6 Cellular mechanisms: host defence

OVERVIEW

Everyone has experienced an inflammatory episode at some time or other and will be familiar with the characteristic redness, swelling, heat, pain and loss of function that this generally entails. In this chapter we list the cellular players involved in the host defence response and explain the bare bones of this crucial and sophisticated mechanism; inflammatory mediators are considered separately in Chapter 17. Understanding these cellular responses and their functions provides an essential basis for understanding the actions of anti-inflammatory and immunosuppressant drugs—a major class of therapeutic agents (see Ch. 26).

The inflammatory response

- The inflammatory response occurs in tissues following exposure to a pathogen or other noxious substance.
- It usually has two components: an *innate* non-adaptive response and an *adaptive* (acquired or specific) immunological response.
- These reactions are generally protective, but if inappropriately deployed they are deleterious.
- The normal outcome of the response is healing with or without scarring; alternatively, if the underlying cause persists, chronic inflammation.
- Many of the diseases that require drug treatment involve inflammation. Understanding the action and use of anti-inflammatory and immunosuppressive drugs necessitates understanding the inflammatory reaction.

INTRODUCTION

All living creatures are born into a universe that poses a constant challenge to their physical well-being and survival. Evolution, which has equipped us with homeostatic systems that maintain a stable internal environment in the face of changing external temperatures and fluctuating supplies of food and water, has also provided us with mechanisms for combating the ever-present threat of infection and for promoting healing and restoration to normal function in the event of injury. In mammals, this function is subserved by the *innate* and *acquired* (or *adaptive*) immune systems, working together with a variety of mediators and mechanisms that collectively gives rise to what we term *inflammation*. Generally this response acts to protect us, but occasionally it goes awry, leading to a spectrum of inflammatory diseases, and it is under these circumstances that we need to resort to drug therapy to dampen or abolish the inflammatory response.

The main functions of this host inflammatory response then are *defence* and *repair*—in other words, nothing less than the security of the organism. This response is crucial to survival. If it is defective either through genetic causes (e.g. *leukocyte adhesion deficiency*), infection with organisms that subvert its function (e.g. HIV) or because of immunosuppressant drug therapy, then the outcome can be very serious or even fatal.

Like border security systems in the mundane world, the body has the cellular and molecular equivalents of guards, identity checks, alarm systems and a communication network with which to summon back-up when required. It also has access to an astonishing data bank that memorises precise details of previous illegal immigrants and prevents them from returning. In this discussion it is convenient to divide the host response into two components, although it should be recognised at the outset that these two systems work hand-in-hand. The two principal components are:

1. The *innate*, non-adaptive response, which developed early in evolution and is present in some form or other in most multicellular organisms. This is the first line of defence.
2. The *adaptive* immune response. This appeared much later in evolutionary terms and is found only in vertebrates. It provides the physical basis for our immunological 'memory' and is the second line of defence.

THE INNATE IMMUNE RESPONSE

The innate response is activated immediately following infection or injury.[1] It is a system that is present in virtually all organisms and some of the mammalian gene families that control these responses were first identified in plants and insects.

PATHOGEN RECOGNITION

One of the most important functions of any security system is the ability to establish identity. How does an organism decide whether a cell is a bona fide citizen or an invading pathogen? In the case of the innate response this is achieved through a network of *pattern recognition receptors* (PRRs), found in virtually all organisms. They recognise *pathogen-associated molecular patterns* (PAMPs), common products produced by bacteria, fungi, viruses and so on that these organisms could not readily change to evade detection.

[1]Mucosal epithelial tissues constantly secrete antibacterial proteins and a type of 'all purpose' immunoglobulin (Ig)A as a sort of pre-emptive defensive strategy. One immunologist aptly referred to the innate response as the organism's 'knee jerk' response to infection; it is an excellent description.

The innate immune response

- The innate response occurs immediately on injury or infection. It comprises vascular and cellular elements. Mediators generated by cells or from plasma modify and regulate the magnitude of the response.
- Utilising Toll and other receptors, sentinel cells in body tissues, such as macrophages, mast and dendritic cells, detect specific pathogen-associated molecular patterns. This triggers the release of cytokines, particularly interleukin (IL)-1 and tumour necrosis factor (TNF)-α, as well as various chemokines.
- IL-1 and TNF-α act on local postcapillary venular endothelial cells, causing:
  - vasodilatation and fluid exudation
  - expression of adhesion molecules on the cell surfaces.
- Exudate contains enzyme cascades that generate bradykinin (from kininogen), and C5a and C3a (from complement). Complement activation lyses bacteria.
- C5a and C3a stimulate mast cells to release histamine, which dilates local arterioles.
- Tissue damage and cytokines release prostaglandins PGI_2 and PGE_2 (vasodilators) and leukotriene $(LT)B_4$ (a chemotaxin).
- Cytokines stimulate synthesis of vasodilator nitric oxide, which increases vascular permeability.
- Using adhesion molecules, leukocytes roll on, adhere to and finally migrate through activated vascular endothelium towards the pathogen (attracted by chemokines, IL-8, C5a, and LTB_4), where phagocytosis and killing takes place.

These receptors include G-protein-coupled receptors such as the *FPR* (formyl peptide receptor) family that recognises N-formylated peptides characteristic of bacterial protein synthesis (although these are also liberated from mitochondria during host cell death as well) and cytoplasmic receptors such as the *NOD-like receptors* (**n**ucleotide-binding **o**ligomerization **d**omain-like receptors)—a large family of intracellular proteins that recognise fragments of bacterial proteoglycan.

Among the best-studied of these PRRs are the *Toll-like receptors* (TLRs). The Toll[2] gene was first identified in *Drosophila* in the mid-1990s. Analogous genes were soon found in vertebrates and it was quickly established that as a family, their main job was to detect highly conserved components in pathogens and to signal their presence to the different components of the immune system.

There are approximately 15 TLRs known but only some 10 occur in mammals. They belong to the class of *receptor tyrosine kinases* (see Ch. 3), and are phylogenetically highly conserved. Unlike the antigen receptors on T and B cells that are generated somatically as the cells develop, endowing each lymphocyte clone with a structurally unique receptor, TLRs are encoded in the host DNA. Table 6.1 lists these receptors and the pathogenic products that are recognised, where these are known. There are two types of TLR, located respectively on the cell surface and in endosomes. The latter type generally recognises pathogen RNA/DNA (presumably because they appear in phagosomes), while the former recognises other pathogen components such as cell wall material, endotoxin, etc. Some TLRs also recognise ligands released when host cells are damaged (e.g. heat shock proteins). Presumably this provides an additional way of monitoring damage.

How a single family of receptors can recognise such a wide spectrum of different chemicals is a molecular mystery. Sometimes the problem is solved by recruiting additional 'accessory' binding proteins to assist this process. When activated, Toll receptors dimerise and initiate a complex signalling pathway that activates genes coding for proteins and factors crucial to the deployment of the inflammatory response, many of which we will discuss below. Interestingly from the pharmacological viewpoint, TLR 7 also recognises some synthetic antiviral compounds such as *imidazoquinolones*. The ability of these drugs to provoke TLR activation probably underlies their clinical effectiveness.

TLRs are strategically located on those 'sentinel' cells which are most likely to come into contact with pathogens in the first instance. These include *mast cells*, *macrophages* and *dendritic cells*, all of which are found in tissues throughout the body, as well as some *intestinal epithelial cells* (which are exposed to pathogens in the food that we eat) and other cells.

Having outlined how 'non-self' pathogens are detected by the innate immune system, we can now describe the events that follow the 'raising of the alarm'.

RESPONSES TO PATTERN RECOGNITION

Vascular events

Interaction of a PAMP with TLRs triggers the sentinel cells to respond immediately by producing the main pro-inflammatory *cytokines*, *tumour necrosis factor* (*TNF*)-α and *interleukin* (*IL*)-*1*, as well as other mediators (such as prostaglandins and histamine) that act on the vascular endothelial cells of the postcapillary venules, causing expression of *adhesion molecules* on the intimal surface and an increase in vascular permeability.

White blood cells adhere to the endothelial cells through interactions between their cell surface *integrins* (see below) and adhesion molecules on endothelial cells. This enables them to migrate out of the vessels, attracted by *chemotaxins* generated by the microorganisms or as a result of their interaction with tissues (see below). *Chemokines* released during TLR activation play an important part in this. (Cytokines and chemokines are considered in Ch. 17.)

The initial vascular events include dilatation of the small arterioles, resulting in increased blood flow. This is followed by a slowing and eventually a stasis of blood, and an increase in the permeability of the postcapillary venules with exudation of fluid. The vasodilatation is brought about by mediators including histamine, prostaglandin $(PG)E_2$ and PGI_2 (prostacyclin) produced by the interaction of the microorganism with tissue, some of which act together with cytokines to increase vascular permeability.

The fluid exudate contains the components for four proteolytic enzyme cascades: the *complement system*, the *coagulation system*, the *fibrinolytic system* and the *kinin* system (see

[2]The name, which loosely translates from German as 'Great!' or 'Eureka!', has remained firmly attached to the family.

Table 6.1 The TLR family of pattern recognition receptors (PRRs)

PRR	Pathogen recognised	Ligand	Host cell type	Location
TLR 1	Bacteria	Lipoproteins	Monocyte/macrophages Some dendritic cells B lymphocytes	Surface
TLR 2	Bacteria Bacteria (Gm pos) Parasites Yeast Damaged host cells	Lipoproteins Lipoteichoic acid GPI anchors Cell wall carbohydrates Heat shock proteins	Monocyte/macrophages Some dendritic cells Mast cells	Surface
TLR 3	Virus	dsRNA	Dendritic cells B lymphocytes	Intracellular
TLR 4	Bacteria (Gm neg) Virus Damaged host cells	Lipopolysaccharide Some viral proteins Heat shock proteins Fibrinogen Hyaluronic acid	Monocyte/macrophages Some dendritic cells Mast cells Intestinal epithelium	Surface
TLR 5	Bacteria	Flagellin	Monocyte/macrophages Some dendritic cells Intestinal epithelium	Surface
TLR 6	Mycoplasma Parasites Yeast	Lipoproteins GPI anchors Cell wall carbohydrates	Monocyte/macrophages Mast cells B lymphocytes	Surface
TLR 7	Virus	ssRNA Some synthetic drugs	Monocyte/macrophages Mast cells B lymphocytes	Intracellular
TLR 8	Virus	ssRNA	Monocyte/macrophages Some dendritic cells Mast cells	Intracellular
TLR 9	Virus/bacteria	CpG containing DNA	Monocyte/macrophages Some dendritic cells B lymphocytes	Intracellular
TLR 10	Unknown	Unknown	Monocyte/macrophages B lymphocytes	Surface
TLR 11[a]	*Toxoplasma*	Profilin	Monocyte/macrophages Liver cells Kidney	Surface

[a] TLR 11 is found in mouse but not human. TLR 12–15 are not included as little is known concerning their function.
CpG DNA, unmethylated CG dinucleotide; dsRNA, double stranded RNA; Gm neg/pos, Gram negative/positive (bacteria); GPI, glycosylphosphatidylinositol anchoring proteins; ssRNA, single stranded RNA.

Fig. 6.1). The components of these cascades are proteases that are inactive in their native form but that are activated by proteolytic cleavage, each activated component then activating the next. The exudate is carried by lymphatics to local lymph nodes or lymphoid tissue, where the products of the invading microorganism trigger the adaptive phase of the response.

▼ The *complement system* comprises nine major components, designated C1 to C9. Activation of the cascade is initiated by substances derived from microorganisms, such as yeast cell walls or endotoxins. This pathway of activation is termed the *alternative pathway* (Fig. 6.1) as opposed to the classic pathway that is dealt with later. One of the main events is the enzymatic splitting of C3, giving rise to various peptides, one of which, *C3a* (termed an *anaphylatoxin*) stimulates mast cells to secrete further chemical mediators and can also directly stimulate smooth muscle, while *C3b* (termed an *opsonin*) attaches to the surface of a microorganism, facilitating ingestion by white blood cells. *C5a*, generated enzymatically from C5, also releases mediators from mast cells and is a powerful chemotactic attractant and activator of white blood cells.

The final components in the sequence, complement-derived mediators (C5 to C9) coalesce to form a 'membrane attack complex' that attaches to certain bacterial membranes, leading to lysis. Complement can therefore mediate the destruction of invading bacteria or damage multicellular parasites; however, it may sometimes cause injury to the host. The principal enzymes of the coagulation and fibrinolytic cascades, thrombin and plasmin, can also activate the cascade by hydrolysing C3, as can enzymes released from white blood cells.

The *coagulation system* and the *fibrinolytic system* are described in Chapter 24. Factor XII is activated to XIIa (e.g. by collagen), and the end product, fibrin, laid down during a host–pathogen interaction, may serve to limit the extent of the infection. Thrombin is additionally

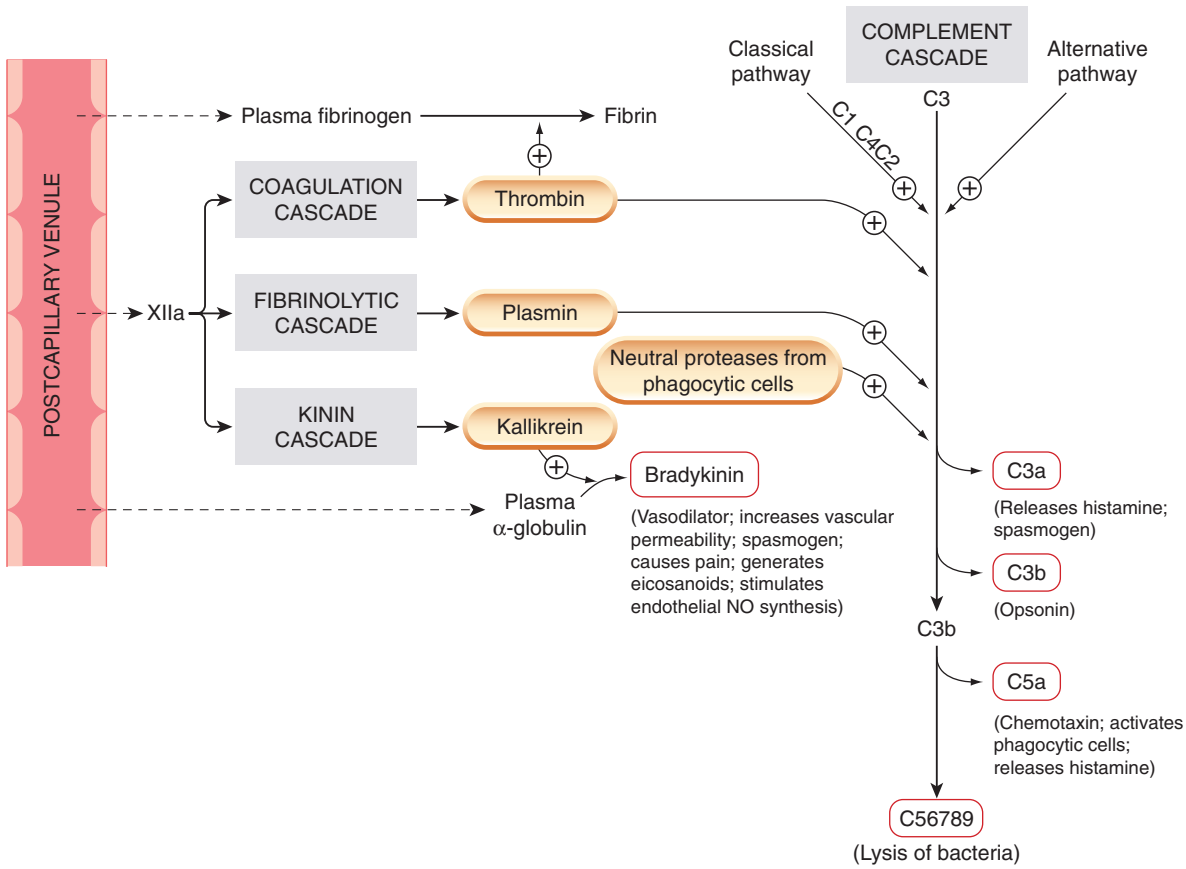

Fig. 6.1 Four enzyme cascades are activated when plasma leaks out into the tissues as a result of the increased vascular permeability of inflammation. Factors causing exudation are depicted in Figure 6.2. Mediators generated are shown in red-bordered boxes. Complement components are indicated by C1, C2, etc. When plasmin is formed, it tends to increase kinin formation and decrease the coagulation cascade. (Adapted from Dale M M, Foreman J C, Fan T-P (eds) 1994 Textbook of immunopharmacology, 3rd edn. Blackwell Scientific, Oxford.)

involved in the activation of the kinin (Fig. 6.1) and, indirectly, the fibrinolytic systems (see Ch. 24).

The *kinin system* is another enzyme cascade relevant to inflammation. It yields several mediators, in particular bradykinin (Fig. 6.1 and see below).

Cellular events

Of the cells involved in inflammation, some (e.g. vascular endothelial cells, mast cells, dendritic cells and tissue macrophages) are normally present in tissues, while other actively motile cells (e.g. leukocytes) gain access from the blood.

Polymorphonuclear leukocytes

Neutrophil polymorphs are the 'shock troops' of inflammation, and are the first of the blood leukocytes to enter an inflamed area (Fig. 6.2). The whole process is cleverly choreographed: under direct observation, the neutrophils may be seen first to *roll* along the activated endothelium, then to *adhere* and finally to *migrate* out of the blood vessel and into the extravascular space. This process is regulated by the successive activation of different families of adhesion molecules (*selectins*, *intercellular adhesion molecule* [ICAM] and *integrins*) on the inflamed endothelium that engage corresponding *counter-ligands* on the neutrophil, capturing it as it rolls along the surface, stabilising its interaction with the endothelial cells and enabling it to migrate out of the vessel (using a further adhesion molecule termed *PECAM*, **p**latelet **e**ndothelium **a**dhesion **m**olecule). The neutrophil is attracted to the invading pathogen by chemicals termed *chemotaxins*, some of which (such as the tripeptide formyl-Met-Leu-Phe) are released by the microorganism, whereas others, such as C5a, are produced locally or released by nearby cells such as macrophages (e.g. chemokines such as IL-8).

