are transferred relatively easily between bacteria, including Enterobacteriaceae. Plasmid-encoded carbapenemases have been detected in strains of Klebsiella pneumoniae (e.g. New Delhi metallo-β-lactamase 1, NDM-1). Strains of MRSA have been described that also have reduced susceptibility to glycopeptides through the development of a relatively impermeable cell wall. Treatment of infectious diseases Key components of treating infection are: • optimising antimicrobial therapy while minimising selection for antimicrobial resistance and the impact on commensal flora • addressing predisposing factors, e.g. glycaemic control in diabetes mellitus; viral load control in HIV-1-associated opportunistic infection • considering adjuvant therapy, e.g. removal of an infected medical device or necrotic tissue • managing complications, e.g. severe sepsis (systemic inflammatory response syndrome, or SIRS, p. 196) and acute kidney injury (p. 411). For communicable disease, treatment must also take into account contacts of the infected patient, and may include IPC interventions such as isolation, antimicrobial prophylaxis, vaccination and contact tracing. Principles of antimicrobial therapy In some situations (e.g.

pneumonia) it is important to start appropriate antimicrobial therapy promptly, whereas in others prior confirmation of the diagnosis and pathogen is preferred. The principles underlying the choice of antimicrobial agent(s) are discussed below. The WHO ‘World Antibiotic Awareness Week’ campaign is a yearly event aimed at highlighting the importance of prudent antimicrobial prescribing (see ‘Further information’). Antimicrobial action and spectrum Antimicrobial agents may kill or inhibit microorganisms by targeting essential and non-essential cellular processes, respectively. The range, or spectrum, of microorganisms that is killed or inhibited by a particular antimicrobial agent needs consideration when selecting therapy. Mechanisms of action of the major classes of antibacterial agent are listed in Box 6.15 and appropriate agents for some common infecting organisms are shown in Box 6.16. In severe infections and/ or immunocompromised patients, it is customary to use bactericidal agents in preference to bacteriostatic agents. Empiric versus targeted therapy Empiric antimicrobial therapy is selected to treat a suspected infection (e.g. meningitis) before the microbiological cause is known. Targeted or ‘directed’ therapy can be prescribed when the pathogen(s) is known. Empirical antimicrobial regimens need to have activity against the range of pathogens that could be causing the infection in question; because broad-spectrum agents affect many more bacteria than needed, they select for antimicrobial resistance. ‘Start Smart – Then Focus’ (Fig. 6.15) describes the principle of converting from empiric therapy to narrow-spectrum targeted therapy. Optimum empiric therapy depends on the site of infection, patient characteristics and local antimicrobial resistance patterns. National or local guidelines are often used to inform antimicrobial prescribing decisions. Combination therapy It is sometimes appropriate to combine antimicrobial agents: • when there is a need to increase clinical effectiveness (e.g. biofilm infections) 6.15 Target and mechanism of action of common antibacterial agents Aminoglycosides, chloramphenicol, macrolides, lincosamides, oxazolidinones • Inhibition of bacterial protein synthesis by binding to subunits of bacterial ribosomes Tetracyclines • Inhibition of protein synthesis by preventing transfer RNA binding to ribosomes Beta-lactams • Inhibition of cell wall peptidoglycan synthesis by competitive inhibition of transpeptidases (‘penicillin-binding proteins’) Cyclic lipopeptide (daptomycin) • Insertion of lipophilic tail into plasma membrane causing depolarisation and cell death Glycopeptides • Inhibition of cell wall peptidoglycan synthesis by forming complexes with D-alanine residues on peptidoglycan precursors Nitroimidazoles • The reduced form of the drug causes strand breaks in DNA Quinolones • Inhibition of DNA replication by binding to DNA topoisomerases (DNA gyrase and topoisomerase IV), preventing supercoiling and uncoiling of DNA Rifamycins • Inhibition of DNA synthesis by inhibiting DNA-dependent RNA polymerase Sulphonamides and trimethoprim • Inhibition of folate synthesis by dihydropteroate synthase (sulphonamides) and dihydrofolate reductase (trimethoprim) inhibition

6.16 Antimicrobial options for common infecting bacteria Organism Antimicrobial options* Gram-positive organisms Enterococcus faecalis Ampicillin, vancomycin/teicoplanin Enterococcus faecium Vancomycin/teicoplanin, linezolid Glycopeptide-resistant enterococci Linezolid, tigecycline, daptomycin MRSA Clindamycin, vancomycin, rifampicin (never used as monotherapy), linezolid, daptomycin, tetracyclines, tigecycline, co-trimoxazole Staphylococcus aureus Flucloxacillin, clindamycin Streptococcus pyogenes Penicillin, clindamycin, vancomycin Streptococcus pneumoniae Penicillin, cephalosporins, levofloxacin, vancomycin Gram-negative organisms Escherichia coli, ‘coliforms’ (enteric Gram-negative bacilli) Amoxicillin, trimethoprim, cefuroxime,

ciprofloxacin, co-amoxiclav Enterobacter spp., Citrobacter spp. Ciprofloxacin, meropenem, ertapenem, aminoglycosides ESBL-producing Enterobacteriaceae Ciprofloxacin, meropenem, ertapenem (if sensitive), temocillin, aminoglycosides Carbapenemase-producing Enterobacteriaceae Ciprofloxacin, aminoglycosides, tigecycline, colistin Haemophilus influenzae Amoxicillin, co-amoxiclav, macrolides, cefuroxime, cefotaxime, ciprofloxacin Legionella pneumophila Azithromycin, levofloxacin, doxycycline Neisseria gonorrhoeae Ceftriaxone/cefixime, spectinomycin Neisseria meningitidis Penicillin, cefotaxime/ceftriaxone, chloramphenicol Pseudomonas aeruginosa Ciprofloxacin, piperacillin–tazobactam, aztreonam, meropenem, aminoglycosides, ceftazidime/cefepime Salmonella typhi Ceftriaxone, azithromycin (uncomplicated typhoid), chloramphenicol (resistance common) Strict anaerobes Bacteroides spp. Metronidazole, clindamycin, co-amoxiclav, piperacillin–tazobactam, meropenem Clostridium difficile Metronidazole, vancomycin (oral), fidaxomicin Clostridium spp. Penicillin, metronidazole, clindamycin Fusobacterium spp. Penicillin, metronidazole, clindamycin Other organisms Chlamydia trachomatis Azithromycin, doxycycline Treponema pallidum Penicillin, doxycycline *Antibiotic selection depends on multiple factors, including local susceptibility patterns, which vary enormously between geographical areas. There are many appropriate alternatives to those listed. (ESBL = extended-spectrum β-lactamase; MRSA = meticillin-resistant Staphylococcus aureus) Fig. 6.15 Stages in the selection and refinement of antimicrobial therapy: ‘Start Smart – Then Focus’. 1 Empiric therapy Based on: • Predicted susceptibility of likely pathogens • Local antimicrobial policies 2 Targeted therapy Based on: • Predicted susceptibility of infecting organism(s) • Local antimicrobial policies 3 Susceptibility-guided therapy Based on: • Susceptibility testing results Antimicrobial susceptibility results Clinical diagnosis Information available: • Organ system involved • Endogenous or exogenous infection • Likely pathogens • Infecting organism(s) • Likely antimicrobial susceptibility Level of knowledge of infecting organism(s) Antimicrobial spectrum of agent(s) used • Antimicrobial susceptibility of infecting organism(s) Laboratory investigations: microbiological diagnosis