Neutrophils can engulf, kill and digest microorganisms. Together with eosinophils, they have surface receptors for C3b, which acts as an *opsonin* that forms a link between neutrophil and invading bacterium. (An even more effective link may be made by antibody; see below.) Neutrophils kill microorganisms by generating toxic oxygen products and other mechanisms, and enzymatic digestion then follows. If the neutrophil is inappropriately activated, these weapons can cause damage to the host's own tissues. When neutrophils have released their toxic chemicals, they undergo apoptosis and must be cleared by macrophages. It is this mass of live and apoptotic neutrophils that constitutes 'pus'.

Mast cells

An important 'sentinel' cell that expresses TLRs, the mast cell also has surface receptors both for IgE and for the

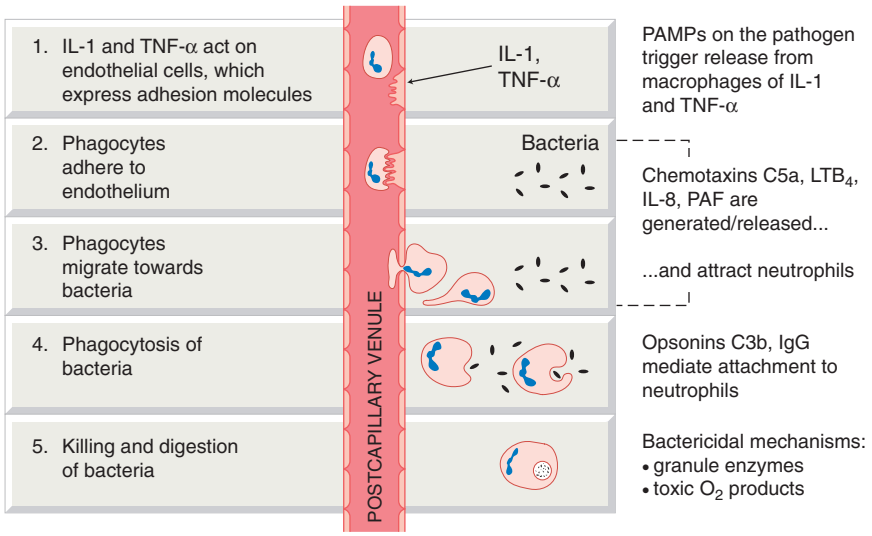

Fig. 6.2 Simplified diagram of the initial events in a local acute inflammatory reaction. Recognition by tissue macrophages of pathogen-associated molecular patterns (PAMPs) on the pathogen triggers release, from tissue macrophages, of the proinflammatory cytokines interleukin (IL)-1 and tumour necrosis factor (TNF)-α. These act on the endothelial cells of postcapillary venules, causing exudation of fluid and expression of adhesion factors (e.g. selectins, integrins) to which counter-ligands on blood-borne neutrophils adhere. Subsequent steps are listed in the figure. C5a and C3b, complement components; IgG, immunoglobulin G; LTB_4, leukotriene B_4; PAF, platelet-activating factor.

complement-derived *anaphylotoxins* C3a and C5a. Ligands acting at these receptors trigger mediator release, as does direct physical damage. One of the main substances released is *histamine*; others include *heparin*, *leukotrienes*, *PGD_2*, *platelet-activating factor (PAF)*, *nerve growth factor* and some *interleukins*. Unusually, mast cells have pre-formed packets of cytokines that they can release when stimulated. This makes them extremely effective triggers of the inflammatory response.

Monocytes/macrophages

Monocytes arrive in inflammatory lesions several hours after the polymorphs. Adhesion to endothelium and migration into the tissue follow a pattern similar to that of the neutrophils (see above), although monocyte chemotaxis utilises additional chemokines, such as MCP-1[3] (which, reasonably enough, stands for **m**onocyte **c**hemoattractant **p**rotein-1) and RANTES (which very *unreasonably* stands for **r**egulated on **a**ctivation **n**ormal **T** cell **e**xpressed and **s**ecreted: immunological nomenclature has excelled itself here!).

Once in tissues, blood monocytes differentiate into macrophages.[4] The resultant 'sentinel' cell has a remarkable range of abilities, being not only a jack-of-all-trades but also master of many (see below). Activation of TLRs stimulates the generation and release of chemokines and other cytokines that act on vascular endothelial cells, attract other leukocytes to the area and give rise to systemic manifestations of the inflammatory response such as fever. Macrophages engulf tissue debris and dead cells, as well as phagocytosing and killing most (but unfortunately not all) microorganisms. They also play an important part in *antigen presentation* (see below). When stimulated by glucocorticoids, macrophages secrete *annexin-1* (a potent anti-inflammatory polypeptide; see Ch. 32), which controls the extent of the local inflammatory reaction.

[3]Human immunodeficiency virus-1 binds to the surface CD4 glycoprotein on monocytes/macrophages but is able to penetrate the cell only after binding also to MCP-1 and RANTES receptors.

[4]Literally 'big eaters', compared with neutrophils, originally called microphages or 'little eaters'.

Dendritic cells

These are present in many tissues, especially when they subserve a barrier function (e.g. the skin, where they are sometimes referred to as *Langerhans cells* after their discoverer). As an important 'sentinel cell' they can recognise the presence of pathogens and when thus activated they can migrate into lymphoid tissue, where they play an important part in antigen presentation (see below).

Eosinophils

These cells have similar capacities to neutrophils but are also 'armed' with a battery of substances stored in their granules, which, when released, kill multicellular parasites (e.g. helminths). These include *eosinophil cationic protein*, a *peroxidase* enzyme, the *eosinophil major basic protein* and a *neurotoxin*. The eosinophil is considered by many to be of primary importance in the pathogenesis of the late phase of asthma where, it is suggested, granule proteins cause damage to bronchiolar epithelium (see Fig. 27.4).

Basophils

Basophils are very similar in many respects to mast cells. Except in certain parasitic infections and hypersensitivity reactions, the basophil content of the tissues is negligible and in health they form only 0.5% of circulating white blood cells.

Vascular endothelial cells

Vascular endothelial cells (see also Chs 22 and 23), originally considered as passive lining cells, are now known to play an active part in inflammation. Small arteriole endothelial cells secrete nitric oxide (NO), causing relaxation of the underlying smooth muscle (see Ch. 20), vasodilatation and increased delivery of plasma and blood cells to the inflamed area. The endothelial cells of the postcapillary venules regulate plasma exudation and thus the delivery of plasma-derived mediators (see Fig. 6.1). Vascular endothelial cells express several adhesion molecules (the ICAM and selectin families; see Fig. 6.2), as well as a variety of receptors including those for histamine, acetylcholine and IL-1. In addition to NO, the cells can synthesise and release the vasodilator agents PGI_2 and PGE_2, the vasoconstrictor agent endothelin, plasminogen activator, PAF and

several cytokines. Endothelial cells also participate in the angiogenesis that occurs during inflammatory resolution, chronic inflammation and cancer (see Chs 5 and 55).

Platelets

Platelets are involved primarily in coagulation and thrombotic phenomena (see Ch. 24) but also play a part in inflammation. They have low-affinity receptors for IgE, and are believed to contribute to the first phase of asthma (Fig. 27.1). In addition to generating thromboxane $(TX)A_2$ and PAF, they can generate free radicals and proinflammatory cationic proteins. Platelet-derived growth factor contributes to the repair processes that follow inflammatory responses or damage to blood vessels.

Natural killer cells

Natural killer (NK) cells are a specialised type of lymphocyte. In an unusual twist to the receptor concept, NK cells kill targets (e.g. virus-infected or tumour cells) that lack ligands for inhibitory receptors on the NK cells themselves. The ligands in question are the *major histocompatibility complex* (MHC) molecules, and any cells lacking these become a target for NK-cell attack, a strategy sometimes called the 'mother turkey strategy'.[5] MHC proteins are expressed on the surface of most host cells and, in simple terms, are specific for that individual, enabling the NK cells to avoid damaging host cells. NK cells have other functions: they are equipped with Fc receptors and, in the presence of antibodies directed against a target cell, they can kill the cell by antibody-dependent cellular cytotoxicity.

THE ADAPTIVE IMMUNE RESPONSE

The adaptive response provides the physical basis for an 'immunological memory'. It provides a more powerful defence than the innate response as well as being highly specific for the invading pathogen. Here we will provide only a simplified outline and stressing those aspects relevant for an understanding of drug action; for more detailed coverage, see Janeway et al. (2004).

The key cells are the *lymphocytes*. These are long-lived cells derived from precursor cells within the bone marrow. Following release into the blood, they mature in the bone or thymus after which they enter the circulation and dwell in the lymphoid tissues such as the lymph nodes and spleen. Here, they are poised to detect, intercept and identify foreign proteins presented to them by *antigen presenting cells* (APCs) such as the macrophage or the dendritic cells. The three main groups of lymphocytes are:

1. *B cells*, which mature in the bone marrow. They are responsible for antibody production, i.e. the *humoral* immune response.
2. *T cells*, which mature in the thymus. They are important in the induction phase of the immune response and in cell-mediated immune reactions.
3. *NK (natural killer) cells*. These are really part of the innate system. They are activated by *interferons* and release cytotoxic granules that destroy target cells identified as 'foreign'.

[5]Richard Dawkins in *River out of Eden*, citing the zoologist Schliedt, explains that the 'rule of thumb a mother turkey uses to recognise nest robbers is a dismayingly brusque one; in the vicinity of the nest, attack anything that moves unless it makes a noise like a baby turkey' (quoted by Kärre & Welsh, 1997).

The adaptive response

- The adaptive (specific, acquired) immunological response boosts the effectiveness of the innate responses. It has two phases, the induction phase and the effector phase, the latter consisting of (i) antibody-mediated and (ii) cell-mediated components.
- During the *induction phase*, naive T cells bearing either the CD4 or the CD8 co-receptors are presented with antigen, triggering proliferation:
  - CD8-bearing T cells develop into cytotoxic T cells that can kill virally infected cells
  - CD4-bearing T-helper (Th) cells are stimulated by different cytokines to develop into Th1, Th2, Th17 or Treg cells
  - *Th1* cells develop into cells that release cytokines that activate macrophages; these cells, along with cytotoxic T cells, control cell-mediated responses
  - *Th2* cells control antibody-mediated responses by stimulating B cells to proliferate, giving rise to antibody-secreting plasma cells and memory cells
  - *Th17* cells are similar to Th1 cells and are important in some human diseases such as rheumatoid arthritis
  - *Treg* cells restrain the development of the immune response.
- The *effector phase* depends on antibody- and cell-mediated responses.
- Antibodies provide:
  - more selective complement activation
  - more effective pathogen phagocytosis
  - more effective attachment to multicellular parasites, facilitating their destruction
  - direct neutralisation of some viruses and of some bacterial toxins.
- Cell-mediated reactions involve:
  - CD8+ cytotoxic T cells that kill virus-infected cells
  - cytokine-releasing CD4+ T cells that enable macrophages to kill intracellular pathogens such as the tubercle bacillus
  - memory cells primed to react rapidly to a known antigen.
- Inappropriately deployed immune reactions are termed *hypersensitivity reactions*.
- Anti-inflammatory and immunosuppressive drugs are used when the normally protective inflammatory and/or immune responses escape control.

Miraculously, T and B lymphocytes express antigen-specific receptors that recognise and react with virtually all foreign proteins and polysaccharides that we are likely to encounter during our lifetime. This receptor repertoire is generated randomly and so would recognise 'self' proteins as well as foreign antigens if it were not that *tolerance* to self antigens is acquired during fetal life by apoptotic deletion of T-cell clones that recognise the host's own tissues. Dendritic cells and macrophages involved in the innate response also have a role in preventing harmful immune reactions against the host's own cells (see below).

The specific immune response occurs in two phases termed the *induction phase* and the *effector phase*.

THE INDUCTION PHASE

During the induction phase, antigen is 'presented' to T cells by macrophages or large *dendritic cells*, and this is followed by complex interactions of those T cells with B cells and other T cells (Fig. 6.3). The antigen may constitute part of an invading pathogen (e.g. the coat of a bacterium) or be released by such an organism (e.g. a bacterial toxin), or it may be a vaccine or a substance introduced experimentally in the laboratory to study the immune response (e.g. the injection of egg albumin into the guinea pig). APCs ingest and proteolytically 'process' the antigen and 'present' it on their surface to lymphocytes in combination with various MHC molecules once they reach local lymph nodes (Fig. 6.4). Two types of lymphocytes 'attend' APCs. They are generally distinguished by the presence, on their surface, of *CD4* or *CD8* receptors. These are *co-receptors* that cooperate with the main antigen-specific receptors in antigen recognition. Macrophages also carry surface CD4 proteins.

The two types of lymphocyte involved in the adaptive response are:

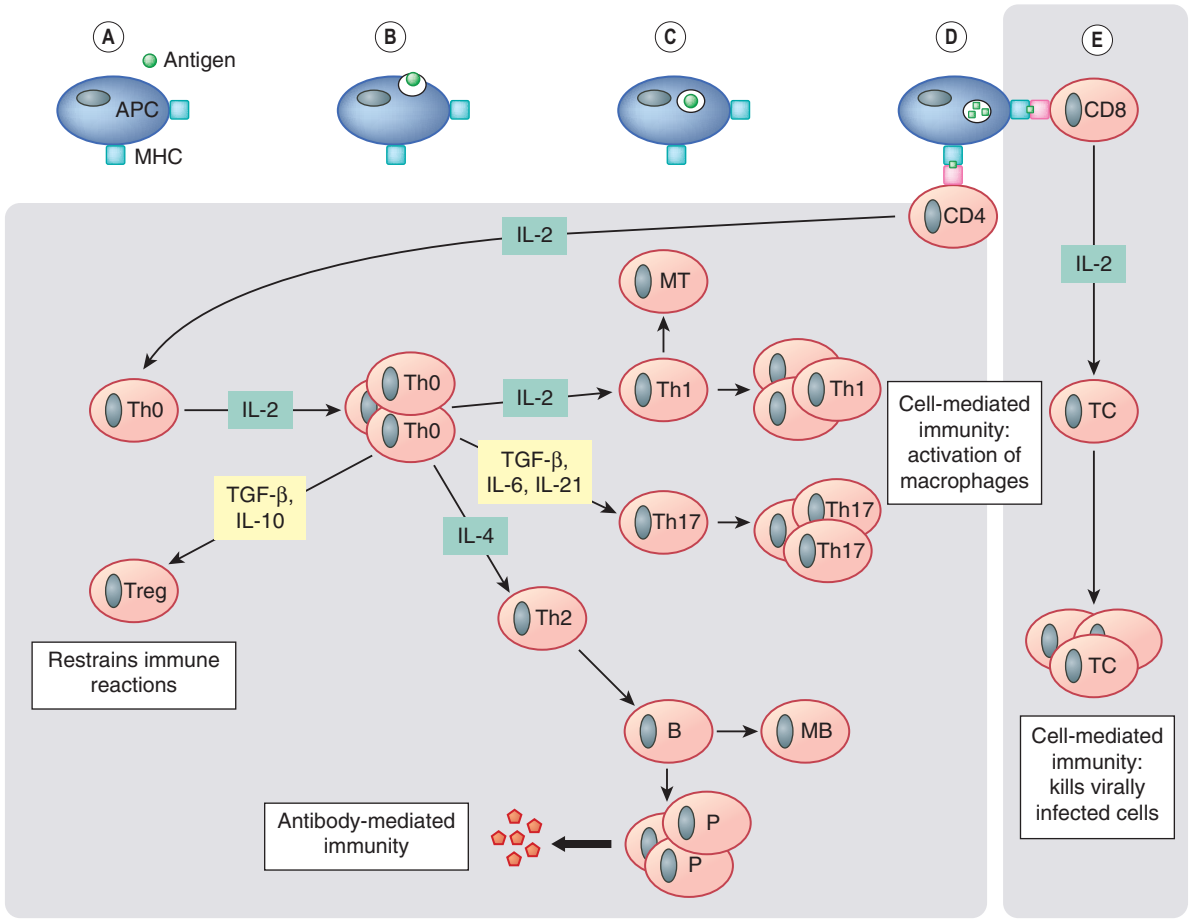

Fig. 6.3 Simplified diagram of the induction and effector phases of lymphocyte activation. Antigen-presenting cells (APCs) ingest and process antigen (A–D) and present fragments to naive, uncommitted CD4 T cells in conjunction with major histocompatibility complex (MHC) class II molecules, or to naive CD8 T cells in conjunction with MHC class I molecules, thus 'arming' them. The armed $CD4^+$ T cells synthesise and express interleukin (IL)-2 receptors and release this cytokine, which stimulates the cells by autocrine action, causing generation and proliferation of T-helper zero (Th0) cells. Autocrine cytokines (e.g. IL-4) cause proliferation of some Th0 cells to give Th2 cells, which are responsible for the development of antibody-mediated immune responses. These Th2 cells cooperate with and activate B cells to proliferate and give rise eventually to memory B cells (MB) and plasma cells (P), which secrete antibodies. Other autocrine cytokines (e.g. IL-2) cause proliferation of Th0 cells to give Th1, Th17 or Treg cells. Th1 and Th17 cells secrete cytokines that activate macrophages (responsible for some cell-mediated immune reactions). Treg cells restrain and inhibit the development of the immune response, thus preventing autoimmunity and excessive immune activation.