 be stopped when there is no longer any clinical evidence of infection. Pharmacokinetics and pharmacodynamics Pharmacokinetics of antimicrobial agents determine whether adequate concentrations are obtained at the sites of infection. Septic patients often have poor gastrointestinal absorption, so the preferred initial route of therapy is intravenous. Knowledge of anticipated antimicrobial drug concentrations at sites of infection is critical. For example, achieving a ‘therapeutic’ blood level of gentamicin is of little practical use in treating meningitis, as CSF penetration of the drug is poor. Knowledge of routes of antimicrobial elimination is also critical; for instance, urinary tract infection is ideally treated with a drug that is excreted unchanged in the urine. Pharmacodynamics describes the relationship between antimicrobial concentration and microbial killing. For many agents, antimicrobial effect can be categorised as ‘concentrationdependent’ or ‘time-dependent’. The concentration of antimicrobial achieved after a single dose is illustrated in Figure 6.17. The maximum concentration achieved is Cmax and the measure of overall exposure is the area under the curve (AUC). The efficacy of antimicrobial agents whose killing is concentration-dependent (e.g. aminoglycosides) increases with the amount by which Cmax exceeds the minimum inhibitory concentration (Cmax : MIC ratio). For this reason, it has become customary to administer aminoglycosides (e.g. gentamicin) infrequently at high doses (e.g. 7 mg/kg) rather than frequently at low doses. This has the added advantage of minimising toxicity by reducing the likelihood Factors promoting antimicrobial resistance include the inappropriate use of antibiotics (e.g. to treat viral infections), inadequate

dosage or unnecessarily prolonged treatment, and use of antimicrobials as growth promoters in agriculture. However, any antimicrobial use exerts a selection pressure that favours the development of resistance. Combination antimicrobial therapy may reduce the emergence of resistance in the target pathogen but not in the normal flora that it also affects. Despite use of combination therapy for M. tuberculosis, multidrug-resistant tuberculosis (MDR-TB, resistant to isoniazid and rifampicin) and extremely drug-resistant tuberculosis (XDR-TB, resistant to isoniazid and rifampicin, any fluoroquinolone and at least one injectable antimicrobial antituberculous agent) have been reported worldwide and are increasing in incidence. The term ‘post-antibiotic era’ has been coined to describe a future in which the acquisition of resistance by bacteria will have been so extensive that antibiotic therapy is rendered useless. A more realistic scenario, which is currently being experienced, is a gradual but inexorable progression of resistance, necessitating the use of ever more toxic and expensive antimicrobials. Duration of therapy Treatment duration reflects the severity of infection and accessibility of the infected site to antimicrobial agents. For most infections, there is limited evidence available to support a specific duration of treatment (Box 6.17). Depending on the indication, initial intravenous therapy can often be switched to oral as soon as the patient is apyrexial and improving. In the absence of specific guidance, antimicrobial therapy should Fig. 6.16 Examples of mechanisms of antimicrobial resistance. (CAT = chloramphenicol acetyltransferase; ESBLs = extended-spectrum β-lactamases; MRSA = meticillin-resistant Staph. aureus; NDM-1 = New Delhi metallo-β-lactamase 1). Impermeability/reduced permeability Carbapenem resistance in Pseudomonas spp. Aminoglycoside resistance in anaerobes (uptake requires O2-dependent transport mechanism) Antimicrobial target Antimicrobial agent Active efflux of antimicrobial agent Tetracycline resistance in Gram-positive and Gram-negative bacteria Fluconazole resistance in Candida spp. Target modification β-lactam resistance in MRSA – altered penicillin-binding protein Glycopeptide resistance in enterococci – altered peptidoglycan amino acid sequence Rifampicin resistance in M. tuberculosis – RNA polymerase mutation Ciprofloxacin resistance in Enterobacteriaceae – DNA gyrase mutation Linezolid resistance in staphylococci and enterococci – 23S rRNA methylation Enzymatic degradation of agent β-lactam resistance in many organisms (penicillinase in Staph. aureus; ESBLs, ampC and NDM-1 in Enterobacteriaceae) Chloramphenicol resistance in staphylococci (CAT)

6.17 Duration of antimicrobial therapy for some common infections* Infection Duration of therapy Viral infections Herpes simplex encephalitis 2–3 weeks Bacterial infections Gonorrhoea Single dose Infective endocarditis (streptococcal, native valve) 4 weeks ± gentamicin for first 2 weeks Infective endocarditis (prosthetic valve) 6 weeks Osteomyelitis 6 weeks Pneumonia (community-acquired, severe) 7–10 days (no organism identified), 14–21 days (Staph. aureus or Legionella spp.) Septic arthritis 2–4 weeks Urinary tract infection (male) 2 weeks Urinary tract infection, upper tract, uncomplicated (female) 7 days Urinary tract infection, lower (female) 3 days Mycobacterial infections Tuberculosis (meningeal) 12 months Tuberculosis (pulmonary) 6 months Fungal infections Invasive pulmonary aspergillosis Until clinical/radiological resolution and reversal of predisposition Candidaemia (acute disseminated) 2 weeks after last positive blood culture and resolution of signs and symptoms *All recommendations are indicative. Actual duration takes into account predisposing factors, specific organisms and antimicrobial susceptibility, adjuvant therapies, current guidelines and clinical response. Fig. 6.17 Antimicrobial pharmacodynamics. The curve represents drug concentrations after a single dose of an antimicrobial agent. Factors that*

determine microbial killing are Cmax : MIC ratio (concentration-dependent killing), time above MIC (time-dependent killing) and AUC : MIC ratio. Time after dose Time above MIC Peak concentration (Cmax) Minimum inhibitory concentration (MIC) Area under the curve (AUC) Concentration of drug accumulation. Conversely, the β-lactam antibiotics and vancomycin exhibit time-dependent killing, and their efficacy depends on Cmax exceeding the MIC for a certain time (which is different for each class of agent). This is reflected in the dosing interval of benzylpenicillin, which is usually given every 4 hours in severe infection (e.g. meningococcal meningitis), and may be administered by continuous infusion. For other antimicrobial agents, the pharmacodynamic relationships are more complex and often less well understood. With some agents, bacterial inhibition persists after antimicrobial exposure (post-antibiotic and post-antibiotic sub-MIC effects). Therapeutic drug monitoring Therapeutic drug monitoring is used to confirm that levels of antimicrobial agents with a low therapeutic index (e.g. aminoglycosides) are not excessive, and that levels of agents with marked pharmacokinetic variability (e.g. vancomycin) are adequate. Specific recommendations for monitoring depend on individual clinical circumstances; for instance, different pre- and post-dose levels of gentamicin are recommended, depending on whether it is being used in traditional divided doses, once daily or for synergy in endocarditis (p. 530). Antimicrobial prophylaxis Antimicrobial prophylaxis is the use of antimicrobial agents to prevent infection. Primary prophylaxis is used to reduce the risk of infection following certain medical procedures (e.g. colonic resection or prosthetic hip insertion), following exposure to a specific pathogen (e.g. Bordetella pertussis) or in specific situations such as post-splenectomy (Box 6.18). It should be 6.18 Recommendations for antimicrobial prophylaxis in adults Infection risk Recommended antimicrobial Bacterial Diphtheria (prevention of secondary cases) Erythromycin Gas gangrene (after high amputation or major trauma) Penicillin or metronidazole Lower gastrointestinal tract surgery Cefuroxime + metronidazole, gentamicin + metronidazole, or co-amoxiclav (single dose only) Meningococcal disease (prevention of secondary cases) Rifampicin or ciprofloxacin Rheumatic fever (prevention of recurrence) Phenoxymethylpenicillin or sulfadiazine Tuberculosis (prevention of secondary cases) Isoniazid ± rifampicin Whooping cough (prevention of secondary cases) Erythromycin Viral HIV, occupational exposure (sharps injury) Combination tenofovir/ emtricitabine and raltegravir. Modified if index case's virus known to be resistant Influenza A (prevention of secondary cases in adults with chronic respiratory, cardiovascular or renal disease, immunosuppression or diabetes mellitus) Oseltamivir Fungal Aspergillosis (in high-risk haematology patients) Posaconazole (voriconazole or itraconazole alternatives if intolerant) Pneumocystis pneumonia (prevention in HIV and other immunosuppressed states) Co-trimoxazole, pentamidine or dapsone Protozoal Malaria (prevention of travelassociated disease) Specific antimalarials depend on travel itinerary (p. 278) *These are based on current UK practice. Recommendations may vary locally or nationally. Antimicrobial prophylaxis for infective endocarditis during dental procedures is not currently recommended in the UK.

 Beta-lactam antibiotics These antibiotics have a β-lactam ring structure and exert a bactericidal action by inhibiting enzymes involved in cell wall synthesis (penicillin-binding proteins, PBPs). They are classified in Box 6.21. Pharmacokinetics • Good drug levels are achieved in lung, kidney, bone, muscle and liver, and in pleural, synovial, pericardial and peritoneal fluids. • CSF levels are low, except when meninges are inflamed. • Activity is not inhibited in abscess (e.g. by low pH and PO2, high protein or neutrophils). • Beta-lactams are subject to an 'inoculum effect' – activity is reduced in the presence of a high organism burden (PBP expression is down-regulated by high organism density). • Generally safe in

pregnancy (except imipenem/cilastatin). Adverse effects Immediate (IgE-mediated) allergic reactions are rare but lifethreatening. Approximately 90% of patients who report a penicillin allergy do not have a true IgE-mediated allergy. Other reactions, such as rashes, fever and haematological effects (e.g. low white cell count), usually follow more prolonged therapy (more than 2 weeks). A large proportion of patients with infectious mononucleosis develop a rash if given aminopenicillins; this does not imply lasting allergy. The relationship between allergy to penicillin and allergy to cephalosporins depends on the specific cephalosporin used; there is significant cross-reactivity with first-generation cephalosporins but cross-reactivity to second-/ third-generation cephalosporins is less common. Avoidance of cephalosporins, however, is recommended in patients who have IgE-mediated penicillin allergy (p. 84). Cross-reactivity between penicillin and carbapenems is rare (approximately 1% by skin testing) and carbapenems may be administered if there are no suitable alternatives and appropriate resuscitation facilities are available. The monobactam aztreonam (p. 121) is the β-lactam least likely to cross-react in patients with penicillin allergy. Gastrointestinal upset and diarrhoea are common, and a mild reversible hepatitis is recognised with many β-lactams. More severe forms of hepatitis can be observed with flucloxacillin and co-amoxiclav. Leucopenia, thrombocytopenia, coagulation associated with minimal adverse effects. In the case of exposure, it may be combined with passive immunisation (see Box 6.12). Secondary prophylaxis is used in patients who have been treated successfully for an infection but remain predisposed to it. It is used in haemato-oncology patients in the context of fungal infection and in HIV-positive individuals with an opportunistic infection until a defined level of immune reconstitution is achieved. Antibacterial agents For details of antibacterial usage in pregnancy and old age, see Boxes 6.19 and 6.20. 6.20 Problems with antimicrobial therapy in old age • Clostridium difficile infection: all antibiotics predispose to some extent, but second- and third-generation cephalosporins, co-amoxiclav and fluoroquinolones (e.g. ciprofloxacin) especially so. • Hypersensitivity reactions: rise in incidence due to increased previous exposure. • Renal impairment: may be significant in old age, despite ‘normal’ creatinine levels (p. 386). • Nephrotoxicity: more likely, e.g. first-generation cephalosporins, aminoglycosides. • Accumulation of β-lactam antibiotics: may result in myoclonus, seizures or coma. • Reduced gastric acid production: gastric pH is higher, which causes increased penicillin absorption. • Reduced hepatic metabolism: results in a higher risk of isoniazid-related hepatotoxicity. • Quinolones: associated with delirium and may increase the risk of seizures. 6.21 Beta-lactam antibiotics Penicillins • Natural penicillins: benzylpenicillin, phenoxymethylpenicillin • Penicillinase-resistant penicillins: meticillin, flucloxacillin, nafcillin, oxacillin • Aminopenicillins: ampicillin, amoxicillin • Carboxy- and ureido-penicillins: ticarcillin, piperacillin, temocillin Cephalosporins • See Box 6.22 Monobactams • Aztreonam Carbapenems • Imipenem, meropenem, ertapenem, doripenem 1Data extracted from Joint Formulary Committee. British National Formulary (online). London: BMJ Group and Pharmaceutical Press; (medicinescomplete. com) [accessed on 16 March 2013]. • Glycopeptides • Linezolid • Meropenem • Penicillins 2Theoretical risk of teratogenicity, not supported by available clinical evidence. 6.19 Antimicrobial agents in pregnancy1 Contraindicated • Chloramphenicol: neonatal ‘grey baby’ syndrome – collapse, hypotension and cyanosis • Fluconazole: teratogenic in high doses • Quinolones: arthropathy in animal studies • Sulphonamides: neonatal haemolysis and methaemoglobinaemia • Tetracyclines, glycylcyclines: skeletal abnormalities in animals in first trimester; fetal dental discoloration and maternal hepatotoxicity with large parenteral doses in second or third trimesters • Trimethoprim: teratogenic in first trimester Relatively contraindicated • Aminoglycosides: potential damage to fetal auditory and vestibular nerves in second and third trimesters • Metronidazole: avoidance of high dosages is recommended2 Not known to be harmful;

use only when necessary • Aciclovir • Cephalosporins • Clarithromycin • Clindamycin • Erythromycin

retain good activity against Strep. pneumoniae and β-haemolytic streptococci. Ceftriaxone is administered once daily and is therefore a suitable agent for outpatient intravenous (parenteral) antimicrobial therapy (OPAT). • Fourth-generation agents, e.g. cefipime, have a broad spectrum of activity, including streptococci and some Gram-negatives, including P. aeruginosa. • Fifth-generation agents, such as ceftobiprole and ceftaroline, have an enhanced spectrum of Gram-positive activity that includes MRSA, and also have activity against Gram-negative bacteria; some, such as ceftobiprole, are active against P. aeruginosa. The spectrum of cephalosporins has also been enhanced by adding β-lactamase inhibitors. Monobactams Aztreonam is the only available monobactam. It is active against Gram-negative bacteria, except ESBL-producing organisms, but inactive against Gram-positive organisms or anaerobes. It is a parenteral-only agent and may be used safely in most penicillin-allergic patients other than those with an allergy to ceftazidime, which shares a common side chain with aztreonam. Carbapenems These intravenous agents have the broadest antibiotic activity of the β-lactam antibiotics, covering most clinically significant bacteria, including anaerobes (e.g. imipenem, meropenem, ertapenem). Macrolide and lincosamide antibiotics Macrolides (e.g. erythromycin, clarithromycin and azithromycin) and lincosamides (e.g. clindamycin) are bacteriostatic agents. Both classes bind to the same component of the ribosome, so they are not administered together. Macrolides are used for Legionella, Mycoplasma, Chlamydia and Bordetella infections. Azithromycin is employed for single-dose/short-course therapy for genitourinary Chlamydia/Mycoplasma spp. infections. Clindamycin is used primarily for skin, soft tissue, bone and joint infections. Pharmacokinetics Macrolides • Variable bioavailability (intravenous and oral preparations available). deficiencies, interstitial nephritis and potentiation of aminoglycosidemediated kidney damage are also recognised (p. 122). Seizures and encephalopathy have been reported, particularly with high doses in the presence of renal insufficiency. Thrombophlebitis occurs in up to 5% of patients receiving parenteral β-lactams. Drug interactions Synergism occurs in combination with aminoglycosides in vitro. Ampicillin decreases the biological effect of oral contraceptives and the whole class is significantly affected by concurrent administration of probenecid, producing a 2–4-fold increase in the peak serum concentration. Penicillins Natural penicillins are primarily effective against Gram-positive organisms (except staphylococci, most of which produce a penicillinase) and anaerobic organisms. Strep. pyogenes has remained sensitive to natural penicillins worldwide. According to the European Antimicrobial Resistance Surveillance Network (EARS-Net), the prevalence of non-susceptibility to penicillin in Strep. pneumoniae in Europe in 2014 varied widely from 0% (Cyprus) to 46.7% (Romania). Penicillinase-resistant penicillins are the mainstay of treatment for infections with Staph. aureus, other than MRSA. However, EARS-Net data from 2014 indicate that MRSA rates in Europe vary widely from 0.9% (Netherlands) to 56% (Romania). Aminopenicillins have the same spectrum of activity as the natural penicillins, with additional Gram-negative cover against Enterobacteriaceae. Amoxicillin has better oral absorption than ampicillin. Unfortunately, resistance to these agents is widespread (57.1% of E. coli Europe-wide in 2014, range 34.7–73%), so they are no longer appropriate for empirical use in Gram-negative infections. In many organisms, resistance is due to β-lactamase production, which can be overcome by the addition of β-lactamase inhibitors (clavulanic acid or sulbactam). Carboxypenicillins (e.g. ticarcillin) and