The armed $CD8^+$ T cells (E) also synthesise and express IL-2 receptors and release IL-2, which stimulates the cells by autocrine action to proliferate and give rise to cytotoxic T cells (TC). These can kill virally infected cells. IL-2 secreted by $CD4^+$ cells also plays a part in stimulating $CD8^+$ cells to proliferate. Note that the 'effector phase' depicted above relates to the 'protective' action of the immune response. When the response is inappropriately deployed—as in chronic inflammatory conditions such as rheumatoid arthritis—the Th1/Th17 component of the immune response is dominant and the activated macrophages release IL-1 and tumour necrosis factor (TNF)-α, which in turn trigger the release of the chemokines and inflammatory cytokines that play a major role in the pathology of the disease. MT and MB, memory T and B cells, respectively.

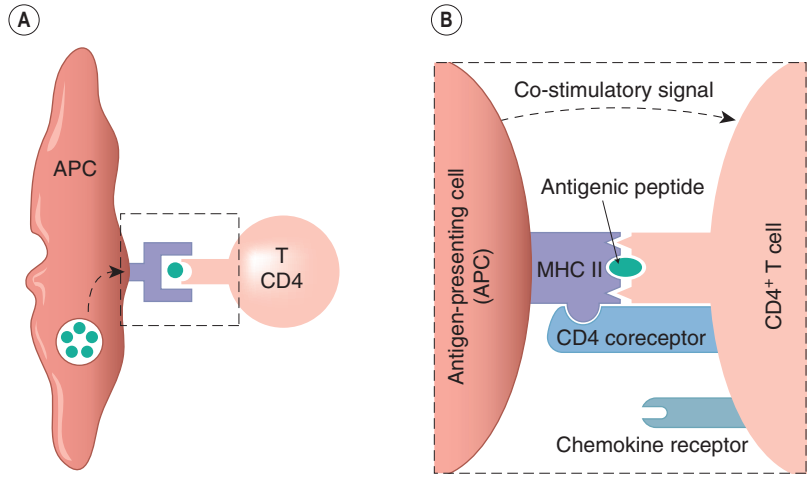

Fig. 6.4 The activation of a T cell by an antigen-presenting cell (APC). [**A**] The APC encounters a foreign protein and this is proteolytically processed into peptide fragments. The activation process then involves three stages. (i) Interaction between the complex of pathogen-derived antigen peptide fragments with major histocompatibility complex (MHC) class II and the antigen-specific receptor on the T cell. [**B**] (ii) Interaction between the CD4 co-receptor on the T cell and an MHC molecule on the APC. (iii) A co-stimulatory signal from the APC to the T cell. The CD4 co-receptor, together with a T-cell chemokine receptor, constitute the main binding sites for the HIV virus (see Fig. 51.3).

1. Uncommitted (naive) $CD4^+$ T-helper (Th) lymphocytes, or T-helper precursor (Thp) cells, in association with class II MHC molecules (see Fig. 6.4).
2. Naive $CD8^+$ T lymphocytes in association with class I MHC molecules.[6]

Activation of a T cell by an APC requires that several signals pass between the two cells at this 'immune synapse' (Fig. 6.4; see Medzhitov & Janeway, 2000). After activation, the T cells both generate IL-2 and acquire IL-2 receptors. Some potent anti-inflammatory drugs block this receptor thus preventing lymphocyte proliferation (see Ch. 26). IL-2 has an *autocrine*[7] action, stimulating proliferation and giving rise to a clone of T cells termed *Th0* cells, which, depending on the prevailing cytokine milieu, give rise to different subsets of armed helper cells. There are four major types of these 'helper cells', each of which generate a characteristic cytokine profile, possess a unique surface marker profile and have different roles in disease. These characteristics are summarised in Table 6.2.

Understanding the relationship between T-cell subsets, their respective cytokine profiles and pathological conditions is expected to highlight ways to manipulate the immune responses for disease prevention and treatment. There are already many experimental models in which modulation of the Th1/Th2 balance with recombinant cytokines or cytokine antagonists alters the outcome of the disease.

THE EFFECTOR PHASE

During the effector phase, the activated B and T lymphocytes differentiate either into *plasma cells* or into *memory cells*. The B plasma cells produce antibodies, which are effective in the extracellular fluid, but which cannot neutralise pathogens within cells. T-cell-mediated immune mechanisms overcome this problem by activating macrophages or directly killing virus-infected host cells. Antigen-sensitive *memory cells* are formed when the clone of lymphocytes that are programmed to respond to an antigen is greatly expanded after the first contact with the organism. They allow a greatly accelerated and more effective response to subsequent antigen exposure. In some cases, the response is so rapid and efficient that, after one exposure, the pathogen can never gain a foothold again. Immunisation procedures make use of this fact.

THE ANTIBODY-MEDIATED (HUMORAL) RESPONSE

There are five main classes of antibody—IgG, IgM, IgE, IgA and IgD—which differ from each other in certain structural respects. All are γ-globulins (immunoglobulins), which both recognise and interact specifically with antigens (i.e. proteins or polysaccharides foreign to the host), as well as activating one or more further components of the host's defence systems.

▼ An antibody is a Y-shaped protein molecule (see Ch. 59) in which the arms of the Y (the Fab portions) are the recognition sites for specific antigens, and the stem of the Y (the Fc portion) activates host defences. The B cells that are responsible for antibody production recognise foreign molecules by means of surface receptors that are essentially the immunoglobulin which that B-cell clone will eventually produce. Mammals harbour a vast number of B-cell clones that produce different antibodies with recognition sites for different antigens.

The induction of antibody-mediated responses varies with the type of antigen. With most antigens, a cooperative process between Th2 cells and B cells is necessary to produce a response. B cells can also present antigen to T cells that then release cytokines that act further on the B cell. The anti-inflammatory glucocorticoids (see Chs 26 and 32) and the immunosuppressive drug **ciclosporin** (see Ch. 26) affect the events at the stage of induction. The cytotoxic immunosuppressive drugs (see Ch. 26) inhibit the proliferation of both B and T cells. Eicosanoids may play a part in controlling these processes as prostaglandins of the E series can inhibit lymphocyte proliferation, probably by inhibiting the release of IL-2.

As you might guess, the ability to make antibodies has huge survival value; children born without this ability

[6]The main reason that it is difficult to transplant organs such as kidneys from one person to another is that their respective MHC molecules are different. Lymphocytes in the recipient will react to non-self (*allogeneic*) MHC molecules in the donor tissue, which is then likely to be rejected by a rapid and powerful immunological reaction.

[7]'Autocrine' signalling means that the mediator acts on the same cell that releases it. 'Paracrine' signalling means that the mediator acts on neighbouring cells.

Table 6.2 Lymphocyte subsets and their role in host defence and relationship to inflammatory disease

Lymphocyte subset	Cytokine stimulus	Main role in adaptive response	Main cytokines produced	Role in disease
Th0	IL-2	To act as a precursor cell type for further differentiation	—	—
Th1	IL-2	'Cell-mediated immunity' Cytokines released from these cells: activate macrophages to phagocytose and kill microorganisms and kill tumour cells; drive proliferation and maturation of the clone into *cytotoxic T cells* that kill virally infected host cells; reciprocally inhibit Th2 cell maturation	IFN-γ, IL-2 and TNF-α	Insulin-dependent diabetes mellitus (Ch. 30), multiple sclerosis, *Helicobacter pylori*-induced peptic ulcer (Ch. 29), aplastic anaemia (Ch. 25) and rheumatoid arthritis (Ch. 26) Allograft rejection
Th2	IL-4	'Humoral' immunity Cytokines released from these cells: stimulate B cells to proliferate and mature into plasma cells producing antibodies; enhance differentiation and activation of eosinophils and reciprocally inhibit Th1/Th17-cell functions. For this reason, they are often thought of as anti-inflammatory	IL-4, IL-5, TGF-β, IL-10 and IL-13	Asthma (Ch. 27) and allergy AIDS progression is associated with loss of Th1 cells and is facilitated by Th2 responses
Th17	TGF-β, IL-6 and IL-21	A specialised type of Th1 cell	IL-17	The response to infection, organ-specific immune responses and in the pathogenesis of diseases such as rheumatoid arthritis and multiple sclerosis
iTreg nTreg	IL-10 and TGF-β Matured in the thymus	Restraining the immune response, preventing auto-immunity and curtailing potentially damaging inflammatory responses	IL-10 and TGF-β	Failure of this mechanism can provoke excessive inflammation

IFN, interferon; IL, interleukin; iTreg, inducible Treg cells; nTreg, normal Treg cells; TGF, transforming growth factor; TNF, tumour necrosis factor.

suffer repeated infections such as pneumonia, skin infections and tonsillitis. Before the days of antibiotics, they died in early childhood, and even today they require regular replacement therapy with immunoglobulin. Apart from their ability to neutralise pathogens, antibodies can boost the effectiveness and specificity of the host's defence reaction in several ways.

Antibodies and complement

Formation of the antigen–antibody complex exposes a binding site for complement on the Fc domain. This activates the complement sequence and sets in train its attendant biological effects (see Fig. 6.1). This route to C3 activation (the *classic pathway*) provides an especially selective way of activating complement in response to a particular pathogen, because the antigen–antibody reaction that initiates it is not only a highly specific recognition event, but also occurs in close association with the pathogen. The lytic property of complement can be used therapeutically: monoclonal antibodies (mAbs) and complement together can be used to rid bone marrow of cancer cells as an adjunct to chemotherapy or radiotherapy (see Ch. 55).

Antibodies and the phagocytosis of bacteria

When antibodies are attached to their antigens on microorganisms by their Fab portions, the Fc domain is exposed. Phagocytic cells (neutrophils and macrophages) express surface receptors for these projecting Fc portions, which serve as a very specific link between microorganism and phagocyte.

Antibodies and cellular cytotoxicity

In some cases, for example with parasitic worms, the invader may be too large to be ingested by phagocytes. Antibody molecules can form a link between parasite and the host's white cells (in this case, eosinophils), which are then able to damage or kill the parasite by surface or extracellular actions. NK cells in conjunction with Fc receptors can also kill antibody-coated target cells (an example of antibody-dependent cell-mediated cytotoxicity).

Antibodies and mast cells or basophils

Mast cells and basophils have receptors for IgE, a particular form of antibody that can attach ('fix') to their cell membranes. When antigen reacts with this cell-fixed antibody, an entire panoply of pharmacologically active mediators is secreted. This very complex reaction is found widely throughout the animal kingdom and presumably offers clear survival value to the host. Having said that, its precise biological significance is not entirely clear, although it may be of importance in association with eosinophil activity as a defence against parasitic worms. When inappropriately triggered by substances not inherently damaging to the host, it is implicated in certain types of allergic reaction (see below) and apparently contributes more to illness than to survival in the modern world.

THE CELL-MEDIATED IMMUNE RESPONSE

Cytotoxic T cells (derived from $CD8^+$ cells) and inflammatory (cytokine-releasing) Th1 cells are attracted to inflammatory sites in a similar manner to neutrophils and macrophages, and are involved in cell-mediated responses (see Fig. 6.3).

Cytotoxic T cells

Armed cytotoxic T cells kill intracellular microorganisms such as viruses. When a virus infects a mammalian cell, there are two aspects to the resulting defensive response. The first step is the expression on the cell surface of peptides derived from the pathogen in association with MHC molecules. The second step is the recognition of the peptide–MHC complex by specific receptors on cytotoxic ($CD8^+$) T cells (Fig. 6.4 shows a similar process for a $CD4^+$ T cell). The cytotoxic T cells then destroy virus-infected cells by programming them to undergo apoptosis. Cooperation with macrophages may be required for killing to occur.

Macrophage-activating $CD4^+$ Th1 cells

Some pathogens (e.g. *Mycobacteria*, *Listeria*) survive and multiply within macrophages after ingestion. Armed $CD4^+$ Th1 cells release cytokines that activate macrophages to kill these intracellular pathogens. Th1 cells also recruit macrophages by releasing cytokines that act on vascular endothelial cells (e.g. TNF-α) and chemokines (e.g. *macrophage chemotactic factor-1; MCP-1*) that attract the macrophages to the sites of infection.

A complex of microorganism-derived peptides plus MHC molecules is expressed on the macrophage surface and is recognised by cytokine-releasing Th1 cells, which then generate cytokines that enable the macrophage to deploy its killing mechanisms. Activated macrophages (with or without intracellular pathogens) are factories for the production of chemical mediators, and can generate and secrete not only many cytokines but also toxic oxygen metabolites and neutral proteases that kill extracellular organisms (e.g. *Pneumocystis carinii* and helminths), complement components, eicosanoids, NO, a fibroblast-stimulating factor, pyrogens and the 'tissue factor' that initiates the extrinsic pathway of the coagulation cascade (Ch. 24), as well as various other coagulation factors. It is primarily the cell-mediated reaction that is responsible for allograft rejection. Macrophages are also important in coordinating the repair processes that must occur for inflammation to 'resolve'.

The specific cell-mediated or humoral immunological response is superimposed on the innate non-specific vascular and cellular reactions described previously, making them not only markedly more effective but much more selective for particular pathogens.

The general events of the inflammatory and hypersensitivity reactions specified above vary in some tissues. For example, in the airway inflammation of asthma, eosinophils and neuropeptides play a particularly significant role (see Ch. 27). In CNS inflammation, there is less neutrophil infiltration and monocyte influx is delayed, possibly because of lack of adhesion molecule expression on CNS vascular endothelium and deficient generation of chemotaxins. It has long been known that some tissues—the CNS parenchyma, the anterior chamber of the eye, and the testis—are *immunologically privileged* sites, in that a foreign antigen introduced directly does not provoke an immune reaction (which could be very disadvantageous to the host). However, introduction elsewhere of an antigen already in the CNS parenchyma will trigger the development of immune/inflammatory responses in the CNS.

SYSTEMIC RESPONSES IN INFLAMMATION

In addition to the local changes in an inflammatory area, there are often general systemic manifestations of inflammatory disease, including fever, an increase in blood leukocytes termed *leukocytosis* (or *neutrophilia* if the increase is in the neutrophils only) and the release from the liver of *acute-phase proteins*. These include C-reactive protein, α_2-macroglobulin, fibrinogen, α_1-antitrypsin and some complement components. While the function of many of these components is still a matter of conjecture, they all seem to have antimicrobial actions. C-reactive protein, for example, binds to some microorganisms, and the resulting complex activates complement. Other proteins scavenge iron (an essential nutrient for invading organisms) or block proteases, perhaps protecting the host against the worst excesses of the inflammatory response.

THE ROLE OF THE NERVOUS SYSTEM IN INFLAMMATION

It has become clear in recent years that the central, autonomic and peripheral nervous systems all play an important part in the regulation of the inflammatory response. This occurs at various levels:

- *The neuroendocrine system.* Adrenocorticotrophic hormone (ACTH), released from the anterior pituitary gland in response to endogenous circadian rhythm or to stress, releases cortisol from the adrenal glands. This hormone plays a crucial role in regulating immune function at all levels, hence the use of glucocorticoid drugs in the treatment of inflammatory disease. This topic is explored fully in Chs 26 and 32.
- *The central nervous sytem.* Surprisingly, cytokines such as IL-1 can signal the development of an inflammatory response directly to the brain through receptors on the vagus nerve. This may elicit an 'inflammatory reflex' and trigger activation of a cholinergic anti-inflammatory pathway. This is a relatively under-researched area: see Tracey (2002) and Sternberg (2006) for interesting discussions of this topic.
- *The autonomic nervous system.* Both the sympathetic and parasympathetic systems can influence the development of the inflammatory response. Generally

speaking, their influence is anti-inflammatory. Receptors for noradrenaline and acetylcholine are found on macrophages and many other cells involved in the immune response although it is not always entirely clear exactly where their ligands originate.

- *Peripheral sensory neurons.* Some sensory neurons release inflammatory neuropeptides when appropriately stimulated. These neurons are fine afferents (capsaicin-sensitive C and Aδ fibres; see Ch. 41) with specific receptors at their peripheral terminals. Kinins, 5-hydroxytryptamine and other chemical mediators generated during inflammation act on these receptors, stimulating the release of neuropeptides such as the tachykinins (neurokinin A, substance P) and calcitonin gene-related peptide (CGRP) which have proinflammatory or algesic actions. The neuropeptides are considered further in Chapter 19.

UNWANTED INFLAMMATORY AND IMMUNE RESPONSES

The immune response has to strike a delicate balance. According to one school of thought, an infection-proof immune system would be a possibility but would come at a serious cost to the host. With approximately 1 trillion potential antigenic sites in the host, such a 'superimmune' system would be some 1000 times more likely to attack the host itself, triggering *autoimmune disease*. In addition, it is not uncommon to find that innocuous substances such as pollen or peanuts sometimes inadvertently activate the immune system. When this happens, the inflammation itself inflicts damage and may be responsible for the major symptoms of the disease—either acutely as in (for example) anaphylaxis, or chronically in (for example) asthma or rheumatoid arthritis. In either case, anti-inflammatory or immunosuppressive therapy may be required.

▼ Unwanted immune responses, termed *allergic* or *hypersensitivity* reactions, have been classified into four types (Janeway et al., 2004).

Type I: immediate or anaphylactic hypersensitivity

▼ *Type I hypersensitivity* (often known simply as 'allergy') occurs in individuals who predominantly exhibit a Th2 rather than a Th1 response to antigen. In these individuals, substances that are not inherently noxious (such as grass pollen, house dust mites, certain foodstuffs or drugs, animal fur and so on) provoke the production of antibodies of the IgE type.[8] These fix on mast cells, in the lung, and also to eosinophils. Subsequent contact with the substance causes the release of histamine, PAF, eicosanoids and cytokines. The effects may be localised to the nose (hay fever), the bronchial tree (the initial phase of asthma), the skin (urticaria) or the gastrointestinal tract. In some cases, the reaction is more generalised and produces anaphylactic shock, which can be severe and life-threatening. Some important unwanted effects of drugs include anaphylactic hypersensitivity responses (see Ch. 57).

Type II: antibody-dependent cytotoxic hypersensitivity

▼ *Type II hypersensitivity* occurs when the mechanisms outlined above are directed against cells within the host that are (or appear to be) foreign. For example, host cells altered by drugs are sometimes mistaken by the immune system for foreign proteins and evoke antibody formation. The antigen–antibody reaction triggers complement activation (and its sequelae) and may promote attack by NK cells. Examples include alteration by drugs of neutrophils, leading to *agranulocytosis* (see Ch. 56), or of platelets, leading to *thrombocytopenic purpura* (Ch. 24). These type II reactions are also implicated in some types of *autoimmune thyroiditis* (e.g. *Hashimoto's disease*; see Ch. 33).

Type III: complex-mediated hypersensitivity

▼ *Type III hypersensitivity* occurs when antibodies react with *soluble* antigens. The antigen–antibody complexes can activate complement or attach to mast cells and stimulate the release of mediators.

An experimental example of this is the Arthus reaction that occurs if a foreign protein is injected subcutaneously into a rabbit or guinea pig with high circulating concentrations of antibody. Within 3–8 h, the area becomes red and swollen because the antigen–antibody complexes precipitate in small blood vessels and activate complement. Neutrophils are attracted and activated (by C5a) to generate toxic oxygen species and to secrete enzymes.

Mast cells are also stimulated by C3a to release mediators. Damage caused by this process is involved in *serum sickness*, caused when antigen persists in the blood after sensitisation, causing a severe reaction, as in the response to mouldy hay (known as *farmer's lung*), and in certain types of autoimmune kidney and arterial disease. Type III hypersensitivity is also implicated in *lupus erythematosus* (a chronic, autoimmune inflammatory disease).

Type IV: cell-mediated hypersensitivity

▼ The prototype of *type IV hypersensitivity* (also known as delayed hypersensitivity) is the *tuberculin reaction*, a local inflammatory response seen when proteins derived from cultures of the tubercle bacillus are injected into the skin of a person who has been sensitised by a previous infection or immunisation. An 'inappropriate' cell-mediated immune response is stimulated, accompanied by infiltration of mononuclear cells and the release of various cytokines. Cell-mediated hypersensitivity is also the basis of the reaction seen in some other infections (e.g. mumps and measles), as well as with mosquito and tick bites. It is also important in the skin reactions to drugs or industrial chemicals (see Ch. 57), where the chemical (termed a *hapten*) combines with proteins in the skin to form the 'foreign' substance that evokes the cell-mediated immune response (Fig. 6.3).

In essence, inappropriately deployed T-cell activity underlies all types of hypersensitivity, initiating types I, II and III, and being involved in both the initiation and the effector phase in type IV. These reactions are the basis of the clinically important group of autoimmune diseases. Immunosuppressive drugs (Ch. 26) and/or glucocorticoids (Ch. 32) are routinely employed to treat such disorders.

THE OUTCOME OF THE INFLAMMATORY RESPONSE

It is important not to lose sight of the fact that the inflammatory response is a defence mechanism and not, ipso facto, a disease. Its role is to restore normal structure and function to the infected or damaged tissue and, in the vast majority of cases, this is what happens. The healing and resolution phase of the inflammatory response is an active process and does not simply 'happen' in the absence of further inflammation. This is an area that we are just beginning to understand, but it is clear that it utilises its own unique palette of mediators and cytokines (including various growth factors, annexin-A1, lipoxins and IL-10; see Ch. 17) to terminate residual inflammation and to promote remodelling and repair of damaged tissue.

In some cases, healing will be complete, but if there has been damage (death of cells, pus formation, ulceration) repair is usually necessary and may result in scarring. If the pathogen persists, the acute response is likely to transform into a chronic inflammatory response. This is a slow, smouldering reaction that can continue indefinitely,

[8]Such individuals are said to be 'atopic', from a Greek word meaning 'out of place'.

destroying tissue and promoting local proliferation of cells and connective tissue. The principal cell types found in areas of chronic inflammation are mononuclear cells and abnormal macrophage-derived cells. During healing or chronic inflammation, growth factors trigger angiogenesis and cause fibroblasts to lay down fibrous tissue. Infection by some microorganisms, such as syphilis, tuberculosis and leprosy, bears the characteristic hallmarks of chronic inflammation from the start. The cellular and mediator components of this type of inflammation are also seen in many, if not most, chronic autoimmune and hypersensitivity diseases, and are important targets for drug action.

REFERENCES AND FURTHER READING

The innate and adaptive responses

Abbas, A.K., Murphy, K.M., Sher, A., 1996. Functional diversity of helper lymphocytes. Nature 383, 787–793. (*Excellent review, helpful diagrams; commendable coverage of Th1 and Th2 cells and their respective cytokine subsets*)

Adams, D.H., Lloyd, A.R., 1997. Chemokines: leukocyte recruitment and activation cytokines. Lancet 349, 490–495. (*Commendable review*)

Delves, P.J., Roitt, I.M., 2000. The immune system. N. Engl. J. Med. 343, 37–49, 108–117. (*A good overview of the immune system – a minitextbook of major areas in immunology; colourful three-dimensional figures*)

Gabay, C., Kushner, I., 1999. Acute phase proteins and other systemic responses to inflammation. N. Engl. J. Med. 340, 448–454. (*Lists the acute-phase proteins and outlines the mechanisms controlling their synthesis and release*)

Kärre, K., Welsh, R.M., 1997. Viral decoy vetoes killer cell. Nature 386, 446–447.

Kay, A.B., 2001. Allergic diseases and their treatment. N. Engl. J. Med. 344, 30–37, 109–113. (*Covers atopy and Th2 cells, the role of Th2 cytokines in allergies, IgE, the main types of allergy and new therapeutic approaches*)

Mackay, C.R., Lanzavecchia, A., Sallusto, F., 1999. Chemoattractant receptors and immune responses. Immunologist 7, 112–118. (*Masterly short review covering the role of chemoattractants in orchestrating immune responses – both the innate reaction and the Th1 and Th2 responses*)

Medzhitov, R., 2001. Toll-like receptors and innate immunity. Nat. Rev. Immunol. 1, 135–145. (*Excellent review of the role of Toll-like receptors in (a) the detection of microbial infection, and (b) the activation of innate non-adaptive responses, which in turn lead to antigen-specific adaptive responses*)

Medzhitov, R., Janeway, C., 2000. Innate immunity. N. Engl. J. Med. 343, 338–344. (*Outstandingly clear coverage of the mechanisms involved in innate immunity and its significance for the adaptive immune response*)

Mills, K.H., 2008. Induction, function and regulation of IL-17-producing T cells. Eur. J. Immunol. 38, 2636–2649. (*This paper covers the biology of Th17 cells – a relatively recent addition to our understanding of Th biology. Accessible and has good diagrams*)

Murphy, P.M., 2001. Viral exploitation and subversion of the immune system through chemokine mimicry. Nat. Immunol. 2, 116–122. (*Excellent description of viral/immune system interaction*)

Panes, J., Perry, M., Granger, D.N., 1999. Leukocyte–endothelial cell adhesion: avenues for therapeutic intervention. Br. J. Pharmacol. 126, 537–550. (*Brief coverage of the principal cell adhesion molecules and factors affecting leukocyte–endothelial adhesion precedes consideration of potential therapeutic targets*)

Parkin, J., Cohen, B., 2001. An overview of the immune system. Lancet 357, 1777–1789. (*A competent, straightforward review covering the role of the immune system in recognising, repelling and eradicating pathogens and in reacting against molecules foreign to the body*)

Sternberg, E.M., 2006. Neural regulation of innate immunity: a coordinated nonspecific host response to pathogens. Nat. Rev. Immunol. 6, 318–328. (*This paper and the next are both excellent and easy-to-read reviews covering the role of the CNS in inflammation. Some good diagrams*)

Takeda, K., Akira, S., 2003. Toll receptors and pathogen resistance. Cell Microbiol. 5, 143–153. (*Useful review and easy to read. Also deals with Toll receptor signalling in some depth*)

Tracey, K.J., 2002. The inflammatory reflex. Nature 420, 853–859.

Vasselon, T., Detmers, P.A., 2002. Toll receptors: a central element in innate immune responses. Infect. Immun. 70, 1033–1041. (*Another comprehensive review on this important topic*)

Walker, C., Zuany-Amorini, C., 2001. New trends in immunotherapy to prevent atopic disease. Trends Pharmacol. Sci. 22, 84–91. (*Discusses potential therapies based on recent advances in the understanding of the immune mechanisms of atopy*)

Wills-Karp, M., Santeliz, J., Karp, C.L., 2001. The germless theory of allergic diseases. Nat. Rev. Immunol. 1, 69–75. (*Discusses the hypothesis that early childhood infections inhibit the tendency to develop allergic disease*)

Books

Dale, M.M., Foreman, J.C., Fan, T.-P. (Eds.), 1994. Textbook of immunopharmacology, third ed. Blackwell Scientific, Oxford. (*Excellent textbook written with second- and third-year medical and science students in mind; contains many sections relevant to this chapter and the next*)

Janeway, C.A., Travers, P., Nolan, A., et al., 2004. Immunobiology: the immune system in health and disease, sixth ed. Churchill Livingstone, Edinburgh. (*Excellent textbook, good diagrams*)

Roitt, I., Brostoff, J., Male, D., 1998. Immunology, ninth ed. Blackwell Science, Oxford. (*Excellent textbook; well illustrated*)

Useful web links

http://www.biochemweb.org/fenteany/research/cell_migration/movement_movies.html. (*If you have never seen a neutrophil in hot pursuit of a bacterium, then you definitely need to look at this online movie. Great fun and highly instructive*)

Method and measurement in pharmacology

7

OVERVIEW

We emphasised in Chapters 2 and 3 that drugs, being molecules, produce their effects by interacting with other molecules. This interaction can lead to effects at all levels of biological organisation, from molecules to human populations (Fig. 7.1).[1]

Gaddum, a pioneering pharmacologist, commented in 1942: 'A branch of science comes of age when it becomes quantitative.' In this chapter, we cover the principles of metrication at the various organisational levels, ranging from laboratory methods to clinical trials. Assessment of drug action at the population level is the concern of *pharmacoepidemiology* and *pharmacoeconomics* (see Ch. 1), disciplines that are beyond the scope of this book.

We consider first the general principles of bioassay, and its extension to studies in human beings; we describe the development of animal models to bridge the predictive gap between animal physiology and human disease; we next discuss aspects of clinical trials used to evaluate therapeutic efficacy in a clinical setting; finally, we consider the principles of balancing benefit and risk. Experimental design and statistical analysis are central to the interpretation of all types of pharmacological data. Kirkwood & Sterne (2003) provide an excellent introduction.

BIOASSAY

Bioassay, defined as the estimation of the concentration or potency of a substance by measurement of the biological response that it produces, has played a key role in the development of pharmacology. Quantitation of drug effects by bioassay is necessary to compare the properties of different substances, or the same substance under different circumstances. It is used:

- to measure the pharmacological activity of new or chemically undefined substances
- to investigate the function of endogenous mediators
- to measure drug toxicity and unwanted effects.

▼ Bioassay plays a key role in the development of new drugs, discussed in Chapter 60.

The use of bioassay to measure the *concentration* of drugs and other active substances in the blood or other body fluids—once an important technology—has now been largely replaced by analytical chemistry techniques.

New hormones and other chemical mediators are often discovered by the biological effects that they produce. The first clue may be the finding that a tissue extract or some other biological sample produces an effect on an assay system. For example, the ability of extracts of the posterior lobe of the pituitary to produce a rise in blood pressure and a contraction of the uterus was observed at the beginning of the 20th century. Quantitative assay procedures based on these actions enabled a standard preparation of the extract to be established by international agreement in 1935. By use of these assays, it was shown that two distinct peptides—vasopressin and oxytocin—were responsible, and they were eventually identified and synthesised in 1953. Biological assay had already revealed much about the synthesis, storage and release of the hormones, and was essential for their purification and identification. Nowadays, it does not take 50 years of laborious bioassays to identify new hormones before they are chemically characterised,[2] but bioassay still plays a key role. The recent growth of *biopharmaceuticals* (see Ch. 59) as registered therapeutic agents has relied on bioassay techniques and the establishment of standard preparations. Biopharmaceuticals, whether derived from natural sources (e.g. monoclonal antibodies, vaccines) or by recombinant DNA technology (e.g. erythropoietin), tend to vary from batch to batch, and need to be standardised with respect to their biological activity. Varying glycosylation patterns, for example, which are not detected by immunoassay techniques, may affect biological activity.

BIOLOGICAL TEST SYSTEMS

Nowadays, an important use of bioassay is to provide information that will predict the effect of the drug in the clinical situation (where the aim is to improve function in patients suffering from the effects of disease). The choice of laboratory test systems (in vitro and in vivo 'models') that provide this predictive link is an important aspect of quantitative pharmacology. As our understanding of drug action at the molecular level advances (Ch. 3), this knowledge, and the technologies underlying it, have greatly extended the range of models that are available for measuring drug effects. By the 1960s, pharmacologists had become adept at using isolated organs and laboratory animals (usually under anaesthesia) for quantitative experiments, and had developed the principles of bioassay to allow reliable measurements to be made with these sometimes difficult and unpredictable test systems.

Bioassays on different test systems may be run in parallel to reveal the profile of activity of an unknown mediator. This was used to great effect by Vane and his colleagues, who studied the generation and destruction of endogenous active substances such as prostanoids (see Ch. 17) by the technique of *cascade superfusion* (Fig. 7.2). In this technique, the sample is run sequentially over a series of test preparations chosen to differentiate between different active constituents of the sample. The pattern of responses produced identifies the active material, and the use of such assay systems for 'on-line' analysis of biological samples has been invaluable in studying the production and fate of short-lived mediators such as prostanoids and nitric oxide (Ch. 20).

[1]Consider the effect of cocaine on organised crime, of organophosphate 'nerve gases' on the stability of dictatorships or of anaesthetics on the feasibility of surgical procedures for examples of molecular interactions that affect the behaviour of populations and societies.

[2]In 1988, a Japanese group (Yanagisawa et al., 1988) described in a single remarkable paper the bioassay, purification, chemical analysis, synthesis and DNA cloning of a new vascular peptide, *endothelin* (see Ch. 19).

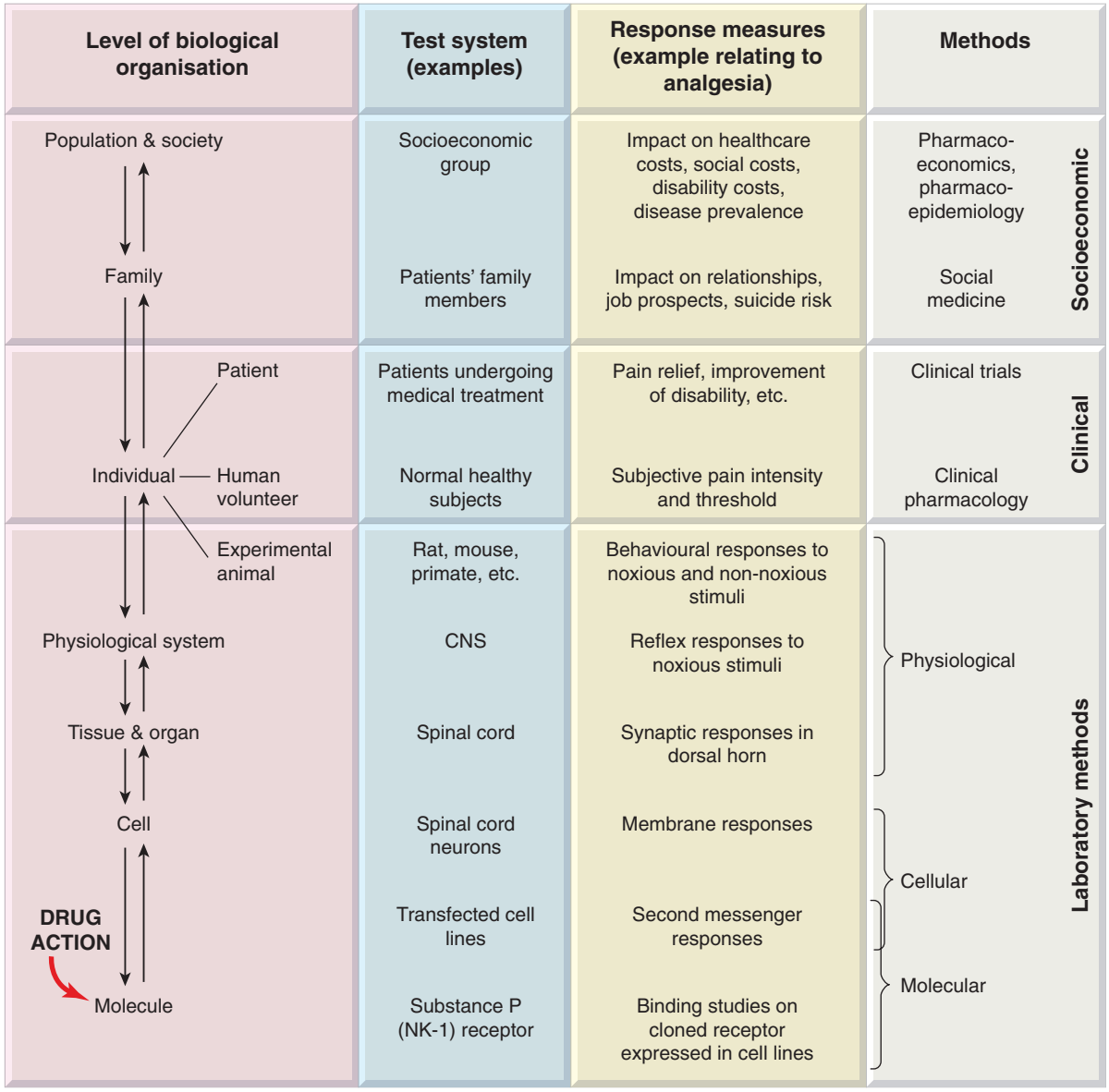

Fig. 7.1 Levels of biological organisation and types of pharmacological measurement.