ureidopenicillins (e.g. piperacillin) are particularly active against Gram-negative organisms, especially Pseudomonas spp., which are resistant to the aminopenicillins. Beta-lactamase inhibitors may be added to extend their spectrum of activity (e.g. piperacillin–tazobactam). Temocillin is derived from ticarcillin; it has good activity against Enterobacteriaceae, including those that produce ESBL enzymes, but poor activity against Pseudomonas aeruginosa and Grampositive bacteria. Cephalosporins and cephamycins Cephalosporins are broad-spectrum agents. Unfortunately, their use is associated with CDI (p. 264). With the exception of ceftobiprole, the group has no activity against enterococci. Only the cephamycins have anti-anaerobic activity. All cephalosporins are inactivated by ESBL. Cephalosporins are arranged in ‘generations’ (Box 6.22). • First-generation compounds have excellent activity against Gram-positive organisms and some activity against Gram-negatives. • Second-generation drugs retain Gram-positive activity but have extended Gram-negative activity. Cephamycins (e.g. cefoxitin), included in this group, are active against anaerobic Gram-negative bacilli. • Third-generation agents further improve anti-Gramnegative cover. For some (e.g. ceftazidime), this is extended to include Pseudomonas spp. Cefotaxime and ceftriaxone have excellent Gram-negative activity and 6.22 Cephalosporins First generation • Cefalexin, cefradine (oral) • Cefazolin (IV) Second generation • Cefuroxime (oral/IV) • Cefaclor (oral) • Cefoxitin (IV) Third generation • Cefixime (oral) • Cefotaxime (IV) • Ceftriaxone (IV) • Ceftazidime (IV) Fourth generation • Cefepime (IV) Fifth generation (also referred to as ‘next generation’) • Ceftobiprole (IV) • Ceftaroline (IV)

122 • PRINCIPLES OF INFECTIOUS DISEASE are < 1 mg/L and 5–10 mg/L (7–10 mg/L with less sensitive organisms, e.g. P. aeruginosa), respectively. • For other aminoglycosides, consult local guidance. Adverse effects • Renal toxicity (usually reversible) accentuated by other nephrotoxic agents. • Cochlear toxicity (permanent) more likely in older people and those with a predisposing mitochondrial gene mutation. • Neuromuscular blockade after rapid intravenous infusion (potentiated by calcium channel blockers, myasthenia gravis and hypomagnesaemia). Spectinomycin Chemically similar to the aminoglycosides and given intramuscularly, spectinomycin was developed to treat strains of Neisseria gonorrhoeae resistant to β-lactam antibiotics. Unfortunately, resistance to spectinomycin is very common. Its only indication is the treatment of gonococcal urethritis in pregnancy or in patients allergic to β-lactam antibiotics. Quinolones and fluoroquinolones These are effective and generally well-tolerated bactericidal agents. The quinolones have purely anti-Gram-negative activity, whereas the fluoroquinolones are broad-spectrum agents (Box 6.23). Ciprofloxacin has anti-pseudomonal activity but resistance emerges rapidly. In 2014, European surveillance showed that resistance to fluoroquinolones in E. coli ranged between 7.8% (Iceland) and 46.4% (Cyprus). Quinolones and fluoroquinolones are used for a variety of common infections, including urinary tract infection and pneumonia, and less common problems like MDR-TB. Pharmacokinetics • Well absorbed after oral administration but delayed by food, antacids, ferrous sulphate and multivitamins. • Wide volume of distribution; tissue concentrations twice those in serum. • Good intracellular penetration, concentrating in phagocytes. Fig. 6.18 Dosing of aminoglycosides using the Hartford nomogram. The nomogram is used to determine the dose interval for 7 mg doses of gentamicin or tobramycin, using measurements of drug levels in plasma 6–14 hours after a single dose. Dose every 48 hours Dose every 36 hours Dose every 24 hours Hours since administration Concentration in plasma (μg/mL)

• Frequency of administration: erythromycin is administered 4 times daily, clarithromycin twice daily, azithromycin once daily. • High protein binding. • Excellent intracellular accumulation.

Lincosamides (e.g. clindamycin) • Good oral bioavailability. • Food has no effect on absorption. • Good bone/joint penetration; limited CSF penetration. Adverse effects • Gastrointestinal upset, especially in young adults (erythromycin 30%). • Cholestatic jaundice with erythromycin estolate. • Prolongation of QT interval can cause torsades de pointes (p. 476). • Clindamycin predisposes to CDI. Aminoglycosides and spectinomycin Aminoglycosides are effective mainly in Gram-negative infections and are therefore commonly used in regimens for intra-abdominal infection. Some aminoglycosides, e.g. amikacin, are important components of therapy for MDR-TB. Because they act synergistically with β-lactam antibiotics they are used in combinations to treat biofilm infections, including infective endocarditis and orthopaedic implant infections. They cause very little local irritation at injection sites and negligible allergic responses. Oto- and nephrotoxicity must be avoided by monitoring of renal function and drug levels and by use of short treatment regimens. Aminoglycosides are not subject to an inoculum effect (p. 120) and they all exhibit a post-antibiotic effect (p. 119). Pharmacokinetics • Negligible oral absorption. • Hydrophilic, so excellent penetration to extracellular fluid in body cavities and serosal fluids. • Very poor intracellular penetration (except hair cells in cochlea and renal cortical cells). • Negligible CSF and corneal penetration and may require intrathecal administration during neurosurgical infections. • Peak plasma levels 30 minutes after infusion. • Monitoring of therapeutic levels required. Gentamicin dosing • Except in certain forms of endocarditis, pregnancy, severe burns, end-stage renal disease and paediatric patients, gentamicin is administered at 7 mg/kg body weight. The appropriate dose interval depends on drug clearance and is determined by reference to the Hartford nomogram (Fig. 6.18). • In streptococcal and enterococcal endocarditis, gentamicin is used with a cell wall active agent (usually a β-lactam), to provide synergy. Commonly used doses are 1 mg/kg 2–3 times daily for enterococcal endocarditis and 3 mg/kg once daily for most strains of oral streptococci. Target pre- and post-dose levels are < 1 mg/L and 3–5 mg/L, respectively, when gentamicin is dosed 3 times daily. • When not used according to the Hartford regimen or for endocarditis, gentamicin is administered twice or 3 times daily at 3–5 mg/kg/day. Target pre- and post-dose levels

Adverse effects • Histamine release due to rapid vancomycin infusion produces a ‘red man’ reaction (rare with modern preparations). • Nephrotoxicity is rare but may occur with concomitant aminoglycoside use, as may ototoxicity. • Teicoplanin can cause rash, bronchospasm, eosinophilia and anaphylaxis. Lipopeptides Daptomycin is a cyclic lipopeptide with bactericidal activity against Gram-positive organisms (including MRSA and GRE) but not Gram-negatives. It is not absorbed orally, and is used intravenously to treat Gram-positive infections, such as soft tissue infections and Staph. aureus infective endocarditis. Daptomycin is not effective for community-acquired pneumonia. Treatment can be associated with increased levels of creatine kinase and eosinophilic pneumonitis. Polymyxins Colistin is a polymyxin antibiotic that binds and disrupts the outer cell membrane of Gram-negative bacteria, including P. aeruginosa and Acinetobacter baumannii. Its use has increased with the emergence and spread of multi-resistant Gram-negative bacteria, including CPEs. It can be administered by oral, intravenous and nebulised routes. Significant adverse effects include neurotoxicity, including encephalopathy, and nephrotoxicity. Folate antagonists These are bacteriostatic antibacterials (p. 109). A combination of a sulphonamide and either trimethoprim or pyrimethamine is most commonly used, which interferes with two consecutive steps in the metabolic pathway. Available combinations include trimethoprim/sulfamethoxazole (co-trimoxazole) and pyrimethamine with either sulfadoxine (used