These 'traditional' assay systems address drug action at the physiological level—roughly, the mid-range of the organisational hierarchy shown in Fig. 7.1. Subsequent developments have extended the range of available models in both directions, towards the molecular and towards the clinical. The introduction of binding assays (Ch. 3) in the 1970s was a significant step towards analysis at the molecular level. Subsequently, use of cell lines engineered to express specific human receptor subtypes has become widespread as a screening tool for drug discovery (see Ch. 60). Indeed, the range of techniques for analysing drug effects at the molecular and cellular levels is now very impressive. Bridging the gap between these levels and effects at the physiological and the therapeutic levels has, however, proved much more difficult, because human illness cannot, in many cases, be accurately reproduced in experimental animals. The use of transgenic animals to model human disease represents a real advance, and is discussed in more detail below.

GENERAL PRINCIPLES OF BIOASSAY

THE USE OF STANDARDS

J H Burn wrote in 1950: 'Pharmacologists today strain at the king's arm, but they swallow the frog, rat and mouse, not to mention the guinea pig and the pigeon.' He was referring to the fact that the 'king's arm' had been long since abandoned as a standard measure of length, whereas drug activity continued to be defined in terms of dose needed to cause, say, vomiting of a pigeon or cardiac arrest in a mouse. A plethora of 'pigeon units', 'mouse units' and the like, which no two laboratories could agree

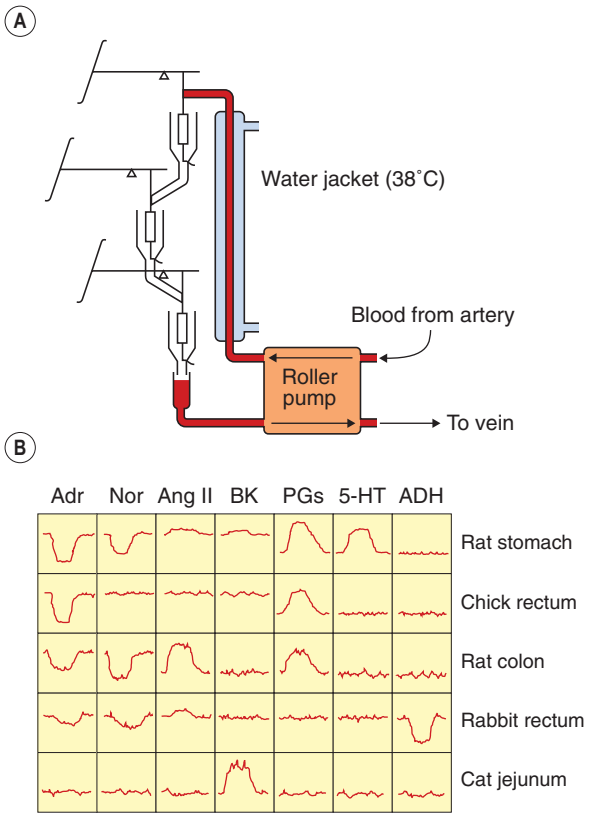

Fig. 7.2 Parallel assay by the cascade superfusion technique. Blood is pumped continuously from the test animal over a succession of test organs, whose responses are measured by a simple transducer system. The response of these organs to a variety of test substances (at 0.1–5 ng/ml) is shown. Each active substance produces a distinct pattern of responses, enabling unknown materials present in the blood to be identified and assayed. 5-HT, 5-hydroxytryptamine; ADH, antidiuretic hormone; Adr, adrenaline (epinephrine); Ang II, angiotensin II; BK, bradykinin; Nor, noradrenaline (norepinephrine); PG, prostaglandin. (From Vane J R 1969 Br J Pharmacol 35: 209–242.)

on, contaminated the literature.[3] Even if two laboratories cannot agree—because their pigeons differ—on the activity in pigeon units of the same sample of an active substance, they should nonetheless be able to agree that preparation X is, say, 3.5 times as active as standard preparation Y on the pigeon test. Biological assays are therefore designed to measure the *relative potency* of two preparations, usually a standard and an unknown. Maintaining stable preparations of various hormones, antisera and other biological materials, as reference standards, is the task of the UK National Board for Biological Standards Control.

[3]More picturesque examples of absolute units of the kind that Burn would have frowned on are the PHI and the mHelen. PHI, cited by Colquhoun (1971), stands for 'purity in heart index' and measures the ability of a virgin pure-in-heart to transform, under appropriate conditions, a he-goat into a youth of surpassing beauty. The mHelen is a unit of beauty, 1 mHelen being sufficient to launch 1 ship.

Bioassay

- Bioassay is the measurement of potency of a drug or unknown mediator from the magnitude of the biological effect that it produces.
- Bioassay normally involves comparison of the unknown preparation with a standard. Estimates that are not based on comparison with standards are liable to vary from laboratory to laboratory.
- Comparisons are best made on the basis of dose–response curves, which allow estimates of the equiactive concentrations of unknown and standard to be used as a basis for the potency comparison. Parallel line assays follow this principle.
- The biological response may be *quantal* (the proportion of tests in which a given all-or-nothing effect is produced) or *graded*. Different statistical procedures are appropriate in each case.
- Different approaches to metrication apply according to the level of biological organisation at which the drug effect needs to be measured. Approaches range through molecular and chemical techniques, in vitro and in vivo animal studies and clinical studies on volunteers and patients, to measurement of effects at the socioeconomic level.

THE DESIGN OF BIOASSAYS

▼ Given the aim of comparing the activity of two preparations, a standard (S) and an unknown (U) on a particular preparation, a bioassay must provide an estimate of the dose or concentration of U that will produce the same biological effect as that of a known dose or concentration of S. As Figure 7.3 shows, provided that the log dose–effect curves for S and U are parallel, the ratio, *M*, of equiactive doses will not depend on the magnitude of response chosen. Thus *M* provides an estimate of the potency ratio of the two preparations. A comparison of the magnitude of the effects produced by equal doses of S and U does not provide an estimate of *M* (see Fig. 7.3).

The main problem with all types of bioassay is that of biological variation, and the design of bioassays is aimed at:

- minimising variation
- avoiding systematic errors resulting from variation
- estimation of the limits of error of the assay result.

Commonly, comparisons are based on analysis of *dose–response curves*, from which the matching doses of S and U are calculated. The use of a logarithmic dose scale means that the curves for S and U will normally be parallel, and the potency ratio (*M*) is estimated from the horizontal distance between the two curves (Fig. 7.3). Assays of this type are known as *parallel line* assays, the minimal design being the 2 + 2 assay, in which two doses of standard (S_1 and S_2) and two of unknown (U_1 and U_2) are used. The doses are chosen to give responses lying on the linear part of the log dose–response curve, and are given repeatedly in randomised order, providing an inherent measure of the variability of the test system, which can be used, by means of straightforward statistical analysis, to estimate the confidence limits of the final result.

Problems arise if the two log dose–response curves are not parallel, for example if the assay is used to compare two drugs whose mechanism of action is not the same, or if one is a partial agonist (see Ch. 2). In this case it is not possible to define the relative potencies of S and U unambiguously in terms of a simple ratio and the experimenter must then face up to the fact that the comparison requires measurement of more than a single dimension of potency. An example of this

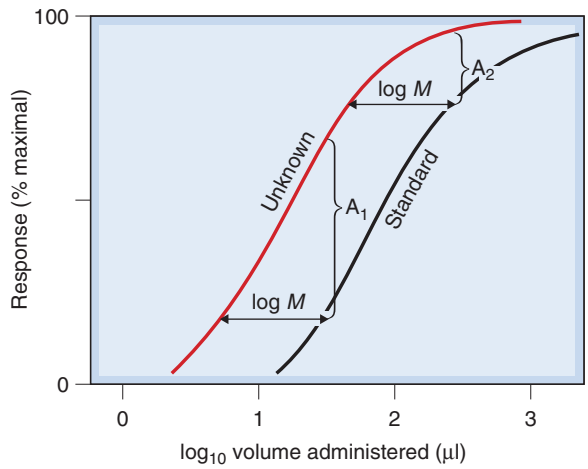

Fig. 7.3 Comparison of the potency of unknown and standard by bioassay. Note that comparing the magnitude of responses produced by the same dose (i.e. volume) of standard and unknown gives no quantitative estimate of their relative potency. (The differences, A_1 and A_2, depend on the dose chosen.) Comparison of equieffective doses of standard and unknown gives a valid measure of their relative potencies. Because the lines are parallel, the magnitude of the effect chosen for the comparison is immaterial; i.e. log *M* is the same at all points on the curves.

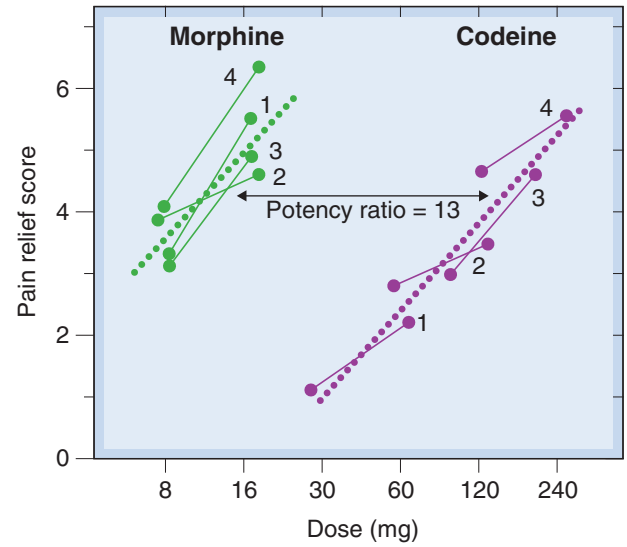

Fig. 7.4 Assay of morphine and codeine as analgesics in humans. Each of four patients (numbered 1–4) was given, on successive occasions in random order, four different treatments (high and low morphine, and high and low codeine) by intramuscular injection, and the subjective pain relief score calculated for each. The calculated regression lines gave a potency ratio estimate of 13 for the two drugs. (After Houde R W et al. 1965. In: Analgetics. Academic Press, New York.)

kind of difficulty is met when diuretic drugs (Ch. 28) are compared. Some ('low ceiling') diuretics are capable of producing only a small diuretic effect, no matter how much is given; others ('high ceiling') can produce a very intense diuresis (described as 'torrential' by authors with vivid imaginations). A comparison of two such drugs requires not only a measure of the doses needed to produce an equal low-level diuretic effect, but also a measure of the relative heights of the ceilings.

A simple example of an experiment to compare two analgesic drugs, morphine and codeine (see Ch. 41) in humans, based on a modified 2 + 2 design is shown in Figure 7.4. Each of the four doses was given on different occasions to each of the four subjects, the order being randomised and both subject and observer being unaware of the dose given. Subjective pain relief was assessed by a trained observer, and the results showed morphine to be 13 times as potent as codeine. This, of course, does not prove its superiority, but merely shows that a smaller dose is needed to produce the same effect. Such a measurement is, however, an essential preliminary to assessing the relative therapeutic merits of the two drugs, for any comparison of other factors, such as side effects, duration of action, tolerance or dependence, needs to be done on the basis of doses that are equiactive as analgesics.

ANIMAL MODELS OF DISEASE

There are many examples where simple intuitive models predict with fair accuracy therapeutic efficacy in humans. Ferrets vomit when placed in swaying cages, and drugs that prevent this are also found to relieve motion sickness and other types of nausea in humans. Irritant chemicals injected into rats' paws cause them to become swollen and tender, and this model predicts very well the efficacy of drugs used for symptomatic relief in inflammatory conditions such as rheumatoid arthritis in humans. As discussed elsewhere in this book, models for many important disorders, such as epilepsy, diabetes, hypertension and gastric ulceration, based on knowledge of the physiology of the condition, are available, and have been used successfully to produce new drugs, even though their success in predicting therapeutic efficacy is far from perfect.[4]

Ideally, an animal model should resemble the human disease in the following ways:

1. similar pathophysiological phenotype (*face validity*)
2. similar causation (*construct validity*)
3. similar response to treatment (*predictive validity*).

In practice, there are many difficulties, and the shortcomings of animal models are one of the main roadblocks on the route from basic medical science to improvements in therapy. The difficulties include the following.

- Many diseases, particularly in psychiatry, are defined by phenomena in humans that are difficult or impossible to observe in animals, which rules out face validity. As far as we know, mania or delusions have no counterpart in rats, nor can we recognise in them anything resembling a migraine attack or autism. Pathophysiological similarity is also inapplicable to conditions such as depression or anxiety disorders, where no clear brain pathology has been defined.
- The 'cause' of many human diseases is complex or unknown. To achieve construct validity for many degenerative diseases (e.g. Alzheimer's disease, osteoarthritis, Parkinson's disease), we need to model

[4]There have been many examples of drugs that were highly effective in experimental animals (e.g. in reducing brain damage following cerebral ischaemia) but ineffective in humans (stroke victims). Similarly, substance P antagonists (Ch. 19) are effective in animal tests for analgesia, but they proved inactive when tested in humans. How many errors in the opposite direction may have occurred we shall never know, because such drugs will not have been tested in humans.

Animal models

- Animal models of disease are important for investigating pathogenesis and for the discovery of new therapeutic agents. Animal models generally reproduce imperfectly only certain aspects of human disease states. Models of psychiatric illness are particularly problematic.
- Transgenic animals are produced by introducing mutations into the germ cells of animals (usually mice), which allow new genes to be introduced ('knock-ins') or existing genes to be inactivated ('knockouts') or mutated in a stable strain of animals.
- Transgenic animals are widely used to develop disease models for drug testing. Many such models are now available.
- The induced mutation operates throughout the development and lifetime of the animal, and may be lethal. New techniques of conditional mutagenesis allow the abnormal gene to be switched on or off at a chosen time.

the upstream (causative) factors rather than the downstream (symptomatic) features of the disease, although the latter are the basis of most of the simple physiological models used hitherto. The inflammatory pain model mentioned above lacks construct validity for rheumatoid arthritis, which is an autoimmune disease.

- Relying on response to treatment as a test of predictive validity carries the risk that drugs acting by novel mechanisms could be missed, because the model will have been selected on the basis of its responsiveness to known drugs. With schizophrenia (Ch. 45), for example, it is clear that dopamine antagonists are effective, and many of the models used are designed to assess dopamine antagonism in the brain, rather than other potential mechanisms that need to be targeted if drug discovery is to move on.

GENETIC AND TRANSGENIC ANIMAL MODELS

Nowadays, genetic approaches are increasingly used as an adjunct to conventional physiological and pharmacological approaches to disease modelling.

By selective breeding, it is possible to obtain pure animal strains with characteristics closely resembling certain human diseases. Genetic models of this kind include spontaneously hypertensive rats, genetically obese mice, epilepsy-prone dogs and mice, rats with deficient vasopressin secretion, and many other examples. In many cases, the genes responsible have not been identified.

▼ The obese mouse, which arose from a spontaneous mutation in a mouse-breeding facility, is one of the most widely used models for the study of obesity and type 2 diabetes (see Ch. 30). The phenotype results from inactivation of the *leptin* gene, and shows good face validity (high food intake, gross obesity, impaired blood glucose regulation, vascular complications—features characteristic of human obesity) and good predictive validity (responding to pharmacological intervention similarly to humans), but poor construct validity, since obese humans are not leptin deficient.

Deliberate genetic manipulation of the germline to generate *transgenic animals* (see Rudolph & Moehler, 1999; Offermanns & Hein, 2004) is of growing importance as a means of replicating human disease states in experimental animals, and thereby providing animal models that are expected to be more predictive of therapeutic drug effects in humans. This versatile technology, first reported in 1980, can be used in many different ways, for example:

- to inactivate individual genes, or mutate them to pathological forms
- to introduce new (e.g. human) genes
- to overexpress genes by inserting additional copies
- to allow gene expression to be controlled by the experimenter.[5]

Currently, most transgenic technologies are applicable in mice but much more difficult in other mammals. Other vertebrates (e.g. zebrafish) and invertebrates (*Drosophila*, *Caenorhabditis elegans*) are increasingly used for drug screening purposes.

Examples of such models include transgenic mice that overexpress mutated forms of the *amyloid precursor protein* or *presenilins*, which are important in the pathogenesis of Alzheimer's disease (see Ch. 39). When they are a few months old, these mice develop pathological lesions and cognitive changes resembling Alzheimer's disease, and provide very useful models with which to test possible new therapeutic approaches to the disease. Another neurodegenerative condition, Parkinson's disease (Ch. 39) has been modelled in transgenic mice that overexpress *synuclein*, a protein found in the brain inclusions that are characteristic of the disease. Transgenic mice with mutations in tumour suppressor genes and oncogenes (see Ch. 5) are widely used as models for human cancers. Mice in which the gene for a particular adenosine receptor subtype has been inactivated show distinct behavioural and cardiovascular abnormalities, such as increased aggression, reduced response to noxious stimuli and raised blood pressure. These findings serve to pinpoint the physiological role of this receptor, whose function was hitherto unknown, and to suggest new ways in which agonists or antagonists for these receptors might be developed for therapeutic use (e.g. to reduce aggressive behaviour or to treat hypertension). Transgenic mice can, however, be misleading in relation to human disease. For example, the gene defect responsible for causing cystic fibrosis (a disease affecting mainly the lungs in humans), when reproduced in mice, causes a disorder that mainly affects the intestine.