to treat malaria) or sulfadiazine (used in toxoplasmosis). Co-trimoxazole is the first-line drug for Pneumocystis jirovecii infection, the second-line drug for treatment and prevention of B. pertussis (whooping cough) infection, and is also used for a variety of other infections, including Staph. aureus. Dapsone is used to treat leprosy (Hansen's disease) and to prevent toxoplasmosis and pneumocystis when patients are intolerant of other medications. Folinic acid should be given to prevent myelosuppression if these drugs are used long-term or unavoidably in early pregnancy. Pharmacokinetics • Well absorbed orally. • Sulphonamides are hydrophilic, distributing well to the extracellular fluid. • Trimethoprim is lipophilic with high tissue concentrations. Adverse effects • Trimethoprim is generally well tolerated, with few adverse effects. • Sulphonamides and dapsone may cause haemolysis in glucose-6-phosphate dehydrogenase deficiency (p. 948). • Sulphonamides and dapsone cause skin and mucocutaneous reactions, including Stevens–Johnson syndrome and 'dapsone syndrome' (rash, fever and lymphadenopathy). • Dapsone causes methaemoglobinaemia (p. 135) and peripheral neuropathy. Adverse effects • Gastrointestinal side-effects in 1–5%. • Rare skin reactions (phototoxicity). • Tendinitis and Achilles tendon rupture, especially in older people. • Central nervous system effects (delirium, tremor, dizziness and occasional seizures in 5–12%), especially in older people. • Reduces clearance of xanthines and theophyllines, potentially inducing insomnia and increased seizure potential. • Prolongation of QT interval on ECG, cardiac arrhythmias. • Ciprofloxacin use is associated with the acquisition of MRSA and emergence of C. difficile ribotype 027 (p. 264). Glycopeptides Glycopeptides (vancomycin and teicoplanin) are effective against Gram-positive organisms only, and are used against MRSA and ampicillin-resistant enterococci. Some staphylococci and enterococci demonstrate intermediate sensitivity or resistance. Vancomycin use should be restricted to limit emergence of resistant strains. Teicoplanin is not available in all countries. Neither drug is absorbed after oral administration but vancomycin is used orally to treat CDI. Pharmacokinetics Vancomycin • Administered by slow intravenous infusion, good tissue distribution and short half-life. • Enters the CSF only in the presence of inflammation and may require intrathecal administration during neurosurgical infections. • Therapeutic monitoring of intravenous vancomycin is recommended, to maintain pre-dose levels of > 10 mg/L (15–20 mg/L in serious staphylococcal infections). Teicoplanin • Long half-life allows once-daily dosing. Agent Route of administration Typical antimicrobial spectrum Quinolones Nalidixic acid Oral Enteric Gram-negative bacilli (not Pseudomonas aeruginosa) Fluoroquinolones Ciprofloxacin Norfloxacin Ofloxacin IV/oral Oral IV/oral/topical Enteric Gram-negative bacilli, P. aeruginosa, Haemophilus spp., 'atypical' respiratory pathogens* Levofloxacin (L-isomer of ofloxacin) IV/oral Haemophilus spp., Strep. pneumoniae, 'atypical' respiratory pathogens* Moxifloxacin Oral Strep. pneumoniae, Staph. aureus, 'atypical' respiratory pathogens*, Mycobacteria and anaerobes } 6.23 Quinolones and fluoroquinolones *'Atypical' pathogens include Mycoplasma pneumoniae and Legionella spp. Fluoroquinolones have variable activity against M. tuberculosis and other mycobacteria.

 H. influenzae, Strep. pneumoniae and N. meningitidis. It has a very broad spectrum of activity against aerobic and anaerobic organisms, spirochaetes, Rickettsia, Chlamydia and Mycoplasma spp. It competes with macrolides and lincosamides for ribosomal binding sites, so should not be used in combination with these agents. Significant adverse effects are 'grey baby' syndrome in infants (cyanosis and circulatory collapse due to inability to conjugate drug and excrete the active form in urine); reversible dose-dependent bone marrow depression in adults receiving high cumulative doses; and severe aplastic anaemia in 1 in 25 000–40 000 exposures (unrelated to dose, duration of therapy or route of administration).

Oxazolidinones Linezolid and tedizolid are examples and their good activity against Gram-positive organisms means they are often used to treat skin and soft tissue infections. They may also be used in infection caused by resistant Gram-positive bacteria, including MRSA and GRE. Administration can be intravenous or oral. Common linezolid adverse effects include mild gastrointestinal upset and tongue discoloration. Myelodysplasia and peripheral and optic neuropathy can occur with prolonged use. Linezolid has monoamine oxidase inhibitor (MAOI) activity, and co-administration with other MAOIs or serotonin re-uptake inhibitors should be avoided, as this may precipitate a ‘serotonin syndrome’ (p. 1199). Other antibacterial agents Fusidic acid This antibiotic, active against Gram-positive bacteria, is available in intravenous, oral or topical formulations. It is lipid-soluble and distributes well to tissues. Its antibacterial activity is, however, unpredictable. Fusidic acid is used in combination, typically with antistaphylococcal penicillins, or for MRSA with clindamycin or rifampicin. It interacts with coumarin derivatives and oral contraceptives. Nitrofurantoin This drug has very rapid renal elimination and is active against aerobic Gram-negative and Gram-positive bacteria, including enterococci. It is used only for treatment of urinary tract infection, being generally safe in pregnancy and childhood. With prolonged treatment, however, it can cause eosinophilic lung infiltrates, fever, pulmonary fibrosis, peripheral neuropathy, hepatitis and haemolytic anaemia so its use must be carefully balanced against risks. Fidaxomicin Fidaxomicin is an inhibitor of RNA synthesis, and was introduced for the treatment of CDI in 2012. In non-severe CDI it is noninferior to oral vancomycin and is associated with a lower recurrence rate. Its effectiveness has not been assessed in severe CDI. Fosfomycin Fosfomycin acts by inhibiting cell wall synthesis. It has activity against Gram-negative but also Gram-positive bacteria and can demonstrate in vitro synergy against MRSA when combined with other antimicrobials. It is used for treatment of urinary tract infections but can be employed in other situations against multi-resistant bacteria. Tetracyclines and glycylcyclines Tetracyclines Of this mainly bacteriostatic class, the newer drugs doxycycline and minocycline show better absorption and distribution than older ones. Many streptococci and Gram-negative bacteria are now resistant, in part due to their use in animals (which is banned in Europe). Tetracyclines are indicated for Mycoplasma spp., Chlamydia spp., Rickettsia spp., Coxiella spp., Bartonella spp., Borrelia spp., Helicobacter pylori, Treponema pallidum and atypical mycobacterial infections. Tetracyclines can also be used for malaria prevention. Pharmacokinetics • Best oral absorption is in the fasting state (doxycycline is 100% absorbed unless gastric pH rises) and absorption is inhibited by cations, e.g. calcium or iron, which should not be administered at the same time. Adverse effects • All tetracyclines except doxycycline are contraindicated in renal failure. • Dizziness with minocycline. • Binding to metallic ions in bones and teeth causes discoloration (avoid in children and pregnancy) and enamel hypoplasia. • Oesophagitis/oesophageal ulcers with doxycycline. • Phototoxic skin reactions. Glycylcyclines (tigecycline) Chemical modification of tetracycline has produced tigecycline, a broad-spectrum, parenteral-only antibiotic with activity against resistant Gram-positive and Gram-negative pathogens, such as MRSA and ESBL (but excluding Pseudomonas spp.). Re-analysis of trial data has shown that there was excess mortality following tigecycline treatment as opposed to comparator antibiotics, so tigecycline should be used only when there has been adequate assessment of risk versus benefit. Nitroimidazoles Nitroimidazoles are highly active against strictly anaerobic bacteria, especially Bacteroides fragilis, C. difficile and other Clostridium spp. They also have significant antiprotozoal activity against amoebae and Giardia lamblia. Pharmacokinetics • Almost completely absorbed after oral administration (60% after rectal administration). • Well distributed, especially to brain and CSF. • Safe in pregnancy. Adverse effects • Metallic taste (dose-dependent). • Severe vomiting if taken with alcohol – ‘Antabuse

effect’. • Peripheral neuropathy with prolonged use. Phenicols Chloramphenicol is the only example in clinical use. In developed countries its use tends to be reserved for severe and lifethreatening infections when other antibiotics are either unavailable or impractical, largely because of concerns about toxicity. It is bacteriostatic to most organisms but apparently bactericidal to