PHARMACOLOGICAL STUDIES IN HUMANS

Studies involving human subjects range from experimental pharmacodynamic or pharmacokinetic investigations to formal clinical trials. Non-invasive recording methods, such as *functional magnetic resonance imaging* to measure

[5]With conventional transgenic technology, the genetic abnormality is expressed throughout development, sometimes proving lethal or causing major developmental abnormalities. *Conditional transgenesis* is now possible (see Ristevski, 2005), allowing the transgene to remain silent until triggered by the administration of a chemical promoter (e.g. the tetracycline analogue, **doxycycline**, in the most widely used *Cre-Lox* conditional system). This avoids the complications of developmental effects and long-term adaptations, and may allow adult disease to be modelled more accurately.

regional blood flow in the brain (a surrogate for neuronal activity) and *ultrasonography* to measure cardiac performance, have greatly extended the range of what is possible. The scientific principles underlying experimental work in humans, designed, for example, to check whether mechanisms that operate in other species also apply to humans, or to take advantage of the much broader response capabilities of a person compared with a rat, are the same as for animals, but the ethical and safety issues are paramount, and ethical committees associated with all medical research centres tightly control the type of experiment that can be done, weighing up not only safety and ethical issues, but also the scientific importance of the proposed study. At the other end of the spectrum of experimentation on humans are formal *clinical trials*, often involving thousands of patients, aimed at answering specific questions regarding the efficacy and safety of new drugs.

CLINICAL TRIALS

Clinical trials are an important and highly specialised form of biological assay, designed specifically to measure therapeutic efficacy. The need to use patients undergoing treatment for experimental purposes raises serious ethical considerations, and imposes many restrictions. Here, we discuss some of the basic principles involved in clinical trials; the role of such trials in the course of drug development is described in Chapter 60.

A clinical trial is a method for comparing objectively, by a prospective study, the results of two or more therapeutic procedures. For new drugs, this is carried out during phase III of clinical development (Ch. 60). It is important to realise that, until about 50 years ago, methods of treatment were chosen on the basis of clinical impression and personal experience rather than objective testing.[6] Although many drugs, with undoubted effectiveness, remain in use without ever having been subjected to a controlled clinical trial, any new drug is now required to have been tested in this way before being licensed for general clinical use.[7]

On the other hand, **digitalis** (see Ch. 21) was used for 200 years to treat cardiac failure before a controlled trial showed it to be of very limited value except in a particular type of patient.

A good account of the principles and organisation of clinical trials is given by Friedman et al. (1996). A clinical trial aims to compare the response of a test group of patients receiving a new treatment (A) with that of a control group receiving an existing 'standard' treatment (B). Treatment A might be a new drug or a new combination of existing drugs, or any other kind of therapeutic intervention, such as a surgical operation, a diet, physiotherapy and so on. The standard against which it is judged (treatment B) might be a currently used drug treatment or (if there is no currently available effective treatment) a placebo or no treatment at all.

The use of controls is crucial in clinical trials. Claims of therapeutic efficacy based on reports that, for example, 16 out of 20 patients receiving drug X got better within 2 weeks are of no value without a knowledge of how 20 patients receiving no treatment, or a different treatment, would have fared. Usually, the controls are provided by a separate group of patients from those receiving the test treatment, but sometimes a crossover design is possible in which the same patients are switched from test to control treatment or vice versa, and the results compared. Randomisation is essential to avoid bias in assigning individual patients to test or control groups. Hence, the *randomised controlled clinical trial* is now regarded as the essential tool for assessing clinical efficacy of new drugs.

Concern inevitably arises over the ethics of assigning patients at random to particular treatment groups (or to no treatment). However, the reason for setting up a trial is that doubt exists whether the test treatment offers greater benefit than the control treatment. All would agree on the principle of informed consent,[8] whereby each patient must be told the nature and risks of the trial, and agree to participate on the basis that he or she will be randomly and unknowingly assigned to either the test or the control group.

Unlike the kind of bioassay discussed earlier, the clinical trial does not normally give any information about potency or the form of the dose–response curve, but merely compares the response produced by two stipulated therapeutic regimens. *Survival curves* provide one commonly used measure. Figure 7.5 shows rates of disease-free survival in two groups of breast cancer patients treated with conventional chemotherapy with and without the addition of **paclitaxel** (see Ch. 55). The divergence of the curves shows that paclitaxel significantly improved the clinical response. Additional questions may be posed, such as the prevalence and severity of side effects, or whether the treatment works better or worse in particular classes of patient, but only at the expense of added complexity and numbers of patients, and most trials are kept as simple as possible. The investigator must decide in advance what dose to use and how often to give it, and the trial will reveal only whether the chosen regimen performed better or worse than the control treatment. It will not say whether increasing or decreasing the dose would have improved the response; another trial would be needed to ascertain that. The basic question posed by a clinical trial is thus simpler than that addressed by most conventional bioassays. However, the

[6]Not exclusively. James Lind conducted a controlled trial in 1753 on 12 mariners, which showed that oranges and lemons offered protection against scurvy. However, 40 years passed before the British Navy acted on his advice, and a further century before the US Navy did.

[7]It is fashionable in some quarters to argue that to require evidence of efficacy of therapeutic procedures in the form of a controlled trial runs counter to the doctrines of 'holistic' medicine. This is a fundamentally antiscientific view, for science advances only by generating predictions from hypotheses and by subjecting the predictions to experimental test. 'Alternative' medical procedures, such as homeopathy, aromatherapy, acupuncture or 'detox', have rarely been so tested, and where they have they generally lack efficacy. Standing up for the scientific approach is the *evidence-based medicine* movement (see Sackett et al., 1996), which sets out strict criteria for assessing therapeutic efficacy, based on randomised, controlled clinical trials, and urges scepticism about therapeutic doctrines whose efficacy has not been so demonstrated.

[8]Even this can be contentious, because patients who are unconscious, demented or mentally ill are unable to give such consent, yet no one would want to preclude trials that might offer improved therapies to these needy patients. Clinical trials in children are particularly problematic but are necessary if the treatment of childhood diseases is to be placed on the same evidence base as is judged appropriate for adults. There are many examples where experience has shown that children respond differently from adults, and there is now increasing pressure on pharmaceutical companies to perform trials in children, despite the difficulties of carrying out such studies. The same concerns apply to trials in elderly patients.

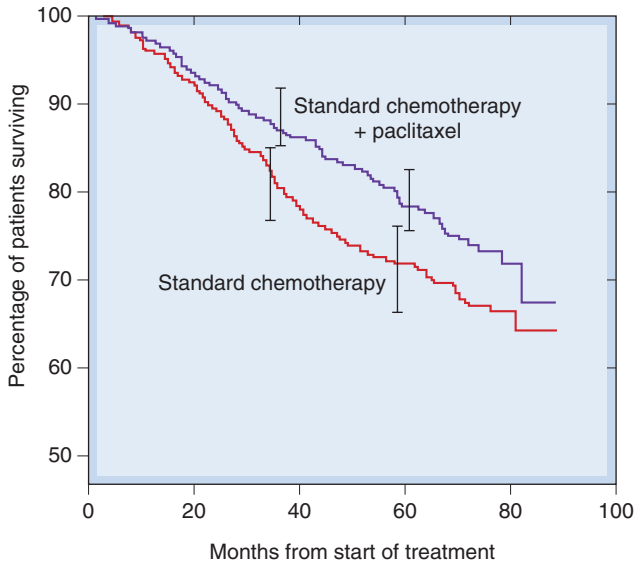

Fig. 7.5 Disease-free survival curves followed for 8 years in matched groups of breast cancer patients treated with a standard chemotherapy regime alone (629 patients), or with addition of paclitaxel (613 patients), showing a highly significant ($P = 0.006$) improvement with paclitaxel. Error bars represent 95% confidence intervals. (Redrawn from Martin et al. 2008 J Natl Cancer Inst 100:805–814.)

organisation of clinical trials, with controls against bias, is immeasurably more complicated, time-consuming and expensive than that of any laboratory-based assay.

AVOIDANCE OF BIAS

There are two main strategies that aim to minimise bias in clinical trials, namely:

1. randomisation
2. the double-blind technique.

If two treatments, A and B, are being compared on a series of selected patients, the simplest form of randomisation is to allocate each patient to A or B by reference to a series of random numbers. One difficulty with simple randomisation, particularly if the groups are small, is that the two groups may turn out to be ill-matched with respect to characteristics such as age, sex or disease severity. *Stratified randomisation* is often used to avoid the difficulty. Thus the subjects might be divided into age categories, random allocation to A or B being used within each category. It is possible to treat two or more characteristics of the trial population in this way, but the number of strata can quickly become large, and the process is self-defeating when the number of subjects in each becomes too small. As well as avoiding error resulting from imbalance of groups assigned to A and B, stratification can also allow more sophisticated conclusions to be reached. B might, for example, prove to be better than A in a particular group of patients even if it is not significantly better overall.

The double-blind technique, which means that neither subject nor investigator is aware at the time of the assessment which treatment is being used, is intended to minimise subjective bias. It has been repeatedly shown that, with the best will in the world, subjects and investigators both contribute to bias if they know which treatment is which, so the use of a double-blind technique is an important safeguard. It is not always possible, however. A dietary regimen or a surgical operation, for example, can seldom be disguised, and even with drugs, pharmacological effects may reveal to patients what they are taking and predispose them to report accordingly.[9] In general, however, the use of a double-blind procedure, with precautions if necessary to disguise such clues as the taste or appearance of the two drugs, is an important principle.[10]

THE SIZE OF THE SAMPLE

Both ethical and financial considerations dictate that the trial should involve the minimum number of subjects, and much statistical thought has gone into the problem of deciding in advance how many subjects will be required to produce a useful result. The results of a trial cannot, by their nature, be absolutely conclusive. This is because it is based on a sample of patients, and there is always a chance that the sample was atypical of the population from which it came. Two types of erroneous conclusion are possible, referred to as *type I* and *type II errors*. A type I error occurs if a difference is found between A and B when none actually exists (false positive). A type II error occurs if no difference is found although A and B do actually differ (false negative). A major factor that determines the size of sample needed is the degree of certainty the investigator seeks in avoiding either type of error. The probability of incurring a type I error is expressed as the *significance* of the result. To say that A and B are different at the $P < 0.05$ level of significance means that the probability of obtaining a false positive result (i.e. incurring a type I error) is less than 1 in 20. For most purposes, this level of significance is considered acceptable as a basis for drawing conclusions.

The probability of avoiding a type II error (i.e. failing to detect a real difference between A and B) is termed the *power* of the trial. We tend to regard type II errors more leniently than type I errors, and trials are often designed with a power of 0.8–0.9. To increase the significance and the power of a trial requires more patients. The second factor that determines the sample size required is the magnitude of difference between A and B that is regarded as clinically significant. For example, to detect that a given treatment reduces the mortality in a certain condition by at least 10 percentage points, say from 50% (in the control group) to 40% (in the treated group), would require 850 subjects, assuming that we wanted to achieve a $P < 0.05$ level of significance and a power of 0.9. If we were content only to reveal a reduction by 20 percentage points (and very likely miss a reduction by 10 points), only 210 subjects would be needed. In this example, missing a real 10-point

[9]The distinction between a true pharmacological response and a beneficial clinical effect produced by the knowledge (based on the pharmacological effects that the drug produces) that an active drug is being administered is not easy to draw, and we should not expect a mere clinical trial to resolve such a fine semantic issue.

[10]Maintaining the blind can be problematic. In an attempt to determine whether **melatonin** is effective in countering jet lag, a pharmacologist selected a group of fellow pharmacologists attending a congress in Australia, providing them with unlabelled capsules of melatonin or placebo, with a jet lag questionnaire to fill in when they arrived. Many of them (one of the authors included), with analytical resources easily to hand, opened the capsules and consigned them to the bin if they contained placebo. Pharmacologists are only human.

reduction in mortality could result in abandonment of a treatment that would save 100 lives for every 1000 patients treated—an extremely serious mistake from society's point of view. This simple example emphasises the need to assess clinical benefit (which is often difficult to quantify) in parallel with statistical considerations (which are fairly straightforward) in planning trials.

▼ A trial may give a significant result before the planned number of patients have been enrolled, so it is common for interim analyses to be carried out (by an independent team so that the trial team remains unaware of the results). If this analysis gives a conclusive result, or if it shows that continuation is unlikely to give a conclusive result, the trial can be terminated, thus reducing the number of subjects tested. In one such large-scale trial (Beta-blocker Heart Attack Trial Research Group, 1982) of the value of long-term treatment with the β-adrenoceptor-blocking drug **propranolol** (Ch. 14) following heart attacks, the interim results showed a significant reduction in mortality, which led to the early termination of the trial. In sequential trials, the results are computed case by case (each case being paired with a control) as the trial proceeds, and the trial stopped as soon as a result (at a predetermined level of significance) is achieved.

Various 'hybrid' trial designs, which have the advantage of sequential trials in minimising the number of patients needed but do not require strict pairing of subjects, have been devised (see Friedman et al., 1996).

Recently, the tendency has been to perform very large-scale trials, to allow several different treatment protocols, in various different patient groups to be compared. An example is the ALLHAT trial of various antihypertensive and lipid-lowering drugs to improve the outcome in cardiovascular disease (see Ch. 22). This ran from 1994 to 2002, cost US$130 million, and involved more than 42 000 patients in 623 treatment centres, with an army of coordinators and managers to keep it on track. One of its several far-reaching conclusions was that a cheap and familiar diuretic drug in use for more than 50 years was more effective than more recent and expensive antihypertensive drugs.[11]

CLINICAL OUTCOME MEASURES

The measurement of clinical outcome can be a complicated business, and is becoming increasingly so as society becomes more preoccupied with assessing the efficacy of therapeutic procedures in terms of improved quality of life, and societal and economic benefit, rather than in terms of objective clinical effects, such as lowering of blood pressure, improved airways conductance or increased life expectancy. Various scales for assessing 'health-related quality of life' have been devised and tested (see Walley & Haycocks, 1997), and the tendency is to combine these with measures of life expectancy to arrive at the measure 'quality-adjusted life years' (QALYs) as an overall measure of therapeutic efficacy, which attempts to combine both survival time and relief from suffering in assessing overall benefit.[12] In planning clinical trials, it is necessary to decide the purpose of the trial in advance, and to define the outcome measures accordingly.

FREQUENTIST AND BAYESIAN APPROACHES

▼ The conventional approach to analysis of scientific data (including clinical trials data) is known as 'frequentist' and is based on a *null hypothesis*, for example of the form: treatment A is no more effective than treatment B. Rejection of the hypothesis implies that A is more effective than B. Suppose that a trial shows, on average, that patients treated with A live longer than patients treated with B. Conventional frequentist statistics addresses the question: *If A were actually no more effective than B, what is the probability* (P) *of obtaining the results that were actually obtained in the trial?* In other words, given that treatment A is no better than B, how often, had we repeated the trial many times, would we have obtained results suggesting that A is better? If this probability is low (say, less than 0.05), we reject the null hypothesis and conclude that A is most likely better. If P is larger, the results could quite easily have been obtained without there being any true difference between A and B, and we cannot reject the null hypothesis.

If we have no prior reason for thinking that A will be better than B, the frequentist approach is perfectly appropriate, and it is the usual principle on which trials of unknown drugs are based. But often, in real life, there will be good reason, based on previous trials or clinical experience, to believe that A is actually better than B. Using a *Bayesian approach* allows this to be taken into account formally and explicitly by defining a *prior probability* for the effect of A. The data from the new trial, which can be smaller than a conventional trial, are then statistically superimposed on the prior probability curve to produce a *posterior probability* curve, in effect an update of the prior probability curve that takes account of the new data. The Bayesian approach is controversial, depending as it does on expressing the (often subjective) prior assumption in explicit mathematical terms, and the statistical analysis is complex. Nevertheless, it can be argued that to ignore altogether prior knowledge and experience when interpreting new data is unjustified, and even unethical, and the Bayesian approach is consequently gaining acceptance.

For an explanation of the principles underlying Bayesian approaches, which are being increasingly applied to clinical trials, see Spiegelhalter et al. (1999) and Lilford & Braunholtz (2000).

PLACEBOS

▼ A placebo is a dummy medicine containing no active ingredient (or alternatively, a dummy surgical procedure, diet or other kind of therapeutic intervention), which the patient believes is (or could be, in the context of a controlled trial) the real thing. The 'placebo response' is widely believed to be a powerful therapeutic effect,[13] producing a significant beneficial effect in about one-third of patients. While many clinical trials include a placebo group that shows improvement, few have compared this group directly with untreated controls. A survey of these trial results (Hróbjartsson & Grøtsche, 2001) concluded (controversially) that the placebo effect was generally insignificant, except in the case of pain relief, where it was small but significant. They concluded that the popular belief in the strength of the placebo effect is misplaced, and probably reflects in part the tendency of many symptoms to improve spontaneously and in part the reporting bias of patients who want to please their doctors. The ethical case for using placebos as therapy, which has been the subject of much public discussion, may therefore be weaker than has been argued. The risks of placebo therapies should not be underestimated. The use of active medicines may be delayed. The necessary element of deception risks undermining the confidence of patients in the integrity of doctors. A state of 'therapy dependence' may be produced in people who are not ill, because there is no way of assessing whether a patient still 'needs' the placebo.

[11]Though without much impact so far on prescribing habits, owing to the marketing muscle of pharmaceutical companies.