neurotoxicity); and paraminosalicylic acid (which causes rashes and gastrointestinal upset). Linezolid may also be used and has good CSF penetration, while meropenem with co-amoxiclav is occasionally chosen. New drugs developed for XDR-TB include delamanid and bedaquiline; their adverse effects include QT prolongation and cardiac arrhythmias. Their co-administration with other agents with a similar side-effect profile (e.g. fluoroquinolones) therefore requires careful risk assessment. Clofazimine Clofazimine is used against M. leprae and resistant strains of M. tuberculosis. Its mode of action may involve induction of oxidative stress and it is weakly bactericidal. Oral absorption is variable and it is excreted in the bile. Side-effects include gastrointestinal upset, dry eyes and skin, and skin pigmentation, especially in those with pigmented skin. Antifungal agents See Box 6.24. Antimycobacterial agents Isoniazid Isoniazid is bactericidal for replicating bacteria and bacteriostatic for non-replicating bacteria. It is activated by mycobacterial catalase-peroxidase (KatG) and inhibits the InhA gene product, a reductase involved in mycolic acid synthesis. Mutations in KatG or InhA result in isoniazid resistance, which was reported in 15% of cases of M. tuberculosis infection globally in 2013. Isoniazid is well absorbed orally and metabolised by acetylation in the liver. The major side-effects are hepatitis, neuropathy (ameliorated by co-administration of pyridoxine) and hypersensitivity reactions. Rifampicin Rifampicin inhibits DNA-dependent RNA polymerase and is bactericidal against replicating bacteria. It is also active in necrotic foci, where mycobacteria have low levels of replication, and is therefore important in sterilisation and sputum conversion. Resistance most often involves the β-subunit of RNA polymerase and most often occurs with isoniazid-resistant MDR-TB. Rifampicin is well absorbed orally. It is metabolised by the liver via the microsomal cytochrome P450 system and is one the most potent inducers of multiple P450 isoenzymes, so is subject to extensive drug–drug interactions. Common side-effects include hepatitis, influenza-like symptoms and hypersensitivity reactions. Orange discoloration of urine and body secretions is expected. Pyrazinamide The mechanism of action of pyrazinamide is incompletely defined but includes inhibition of fatty acid synthase and ribosomal trans-translation. Pyrazinamide is often bacteriostatic but can be bactericidal and is active against semidormant bacteria in a low-pH environment. Primary resistance is rare but MDR-TB strains are frequently pyrazinamide-resistant and intrinsic resistance is a feature of Mycobacterium bovis strains. Pyrazinamide is well absorbed orally and metabolised by the liver. Side-effects include nausea, hepatitis, asymptomatic elevation of uric acid and myalgia. Ethambutol Ethambutol is a bacteriostatic agent. It inhibits arabinosyl transferase, which is involved in the synthesis of arabinogalactan, a component of the mycobacterial cell wall. Resistance is usually seen when resistance to other antimycobacterial agents is also present, e.g. in MDR-TB strains. It is orally absorbed but, in contrast to the first-line agents described above, it achieves poor CSF penetration and is renally excreted. The major side-effect is optic neuritis with loss of red–green colour discrimination and impaired visual acuity. It can also cause hepatitis. Streptomycin Streptomycin is an aminoglycoside whose mechanism of action and side-effects are similar to those of other aminoglycosides. It is administered intramuscularly. Other antituberculous agents Second-line agents used in MDR or XDR strains (p. 116) include aminoglycosides (amikacin,

capreomycin or kanamycin) and fluoroquinolones (moxifloxacin or levofloxacin), discussed above. Other established second-line agents administered orally are cycloserine (which causes neurological side-effects); ethionamide or prothionamide (which are not active with InhA-gene-mediated resistance but have reasonable CSF penetration; their side-effect profile includes gastrointestinal disturbance, hepatotoxicity and Agent Usual route(s) of administration Clinically relevant antifungal spectrum Imidazoles Miconazole Econazole Clotrimazole Topical Candida spp., dermatophytes Ketoconazole Topical, oral Malassezia spp., dermatophytes, agents of eumycetoma Triazoles Fluconazole Oral, IV Yeasts (Candida and Cryptococcus spp.) Itraconazole Oral, IV Yeasts, dermatophytes, dimorphic fungi (p. 300), Aspergillus spp. Voriconazole Oral, IV Yeasts and most filamentous fungi (excluding mucoraceous moulds) Posaconazole Oral, IV Yeasts and many filamentous fungi (including most mucoraceous moulds) Isavuconazole Oral, IV Yeasts and many filamentous fungi (variable activity against mucoraceous moulds) Echinocandins Anidulafungin Caspofungin Micafungin IV only Candida spp., Aspergillus spp. (no activity against Cryptococcus spp. or mucoraceous moulds) Polyenes Amphotericin B Nystatin IV Topical Yeasts and most dimorphic and filamentous fungi (including mucoraceous moulds) Others 5-fluorocytosine Oral, IV Yeasts Griseofulvin Oral Dermatophytes Terbinafine Topical, oral Dermatophytes } } 6.24 Antifungal agents

 is similar. Lipid formulations of AmB are used in invasive fungal disease, as empirical therapy in patients with neutropenic fever (p. 1327), and also in visceral leishmaniasis (p. 282). Other antifungal agents Flucytosine Flucytosine (5-fluorocytosine) has particular activity against yeasts. When it is used as monotherapy, acquired resistance develops rapidly, so it should be given in combination with another antifungal agent. Adverse effects include myelosuppression, gastrointestinal upset and hepatitis. Griseofulvin Griseofulvin has been largely superseded by terbinafine and itraconazole for treatment of dermatophyte infections, except in children, for whom these agents remain largely unlicensed. It is deposited in keratin precursor cells, which become resistant to fungal invasion. Terbinafine Terbinafine distributes with high concentration to sebum and skin, with a half-life of more than 1 week. It is used topically for dermatophyte skin infections and orally for onychomycosis. The major adverse reaction is hepatic toxicity (approximately 1: 50 000 cases). Terbinafine is not recommended for breastfeeding mothers. Antiviral agents Most viral infections in immunocompetent individuals resolve without intervention. Antiviral therapy is available for a limited number of infections only (Box 6.25). Antiretroviral agents These agents, used predominantly against HIV, are discussed on page 324. Anti-herpesvirus agents Aciclovir, valaciclovir, penciclovir and famciclovir These antivirals are acyclic analogues of guanosine, which inhibit viral DNA polymerase after being phosphorylated by virus-derived thymidine kinase (TK). Aciclovir is poorly absorbed after oral dosing; better levels are achieved intravenously or by use of the prodrug valaciclovir. Famciclovir is the prodrug of penciclovir. Resistance is mediated by viral TK or polymerase mutations. Ganciclovir Chemical modification of the aciclovir molecule allows preferential phosphorylation by protein kinases of cytomegalovirus (CMV) and other β-herpesviruses (e.g. human herpesvirus (HHV) 6/7) and hence greater inhibition of the DNA polymerase, but at the expense of increased toxicity. Ganciclovir is administered intravenously or as a prodrug (valganciclovir) orally. Cidofovir Cidofovir inhibits viral DNA polymerases with potent activity against CMV, including most ganciclovir-resistant CMV. It also has activity against aciclovir-resistant herpes simplex virus (HSV) and varicella zoster virus (VZV), HHV6 and occasionally adenovirus, poxvirus, papillomavirus or polyoma virus, and may be used to treat these infections in immunocompromised hosts. Azole antifungals The azoles

(imidazoles and triazoles) inhibit synthesis of ergosterol, a constituent of the fungal cell membrane. Side-effects vary but include gastrointestinal upset, hepatitis and rash. Azoles are inhibitors of cytochrome P450 enzymes, so tend to increase exposure to cytochrome P450-metabolised drugs (p. 24). Imidazoles Miconazole, econazole, clotrimazole and ketoconazole are relatively toxic and therefore administered topically. Clotrimazole is used extensively to treat superficial fungal infections. Triazoles are used for systemic treatment because they are less toxic. Triazoles Fluconazole is effective against yeasts (Candida and Cryptococcus spp.) and has a long half-life (approximately 30 hours) and an excellent safety profile. The drug is highly water-soluble and distributes widely to all body sites and tissues, including CSF. Itraconazole is lipophilic and distributes extensively, including to toenails and fingernails. CSF penetration is poor. Because oral absorption is erratic, therapeutic drug monitoring is required. Voriconazole is well absorbed orally but variability in levels requires therapeutic drug monitoring. It is used mainly in aspergillosis (p. 596). Side-effects include photosensitivity, hepatitis and transient retinal toxicity. Posaconazole and isavuconazole are broad-spectrum azoles, with activity against Candida spp., Aspergillus spp. and some mucoraceous moulds. Isavuconazole is non-inferior to voriconazole in the management of invasive aspergillosis and may be considered as an alternative when voriconazole is not tolerated. Echinocandins The echinocandins inhibit β-1,3-glucan synthesis in the fungal cell wall. They have few significant adverse effects. Caspofungin, anidulafungin and micafungin are used to treat systemic candidosis, and caspofungin is also used in aspergillosis. Polyenes Amphotericin B (AmB) deoxycholate causes cell death by binding to ergosterol and damaging the fungal cytoplasmic membrane. Its use in resource-rich countries has been largely supplanted by less toxic agents. Its long half-life enables once-daily administration. CSF penetration is poor. Adverse effects include immediate anaphylaxis, other infusionrelated reactions and nephrotoxicity. Nephrotoxicity may be sufficient to require dialysis and occurs in most patients who are adequately dosed. It may be ameliorated by concomitant infusion of normal saline. Irreversible nephrotoxicity occurs with large cumulative doses of AmB. Nystatin has a similar spectrum of antifungal activity to AmB. Its toxicity limits it to topical use, e.g. in oral and vaginal candidiasis. Lipid formulations of amphotericin B Lipid formulations of AmB have been developed to reduce AmB toxicity and have replaced AmB deoxycholate in many regions. They consist of AmB encapsulated in liposomes (liposomal AmB, L-AmB) or complexed with phospholipids (AmB lipid complex, ABLC). The drug becomes active on dissociating from its lipid component. Adverse effects are similar to, but considerably less frequent than, those with AmB deoxycholate, and efficacy

cases of influenza, e.g. in intensive care units. It is now approved for use in adults in a number of countries. An intravenous formulation of zanamivir is also in development for critically ill patients. Laninamivir is approved as an intranasal formulation in Japan. Amantadine and rimantadine These drugs reduce replication of influenza A by inhibition of viral M2 protein ion channel function, which is required for uncoating (see Fig. 6.2). Resistance develops rapidly and is widespread, and amantadine and rimantadine should be used only if the prevalence of resistance locally is known to be low. They are no longer recommended for treatment or prophylaxis in the UK or USA, having been superseded by zanamivir and oseltamivir. However, they may still be indicated to treat oseltamivir-resistant influenza A in patients unable to take zanamivir (e.g. ventilated patients). Other agents used to treat viruses Antiviral agents used to treat hepatitis B and C virus are discussed on pages 875 and 878, and those used against HIV-1 are described on page 324.

Foscarnet This analogue of inorganic pyrophosphate acts as a non-competitive inhibitor of HSV, VZV, HHV6/7 or CMV DNA polymerase. It does not require significant intracellular phosphorylation and so may be effective when HSV or CMV resistance is due to altered drug phosphorylation. It has variable CSF penetration. Anti-influenza agents Zanamivir and oseltamivir These agents inhibit influenza A and B neuraminidase, which is required for release of virus from infected cells (see Fig. 6.2, p. 101). They are used in the treatment and prophylaxis of influenza. Administration within 48 hours of disease onset reduces the duration of symptoms by approximately 1–112 days. In the UK, their use is limited mainly to adults with chronic respiratory or renal disease, significant cardiovascular disease, immunosuppression or diabetes mellitus, during known outbreaks. Peramivir has been developed as a distinct chemical structure, which means that it retains activity against some oseltamivir- and zanamivir-resistant strains. It has poor oral bioavailability and has been developed as an intravenous or intramuscular formulation for treatment of severe Drug Route(s) of administration Indications Significant side-effects Antiretroviral therapy (ART, p. 324) Oral HIV infection (including AIDS) CNS symptoms, anaemia, lipodystrophy Anti-herpesvirus agents Aciclovir Topical/oral/IV Herpes zoster Chickenpox (esp. in immunosuppressed) Herpes simplex infections: encephalitis (IV only), genital tract, oral, ophthalmic Significant side-effects rare Hepatitis, renal impairment and neurotoxicity reported rarely Valaciclovir Oral Herpes zoster, herpes simplex As for aciclovir Famciclovir Oral Herpes zoster, herpes simplex (genital) As for aciclovir Penciclovir Topical Labial herpes simplex Local irritation Ganciclovir IV Treatment and prevention of CMV infection in immunosuppressed Gastrointestinal symptoms, liver dysfunction, neurotoxicity, myelosuppression, renal impairment, fever, rash, phlebitis at infusion sites Potential teratogenicity Valganciclovir Oral Treatment and prevention of CMV infection in immunosuppressed As for ganciclovir but neutropenia is predominant Cidofovir IV/topical HIV-associated CMV infections and occasionally other viruses (see text) Renal impairment, neutropenia Foscarnet IV CMV and aciclovir-resistant HSV and VZV infections in immunosuppressed Gastrointestinal symptoms, renal impairment, electrolyte disturbances, genital ulceration, neurotoxicity Anti-influenza agents Zanamivir Inhalation Influenza A and B Allergic reactions (very rare) Oseltamivir Oral Influenza A and B Gastrointestinal side-effects, rash, hepatitis (very rare) Peramivir IV, IM Amantadine, rimantadine Oral Influenza A (but see text) CNS symptoms, nausea Agents used in other virus infections* Ribavirin Oral/IV/inhalation Lassa fever (IV) RSV infection in infants (inhalation) Haemolytic anaemia, cough, dyspnoea, bronchospasm and ocular irritation (when given by inhalation) } 6.25 Antiviral agents *Antiviral agents used in viral hepatitis are discussed on pages 875 and 878. (AIDS = acquired immunodeficiency syndrome; CMV = cytomegalovirus; CNS = central nervous system; HIV = human immunodeficiency virus; HSV = herpes simplex virus; IM = intramuscular; IV = intravenous; RSV = respiratory syncytial virus; VZV = varicella zoster virus)

128 • PRINCIPLES OF INFECTIOUS DISEASE Lumefantrine Lumefantrine is used in combination with artemether to treat uncomplicated falciparum malaria, including chloroquine-resistant strains. Its mechanism of action is unknown. Significant adverse effects are uncommon. Drugs used in trypanosomiasis Benznidazole Benznidazole is an oral agent used to treat South American trypanosomiasis (Chagas’ disease, p. 279). Significant and common adverse effects include dose-related peripheral neuropathy, purpuric rash and granulocytopenia. Eflornithine Eflornithine inhibits biosynthesis of polyamines by ornithine decarboxylase inhibition, and is used in West African trypanosomiasis (T. brucei gambiense infection) of the central nervous system. It is administered as an intravenous infusion 4 times daily, which may be logistically difficult in the geographical areas affected by this disease. Significant adverse effects are common and include convulsions,