[12]As may be imagined, trading off duration and quality of life raises issues about which many of us feel decidedly squeamish. Not so economists, however. They approach the problem by asking such questions as: 'How many years of life would you be prepared to sacrifice in order to live the rest of your life free of the disability you are currently experiencing?' Or, even more disturbingly: 'If you could gamble on surviving free of disability for your normal lifespan, or (if you lose the gamble) dying immediately, what odds would you accept?' Imagine being asked this by your doctor. 'But I only wanted something for my sore throat,' you protest weakly.

[13]Its opposite, the *nocebo effect*, describes the adverse effects reported with dummy medicines.

META-ANALYSIS

▼ It is possible, by the use of statistical techniques, to combine the data obtained in several individual trials (provided each has been conducted according to a randomised design) in order to gain greater power and significance. This procedure, known as meta-analysis or overview analysis, can be very useful in arriving at a conclusion on the basis of several published trials, of which some claimed superiority of the test treatment over the control while others did not. As an objective procedure, it is certainly preferable to the 'take your pick' approach to conclusion forming adopted by most human beings when confronted with contradictory data. It has several drawbacks, however (see Naylor, 1997), the main one being 'publication bias', because negative studies are generally considered less interesting, and are therefore less likely to be published, than positive studies. Double counting, caused by the same data being incorporated into more than one trial report, is another problem.

The organisation of large-scale clinical trials involving hundreds or thousands of patients at many different centres is a massive and expensive undertaking that makes up one of the major costs of developing a new drug, and can easily go wrong.

An early large trial (Anturane Reinfarction Trial Research Group, 1978) involved 1620 patients at 26 research centres in the USA and Canada, 98 collaborating researchers and a formidable list of organising committees, including two independent audit committees to check that the work was being carried out in conformity with the strict protocols established. The conclusion was that the drug under test (**sulfinpyrazone**) reduced by almost one-half the mortality from repeat heart attacks in the 8-month period after a first attack, and could save many lives. The US Food and Drug Administration, however, refused to grant a licence for the use of the drug, criticising the trial as unreliable and biased in several respects. Their independent analysis of the data showed the beneficial effect of the drug to be slight and insignificant. Further analysis and further trials, however, supported the original conclusion, but by then the efficacy of aspirin in this condition had been established, so the use of sulfinpyrazone never found favour. Much larger trials are now regularly conducted, exemplified by the ALLHAT trial mentioned above (p. 96).

BALANCING BENEFIT AND RISK

THERAPEUTIC INDEX

▼ The concept of *therapeutic index* aims to provide a measure of the margin of safety of a drug, by drawing attention to the relationship between the effective and toxic doses:

$$\text{Therapeutic index} = LD_{50}/ED_{50}$$

where LD_{50} is the dose that is lethal in 50% of the population, and ED_{50} is the dose that is 'effective' in 50%. Obviously, it can only be measured in animals, and it is not a useful guide to the safety of a drug in clinical use for several reasons:

- LD_{50} does not reflect the incidence of adverse effects in the therapeutic setting.[14]
- ED_{50} depends on what measure of effectiveness is used. For example, the ED_{50} for aspirin used for a mild headache is much lower than for aspirin as an antirheumatic drug.
- Both efficacy and toxicity show variability between individuals, for various reasons (see Ch. 56). Such variability in the effective dose or the toxic dose of a drug makes it inherently less predictable, and therefore less safe, although this is not reflected in the therapeutic index.

[14]Ironically, thalidomide—probably the most harmful drug ever marketed—was promoted specifically on the basis of its exceptionally high therapeutic index (i.e. it killed rats only when given in extremely large doses).

Clinical trials

- A clinical trial is a special type of bioassay done to compare the clinical efficacy of a new drug or procedure with that of a known drug or procedure (or a placebo).
- Generally, the aim is a straight comparison of unknown (A) with standard (B) at a single dose level. The result may be: 'B better than A', 'B worse than A', or 'No difference detected'. Efficacy, not potency, is compared.
- To avoid bias, clinical trials should be:
  - *controlled* (comparison of A with B, rather than study of A alone)
  - *randomised* (assignment of subjects to A or B on a random basis)
  - *double-blind* (neither subject nor assessor knows whether A or B is being used).
- Type I errors (concluding that A is better than B when the difference is actually due to chance) and type II errors (concluding that A is not different from B because a real difference has escaped detection) can occur; the likelihood of either kind of error decreases as the sample size and number of end-point events is increased.
- Interim analysis of data, carried out by an independent group, may be used as a basis for terminating a trial prematurely if the data are already conclusive, or if a clear result is unlikely to be reached.
- All experiments on human subjects require approval by an independent ethical committee.
- Clinical trials require very careful planning and execution, and are inevitably expensive.
- Clinical outcome measures may comprise:
  - physiological measures (e.g. blood pressure, liver function tests, airways function)
  - subjective assessments (e.g. pain relief, mood)
  - long-term outcome (e.g. survival or freedom from recurrence)
  - overall '*quality of life*' measures
  - '*quality-adjusted life years*' (QALYs), which combine survival with quality of life.
- Meta-analysis is a statistical technique used to pool the data from several independent trials.

OTHER MEASURES OF BENEFIT AND RISK

▼ Alternative ways of quantifying the benefits and risks of drugs in clinical use have received much attention. One useful approach is to estimate from clinical trial data the proportion of test and control patients who will experience (a) a defined level of clinical benefit (e.g. survival beyond 2 years, pain relief to a certain predetermined level, slowing of cognitive decline by a given amount) and (b) adverse effects of defined degree. These estimates of proportions of patients showing beneficial or harmful reactions can be expressed as *number needed to treat* (NNT; i.e. the number of patients who need to be treated in order for one to show the given effect, whether beneficial or adverse). For example, in a recent study of pain relief by antidepressant drugs compared with placebo, the findings were: for benefit

(a defined level of pain relief), NNT = 3; for minor unwanted effects, NNT = 3; for major adverse effects, NNT = 22. Thus of 100 patients treated with the drug, on average 33 will experience pain relief, 33 will experience minor unwanted effects, and 4 or 5 will experience major adverse effects, information that is helpful in guiding therapeutic choices. One advantage of this type of analysis is that it can take into account the underlying disease severity in quantifying benefit. Thus if drug A halves the mortality of an often fatal disease (reducing it from 50% to 25%, say), the NNT to save one life is 4; if drug B halves the mortality of a rarely fatal disease (reducing it from 5% to 2.5%, say), the NNT to save one life is 40. Notwithstanding other considerations, drug A is judged to be more valuable than drug B, even though both reduce mortality by one-half. Furthermore, the clinician must realise that to save one life with drug B, 40 patients must be exposed to a risk of adverse effects, whereas only 4 are exposed for each life saved with drug A.

Determination of risk and benefit

- *Therapeutic index* (lethal dose for 50% of the population divided by effective dose for 50%) is unsatisfactory as a measure of drug safety because:
  - it is based on animal toxicity data, which may not reflect forms of toxicity or adverse reactions that are important clinically
  - it takes no account of idiosyncratic toxic reactions.
- More sophisticated measures of risk–benefit analysis for drugs in clinical use are available, and include the *number needed to treat* (NNT) principle.

REFERENCES AND FURTHER READING

General references

Colquhoun, D., 1971. Lectures on biostatistics. Oxford University Press, Oxford. (*Standard textbook*)

Kirkwood, B.R., Sterne, J.A.C., 2003. Medical statistics, second ed. Blackwell, Malden. (*Clear introductory textbook covering statistical principles and methods*)

Lilford, R.J., Braunholtz, D., 2000. Who's afraid of Thomas Bayes? J. Epidemiol. Community Health 54, 731–739. (*Explains the principles of Bayesian analysis in a non-mathematical way*)

Walley, T., Haycocks, A., 1997. Pharmacoeconomics: basic concepts and terminology. Br. J. Clin. Pharmacol. 43, 343–348. (*Useful introduction to analytical principles that are becoming increasingly important for therapeutic policy makers*)

Yanagisawa, M., Kurihara, H., Kimura, S., et al., 1988. A novel potent vasoconstrictor peptide produced by vascular endothelial cells. Nature 332, 411–415. (*The first paper describing endothelin—a remarkably full characterisation of an important new mediator*)

Animal models

Offermanns, S., Hein, L. (Eds.), 2004. Transgenic models in pharmacology. Handb. Exp. Pharmacol. 159. (*A comprehensive series of review articles describing transgenic mouse models used to study different pharmacological mechanisms and disease states*)

Ristevski, S., 2005. Making better transgenic models: conditional, temporal, and spatial approaches. Mol. Biotechnol. 29, 153–164. (*Description of methods for controlling transgene expression*)

Rudolph, U., Moehler, H., 1999. Genetically modified animals in pharmacological research: future trends. Eur. J. Pharmacol. 375, 327–337. (*Good review of uses of transgenic animals in pharmacological research, including application to disease models*)

Clinical trials

Anturane Reinfarction Trial Research Group, 1978. Sulfinpyrazone in the prevention of cardiac death after myocardial infarction. N. Engl. J. Med. 298, 289–295. (*Example of an early large-scale clinical trial*)

Beta-blocker Heart Attack Trial Research Group, 1982. A randomised trial of propranolol in patients with acute myocardial infarction. 1. Mortality results. JAMA 247, 1707–1714. (*A trial that was terminated early when clear evidence of benefit emerged*)

Friedman, L.M., Furberg, C.D., DeMets, D.L., 1996. Fundamentals of clinical trials, third ed. Mosby, St Louis. (*Standard textbook*)

Hróbjartsson, A., Gøtzsche, P.C., 2001. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. N. Engl. J. Med. 344, 1594–1601. (*An important meta-analysis of clinical trial data, which shows, contrary to common belief, that placebos in general have no significant effect on clinical outcome, except—to a small degree—in pain relief trials. Confirmed in an extended analysis: J. Int. Med. 2004, 256, 91–100*)

Naylor, C.D., 1997. Meta-analysis and the meta-epidemiology of clinical research. Br. Med. J. 315, 617–619. (*Thoughtful review on the strengths and weaknesses of meta-analysis*)

Sackett, D.L., Rosenburg, W.M.C., Muir-Gray, J.A., et al., 1996. Evidence-based medicine: what it is and what it isn't. Br. Med. J. 312, 71–72. (*Balanced account of the value of evidence-based medicine—an important recent trend in medical thinking*)

Spiegelhalter, D.J., Myles, J.P., Jones, D.R., Abrams, K.R., 1999. An introduction to Bayesian methods in health technology assessment. Br. Med. J. 319, 508–512. (*Short non-mathematical explanation of the Bayesian approach to data analysis*)

8 Drug absorption and distribution

OVERVIEW

The physical processes of diffusion, penetration of membranes, binding to plasma protein and partition into fat and other tissues underlie the absorption and distribution of drugs. These processes are described, followed by more specific coverage of the process of drug absorption and related practical issue of routes of drug administration, and of the distribution of drugs into different bodily compartments. There is a short final section on special drug delivery systems designed to deliver drugs efficiently and selectively to their sites of action.

INTRODUCTION

Drug disposition is divided into four stages designated by the acronym 'ADME':

- **A**bsorption from the site of administration
- **D**istribution within the body
- **M**etabolism
- **E**xcretion.

Absorption and distribution are considered here, together with routes of administration. Metabolism and excretion are covered in Chapter 9. We begin with a description of the physical processes that underlie drug disposition.

PHYSICAL PROCESSES UNDERLYING DRUG DISPOSITION

Drug molecules move around the body in two ways:

- bulk flow (i.e. in the bloodstream, lymphatics or cerebrospinal fluid)
- diffusion (i.e. molecule by molecule, over short distances).

The chemical nature of a drug makes no difference to its transfer by bulk flow. The cardiovascular system provides a rapid long-distance distribution system. In contrast, diffusional characteristics differ markedly between different drugs. In particular, ability to cross hydrophobic diffusion barriers is strongly influenced by lipid solubility. Aqueous diffusion is part of the overall mechanism of drug transport, because it is this process that delivers drug molecules to and from the non-aqueous barriers. The rate of diffusion of a substance depends mainly on its molecular size, the diffusion coefficient for small molecules being inversely proportional to the square root of molecular weight. Consequently, while large molecules diffuse more slowly than small ones, the variation with molecular weight is modest. Many drugs fall within the molecular weight range 200–1000 Da, and variations in aqueous diffusion rate have only a small effect on their overall pharmacokinetic behaviour. For most purposes, we can regard the body as a series of interconnected well-stirred compartments within each of which the drug concentration is uniform. It is movement between compartments, generally involving penetration of non-aqueous diffusion barriers, that determines where, and for how long, a drug will be present in the body after it has been administered. The analysis of drug movements with the help of a simple compartmental model is discussed in Chapter 9.

THE MOVEMENT OF DRUG MOLECULES ACROSS CELL BARRIERS

Cell membranes form the barriers between aqueous compartments in the body. A single layer of membrane separates the intracellular from the extracellular compartments. An epithelial barrier, such as the gastrointestinal mucosa or renal tubule, consists of a layer of cells tightly connected to each other so that molecules must traverse at least two cell membranes (inner and outer) to pass from one side to the other. Vascular endothelium is more complicated, its anatomical disposition and permeability varying from one tissue to another. Gaps between endothelial cells are packed with a loose matrix of proteins that act as filters, retaining large molecules and letting smaller ones through. The cut-off of molecular size is not exact: water permeates rapidly whereas molecules of 80 000–100 000 Da permeate very slowly. In some organs, especially the central nervous system (CNS) and the placenta, there are tight junctions between the cells, and the endothelium is encased in an impermeable layer of periendothelial cells (*pericytes*). These features prevent potentially harmful molecules from leaking from the blood into these organs and have major pharmacokinetic consequences for drug distribution.[1]

In other organs (e.g. the liver and spleen), endothelium is discontinuous, allowing free passage between cells. In the liver, hepatocytes form the barrier between intra- and extravascular compartments and take on several endothelial cell functions. Fenestrated endothelium occurs in endocrine glands, facilitating transfer to the bloodstream of hormones or other molecules through pores in the endothelium. Formation of fenestrated endothelium (*angiogenesis*) is controlled by a specific endocrine gland-derived vascular endothelial growth factor (dubbed EG-VEGF). Endothelial cells lining postcapillary venules have specialised functions relating to leukocyte migration and inflammation: the sophistication of the intercellular junction can be appreciated from the observation that leukocyte migration can occur without any detectable leak of water or small ions (see Ch. 16).

[1]This is illustrated by strain and species differences. For example, collie dogs lack the multidrug resistance gene (*mdr1*) and a P-glycoprotein that contributes importantly to the blood–brain barrier, with consequences for veterinary medicine because **ivermectin** (an anthelminthic drug, Ch. 54,) is consequently severely neurotoxic in the many breeds with collie ancestry (see Neff et al., 2004).

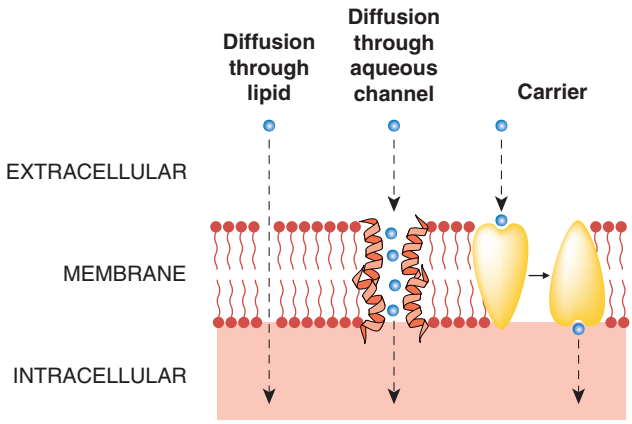

Fig. 8.1 Routes by which solutes can traverse cell membranes. (Molecules can also cross cellular barriers by pinocytosis.)

There are four main ways by which small molecules cross cell membranes (Fig. 8.1):

1. by diffusing directly through the lipid
2. by diffusing through aqueous pores formed by special proteins (*aquaporins*) that traverse the lipid
3. by combination with a *solute carrier* (SLC) or other membrane transporter
4. by *pinocytosis*.

Of these routes, diffusion through lipid and carrier-mediated transport are particularly important in relation to pharmacokinetic mechanisms. Diffusion through aquaporins (membrane glycoproteins that can be blocked by mercurial reagents such as *para*-chloromercurobenzene sulfonate) is probably important in the transfer of gases such as carbon dioxide, but the pores are too small in diameter (about 0.4 nm) to allow most drug molecules (which usually exceed 1 nm in diameter) to pass through. Consequently, drug distribution is not notably abnormal in patients with genetic diseases affecting aquaporins. Pinocytosis involves invagination of part of the cell membrane and the trapping within the cell of a small vesicle containing extracellular constituents. The vesicle contents can then be released within the cell, or extruded from its other side. This mechanism is important for the transport of some macromolecules (e.g. **insulin**, which crosses the blood–brain barrier by this process), but not for small molecules.

Diffusion through lipid and carrier-mediated transport will now be discussed in more detail.

DIFFUSION THROUGH LIPID

Non-polar molecules (in which electrons are uniformly distributed) dissolve freely in membrane lipids, and consequently diffuse readily across cell membranes. The number of molecules crossing the membrane per unit area in unit time is determined by the *permeability coefficient*, *P*, and the concentration difference across the membrane. Permeant molecules must be present within the membrane in sufficient numbers and must be mobile within the membrane if rapid permeation is to occur. Thus, two physicochemical factors contribute to *P*, namely solubility in the membrane (which can be expressed as a partition coefficient for the substance distributed between the membrane phase and the aqueous environment) and diffusivity, which is a measure of the mobility of molecules within the lipid and is expressed as a diffusion coefficient. The diffusion coefficient varies only slightly between different drugs, as noted above, so the most important variable is the partition coefficient (Fig. 8.2). Consequently, there is a close correlation between lipid solubility and the permeability of the cell membrane to different substances. For this reason, lipid solubility is one of the most important determinants of the pharmacokinetic characteristics of a drug, and many properties—such as rate of absorption from the gut, penetration into different tissues and the extent of renal elimination—can be predicted from knowledge of a drug's lipid solubility.