gastrointestinal upset and bone marrow depression. Melarsoprol This is an arsenical agent, used to treat central nervous system infections in East and West African trypanosomiasis (T. brucei rhodesiense and gambiense). It is administered intravenously. Melarsoprol treatment is associated with peripheral neuropathy and reactive arsenical encephalopathy (RAE), which carries a significant mortality. Nifurtimox Nifurtimox is administered orally to treat South American trypanosomiasis (Chagas' disease). Gastrointestinal and neurological adverse effects are common. Pentamidine isetionate Pentamidine is an inhibitor of DNA replication used in West African trypanosomiasis (T. brucei gambiense) and, to a lesser extent, in visceral and cutaneous leishmaniasis. It is also prescribed in Pneumocystis jirovecii pneumonia. It is administered via intravenous or intramuscular routes. It is a relatively toxic drug, commonly causing rash, renal impairment, profound hypotension (especially on rapid infusion), electrolyte disturbances, blood dyscrasias and hypoglycaemia. Suramin Suramin is a naphthaline dye derivative, used to treat East African trypanosomiasis (T. brucei rhodesiense). It is administered intravenously. Adverse effects are common and include rash, gastrointestinal disturbance, blood dyscrasias, peripheral neuropathies and renal impairment. Other antiprotozoal agents Pentavalent antimonials Sodium stibogluconate and meglumine antimoniate inhibit protozoal glycolysis by phosphofructokinase inhibition. They are used parenterally (intravenous or intramuscular) to treat leishmaniasis. Adverse effects include arthralgia, myalgias, raised hepatic transaminases, pancreatitis and electrocardiogram changes. Severe cardiotoxicity leading to death is not uncommon. Ribavirin Ribavirin is a guanosine analogue that inhibits nucleic acid synthesis in a variety of viruses. It is used in particular in the treatment of hepatitis C virus but also against certain viral haemorrhagic fevers, e.g. Lassa fever, although it has not been useful against Ebola virus. Antiparasitic agents Antimalarial agents Artemisinin (qinghaosu) derivatives Artemisinin originates from a herb (sweet wormwood, Artemisia annua), which was used in Chinese medicine to treat fever. Its derivatives, artemether and artesunate, were developed for use in malaria in the 1970s. Their mechanism of action is unknown. They are used in the treatment, but not prophylaxis, of malaria, usually in combination with other antimalarials, and are effective against strains of Plasmodium spp. that are resistant to other antimalarials. Artemether is lipid-soluble and may be administered via the intramuscular and oral routes. Artesunate is water-soluble and is administered intravenously or orally. Serious adverse effects are uncommon. Current advice for malaria in pregnancy is that the artemisinin derivatives should be used to treat uncomplicated falciparum malaria in the second and third trimesters, but should not be prescribed in the first trimester until more information becomes available. Atovaquone Atovaquone inhibits mitochondrial function. It is an oral agent, used for treatment and prophylaxis of malaria, in combination with proguanil (see below), without which it is ineffective. It is also employed in the treatment of mild cases of Pneumocystis jirovecii pneumonia, where there is intolerance to co-trimoxazole. Significant adverse effects are uncommon. Folate synthesis inhibitors (proguanil, pyrimethamine–sulfadoxine) Proguanil inhibits dihydrofolate reductase and is used for malaria prophylaxis. Pyrimethamine–sulfadoxine may be used in the treatment of malaria. Quinoline-containing compounds Chloroquine and quinine are believed to act by intraparasitic inhibition of haem polymerisation, resulting in toxic build-up of intracellular haem. The mechanisms of action of other agents in this group (quinidine, amodiaquine, mefloquine, primaquine, etc.) may differ. They are employed in the treatment and prophylaxis of malaria. Primaquine is used for radical cure of malaria due to Plasmodium vivax and P. ovale (destruction of liver hypnozoites). Chloroquine may also be given for extra-intestinal amoebiasis. Chloroquine can cause a pruritus sufficient to compromise adherence to therapy. If used in long-term, high-dose regimens, it causes an irreversible retinopathy. Overdosage leads to

lifethreatening cardiotoxicity. The side-effect profile of mefloquine includes neuropsychiatric effects ranging from mood change, nightmares and agitation to hallucinations and psychosis. Quinine may cause hypoglycaemia and cardiotoxicity, especially when administered parenterally. Primaquine causes haemolysis in people with glucose-6-phosphate dehydrogenase deficiency (p. 948), which should be excluded before therapy. Chloroquine is considered safe in pregnancy but mefloquine should be avoided in the first trimester.

Ivermectin Ivermectin binds to helminth nerve and muscle cell ion channels, causing increased membrane permeability. It is an oral agent, used in Strongyloides infection, filariasis and onchocerciasis. Significant side-effects are uncommon. Niclosamide Niclosamide inhibits oxidative phosphorylation, causing paralysis of helminths. It is an oral agent, used in Taenia saginata and intestinal T. solium infection. Systemic absorption is minimal and it has few significant side-effects. Piperazine Piperazine inhibits neurotransmitter function, causing helminth muscle paralysis. It is an oral agent, used in ascariasis and threadworm (Enterobius vermicularis) infection. Significant adverse effects are uncommon but include neuropsychological reactions such as vertigo, delirium and convulsions. Praziquantel Praziquantel increases membrane permeability to Ca2+, causing violent contraction of worm muscle. It is the drug of choice for schistosomiasis and is also used in T. saginata, T. solium (cysticercosis) and fluke infections (Clonorchis, Paragonimus) and in echinococcosis. It is administered orally and is well absorbed. Adverse effects are usually mild and transient, and include nausea and abdominal pain. Pyrantel pamoate This agent causes spastic paralysis of helminth muscle through a suxamethonium-like action. It is used orally in ascariasis and threadworm infection. Systemic absorption is poor and adverse effects are uncommon. Thiabendazole Thiabendazole inhibits fumarate reductase, which is required for energy production in helminths. It is used orally in Strongyloides infection and topically to treat cutaneous larva migrans. Significant adverse effects are uncommon. Further information Websites cdc.gov Centers for Disease Control and Prevention, Atlanta, USA. Provides information on all aspects of communicable disease, including prophylaxis against malaria. dh.gov.uk UK Department of Health. The publications section provides current UK recommendations for immunisation. ecdc.europa.eu European Centre for Disease Prevention and Control. Includes data on prevalence of antibiotic resistance in Europe. gov.uk/government/organisations/public-health-england Public Health England. Provides information on infectious diseases relating mainly to England, including community infection control. idsociety.org Infectious Diseases Society of America. Publishes up-to-date, evidence-based guidelines. who.int World Health Organization. Provides up-to-date information on global aspects of infectious disease, including outbreak updates. Also has information on the ‘World Antibiotic Awareness Week’ campaign. Diloxanide furoate This oral agent is used to eliminate luminal cysts following treatment of intestinal amoebiasis, or in asymptomatic cyst excreters. The drug is absorbed slowly (enabling luminal persistence) and has no effect in hepatic amoebiasis. It is a relatively non-toxic drug, the most significant adverse effect being flatulence. Iodoquinol (di-iodohydroxyquinoline) Iodoquinol is a quinoline derivative (p. 128) with activity against Entamoeba histolytica cysts and trophozoites. It is used orally to treat asymptomatic cyst excreters or, in association with another amoebicide (e.g. metronidazole), to treat extra-intestinal amoebiasis. Long-term use of this drug is not recommended, as neurological adverse effects include optic neuritis and peripheral neuropathy. Nitazoxanide Nitazoxanide is an inhibitor of pyruvate–ferredoxin oxidoreductasedependent anaerobic energy metabolism in

protozoa. It is a broad-spectrum agent, active against various nematodes, tapeworms, flukes and intestinal protozoa. Nitazoxanide also has activity against some anaerobic bacteria and viruses. It is administered orally in giardiasis and cryptosporidiosis. Adverse effects are usually mild and involve the gastrointestinal tract (e.g. nausea, diarrhoea and abdominal pain). Paromomycin Paromomycin is an aminoglycoside (p. 122) that is used to treat visceral leishmaniasis and intestinal amoebiasis. It is not significantly absorbed when administered orally, and is therefore given orally for intestinal amoebiasis and by intramuscular injection for leishmaniasis. It showed early promise in the treatment of HIV-associated cryptosporidiosis but subsequent trials have demonstrated that this effect is marginal at best. Drugs used against helminths Benzimidazoles (albendazole, mebendazole) These agents act by inhibiting both helminth glucose uptake, causing depletion of glycogen stores, and fumarate reductase. Albendazole is used for hookworm, ascariasis, threadworm, Strongyloides infection, trichinellosis, Taenia solium (cysticercosis) and hydatid disease. Mebendazole is used for hookworm, ascariasis, threadworm and whipworm. The drugs are administered orally. Absorption is relatively poor but is increased by a fatty meal. Significant adverse effects are uncommon. Bithionol Bithionol is used to treat fluke infections with Fasciola hepatica. It is well absorbed orally. Adverse effects are mild (e.g. nausea, vomiting, diarrhoea, rashes) but relatively common (approximately 30%). Diethylcarbamazine Diethylcarbamazine (DEC) is an oral agent used to treat filariasis and loiasis. Treatment of filariasis is often followed by fever, headache, nausea, vomiting, arthralgia and prostration. This is caused by the host response to dying microfilariae, rather than the drug, and may be reduced by pre-treatment with glucocorticoids.

7KLVSDJHLQWHQWLRQDOO\OHIWEODQN