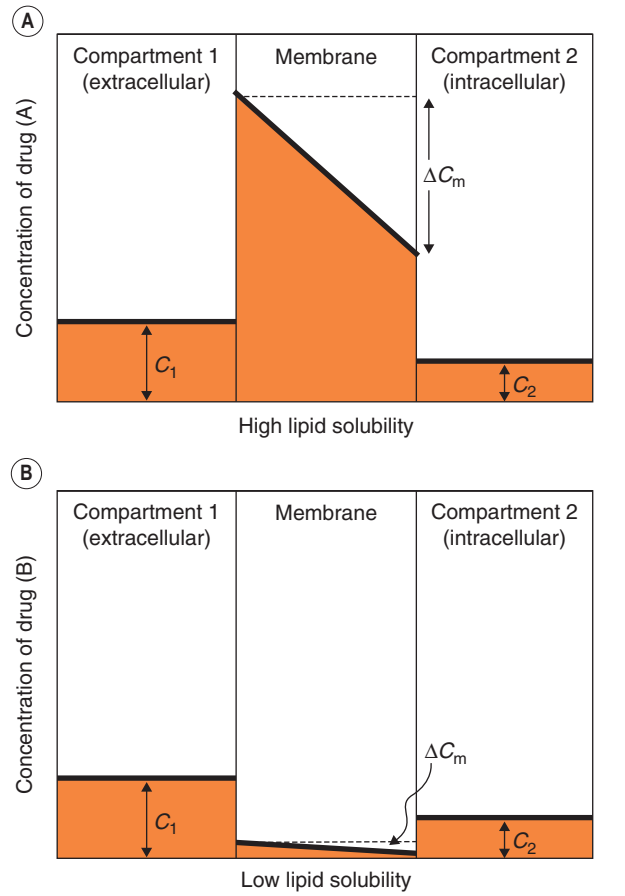

Fig. 8.2 The importance of lipid solubility in membrane permeation. [**A**] and [**B**] Figures show the concentration profile in a lipid membrane separating two aqueous compartments. A lipid-soluble drug (A) is subject to a much larger transmembrane concentration gradient (ΔC_m) than a lipid-insoluble drug (B). It therefore diffuses more rapidly, even though the aqueous concentration gradient (C_1–C_2) is the same in both cases.

pH and ionisation

One important complicating factor in relation to membrane permeation is that many drugs are weak acids or bases, and therefore exist in both unionised and ionised form, the ratio of the two forms varying with pH. For a weak base, the ionisation reaction is:

$$BH^+ \rightleftharpoons^{K_a} B + H^+$$

and the dissociation constant pK_a is given by the Henderson–Hasselbalch equation

$$pK_a = pH + \log_{10}\frac{[BH^+]}{[B]}$$

For a weak acid:

$$AH \rightleftharpoons^{K_a} A^- + H^+$$

$$pK_a = pH + \log_{10}\frac{[AH]}{[A^-]}$$

In either case, the ionised species, BH^+ or A^-, has very low lipid solubility and is virtually unable to permeate membranes except where a specific transport mechanism exists. The lipid solubility of the uncharged species, B or AH, depends on the chemical nature of the drug; for many drugs, the uncharged species is sufficiently lipid soluble to permit rapid membrane permeation, although there are exceptions (e.g. aminoglycoside antibiotics; see Ch. 50) where even the uncharged molecule is insufficiently lipid soluble to cross membranes appreciably. This is usually because of the occurrence of hydrogen-bonding groups (such as hydroxyl in sugar moieties in aminoglycosides) that render the uncharged molecule hydrophilic.

pH partition and ion trapping

Ionisation affects not only the rate at which drugs permeate membranes but also the steady-state distribution of drug molecules between aqueous compartments, if a pH difference exists between them. Figure 8.3 shows how a weak acid (e.g. **aspirin**, pK_a 3.5) and a weak base (e.g. **pethidine**, pK_a 8.6) would be distributed at equilibrium between three body compartments, namely plasma (pH 7.4), alkaline urine (pH 8) and gastric juice (pH 3). Within each compartment, the ratio of ionised to unionised drug is governed by the pK_a of the drug and the pH of that compartment. It is assumed that the unionised species can cross the membrane, and therefore reaches an equal concentration in each compartment. The ionised species is assumed not to cross at all. The result is that, at equilibrium, the total (ionised + unionised) concentration of the drug will be different in the two compartments, with an acidic drug being concentrated in the compartment with high pH ('ion trapping'), and vice versa. The concentration gradients produced by ion trapping can theoretically be very large if there is a large pH difference between compartments. Thus, aspirin would be concentrated more than four-fold with respect to plasma in an alkaline renal tubule, and about 6000-fold in plasma with respect to the acidic gastric contents. Such large gradients are not achieved in reality for two main reasons. First, the attribution of total impermeability to the charged species is not realistic, and even a small permeability will attenuate considerably the concentration difference that can be reached. Second, body compartments rarely approach equilibrium. Neither the gastric contents nor the renal tubular fluid stands still, and the resulting flux of drug molecules reduces the concentration gradients well below the theoretical equilibrium conditions. The pH partition mechanism nonetheless correctly explains some of the qualitative effects of pH changes in different body compartments on the pharmacokinetics of weakly acidic or basic drugs, particularly in relation to renal excretion and to penetration of the blood–brain barrier.

pH partition is not the main determinant of the site of absorption of drugs from the gastrointestinal tract. This is because the enormous absorptive surface area of the villi and microvilli in the ileum compared with the much smaller surface area in the stomach is of overriding importance. Thus, absorption of an acidic drug such as **aspirin** is promoted by drugs that accelerate gastric emptying (e.g. **metoclopramide**) and retarded by drugs that slow gastric emptying (e.g. **propantheline**), despite the fact that the acidic pH of the stomach contents favours absorption of weak acids. Values of pK_a for some common drugs are shown in Figure 8.4.

There are several important consequences of pH partition:

- Free-base trapping of some antimalarial drugs (e.g. **chloroquine**, see Ch. 53) in the acidic environment in the food vacuole of the malaria parasite contributes to the disruption of the haemoglobin digestion pathway that underlies their toxic effect on the parasite.
- Urinary acidification accelerates excretion of weak bases and retards that of weak acids.
- Urinary alkalinisation has the opposite effects: it reduces excretion of weak bases and increases excretion of weak acids.
- Increasing plasma pH (e.g. by administration of sodium bicarbonate) causes weakly acidic drugs to be extracted from the CNS into the plasma. Conversely, reducing plasma pH (e.g. by administration of a carbonic anhydrase inhibitor such as **acetazolamide**) causes weakly acidic drugs to become concentrated in the CNS, increasing their neurotoxicity. This has practical consequences in choosing a means to alkalinise urine in treating aspirin overdose: bicarbonate and acetazolamide each increase urine pH and hence increase salicylate elimination, but bicarbonate reduces whereas acetazolamide increases distribution of salicylate to the CNS.

CARRIER-MEDIATED TRANSPORT

Many cell membranes possess specialised transport mechanisms that regulate entry and exit of physiologically important molecules, such as sugars, amino acids, neurotransmitters and metal ions. They are broadly divided into *solute carrier (SLC) transporters* and *ATP-binding cassette (ABC) transporters*. The former mediate passive movement of solutes down their electrochemical gradient, while the latter are active pumps fuelled by ATP. Over 300 human genes are believed to code these transporters, most of which act mainly on endogenous substrates, but some also transport foreign chemicals ('xenobiotics') including drugs (see Hediger et al., 2004). The role of such transporters in neurotransmitter function is discussed in Chapters 13, 14 and 36.

Organic cation transporters and organic anion transporters

Two structurally related SLC carriers of importance in drug distribution are the organic cation transporters (OCTs) and organic anion transporters (OATs). Generally, such transport systems involve a carrier molecule, i.e. a transmembrane protein that binds one or more molecules or ions, changes conformation and releases them on the other side of the membrane. Such systems may operate purely passively, without any energy source; in this case, they merely facilitate the process of transmembrane equilibration of a single transported species in the direction of

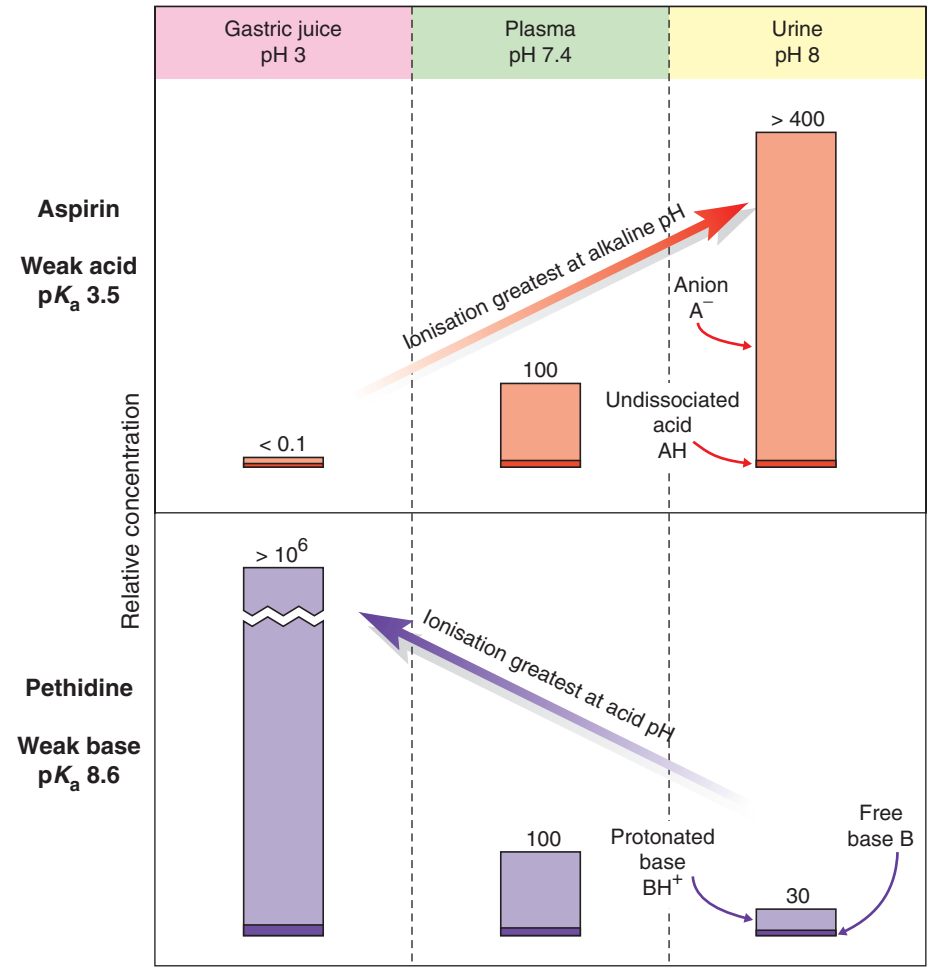

Fig. 8.3 Theoretical partition of a weak acid (aspirin) and a weak base (pethidine) between aqueous compartments (urine, plasma and gastric juice) according to the pH difference between them. Numbers represent relative concentrations (total plasma concentration = 100). It is assumed that the uncharged species in each case can permeate the cellular barrier separating the compartments, and therefore reaches the same concentration in all three. Variations in the fractional ionisation as a function of pH give rise to the large total concentration differences with respect to plasma.

its electrochemical gradient. The mechanism is called facilitated diffusion and the transporter is a 'uniporter'. The OCTs (several families of SLC transporters) translocate dopamine, choline and various drugs including **vecuronium**, **quinine** and **procainamide**. They are uniporters and cause facilitated diffusion down the electrochemical gradient. OCT2 (transporter in proximal tubular cells in the kidney) concentrates drugs such as **cisplatin** (an important anticancer drug) in these cells, an explanation of its selective nephrotoxicity; related drugs (e.g. **carboplatin**, **oxaliplatin**) are not transported by OCT2 and are less nephrotoxic; competition with **cimetidine** for OCT2 offers possible protection against cisplatin nephrotoxicity (Fig. 8.5). Other SLCs are coupled to the electrochemical gradient of Na^+ or other ions across the membrane, generated by ATP-dependent ion pumps (see Ch. 4); in this case, transport can occur against an electrochemical gradient. It may involve exchange of one molecule for another ('antiport') or transport of two molecules together in the same direction ('symport'). The OATs are responsible for the renal secretion of urate, prostaglandins, several vitamins and *p*-amino hippurate, and for drugs such as **probenecid** as well as many antibiotics, antiviral drugs, non-steroidal anti-inflammatory drugs and antineoplastic drugs among others. Uptake is driven by exchange with intracellular dicarboxylic acids (mainly α-ketoglutarate, partly derived from cellular metabolism and partly by co-transport with Na^+ entering cells down its concentration gradient). Metabolic energy is provided by ATP for Na^+/K^+ exchange. Carrier-mediated transport, because it involves a binding step, shows the characteristic of saturation.

Carriers of this type are ubiquitous, and many pharmacological effects are the result of interference with them. Thus nerve terminals have transport mechanisms for accumulating specific neurotransmitters, and there are many examples of drugs that act by inhibiting these transport mechanisms (see Chs 13, 14 and 36). From a general pharmacokinetic point of view, however, the main sites where SLCs, including OCTs and OATs, are expressed and carrier-mediated drug transport is important are:

- the blood–brain barrier
- the gastrointestinal tract
- the renal tubule

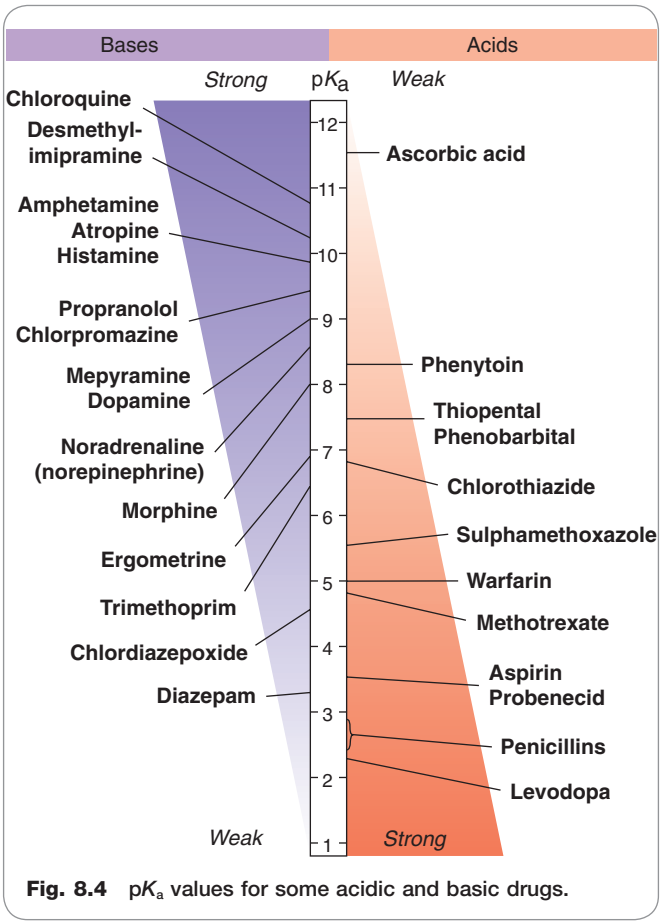

Fig. 8.4 pK_a values for some acidic and basic drugs.

- the biliary tract
- the placenta.

P-glycoprotein transporters

P-glycoproteins (P-gp; P for 'permeability'), which belong to the ABC transporter superfamily, are the second important class of transporters, and responsible for multidrug resistance in cancer cells. They are present in renal tubular brush border membranes, in bile canaliculi, in astrocyte foot processes in brain microvessels, and in the gastrointestinal tract. They play an important part in absorption, distribution and elimination of many drugs, and are often co-located with SLC drug carriers, so that a drug that has been concentrated by, for example, an OAT transporter in the basolateral membrane of a renal tubular cell may then be pumped out of the cell by a P-gp in the luminal membrane.

Polymorphic variation in the genes coding SLCs and P-gp contributes to individual genetic variation in responsiveness to different drugs. OCT1 transports several drugs, including **metformin** (used to treat diabetes; see Ch. 30), into hepatocytes (in contrast to OCT2 which is active in renal proximal tubular cells, see above). Metformin acts partly through intracellular effects within hepatocytes. Single nucleotoide polymorphisms (SNPs; Ch. 56) that impair the function of OCT1 influence the effectiveness of metformin (Fig. 8.6). This is but one example of many genetic influences on drug effectiveness or toxicity via altered activity of carriers that influence drug disposition. Furthermore, induction or competitive inhibition of transport can occur in the presence of a second ligand that binds the carrier, so there is a potential for drug interaction (see Fig. 8.5 and Ch. 56). The characteristics of transport systems are discussed later, when patterns of distribution and elimination in the body as a whole are considered more fully.

In addition to the processes so far described, which govern the transport of drug molecules across the barriers between different aqueous compartments, two additional factors have a major influence on drug distribution and elimination. These are:

- binding to plasma proteins
- partition into body fat and other tissues.

Movement of drugs across cellular barriers

- To traverse cellular barriers (e.g. gastrointestinal mucosa, renal tubule, blood–brain barrier, placenta), drugs have to cross lipid membranes.
- Drugs cross lipid membranes mainly (a) by passive diffusional transfer and (b) by carrier-mediated transfer.
- The main factor that determines the rate of passive diffusional transfer across membranes is a drug's lipid solubility. Molecular weight is less important.
- Many drugs are weak acids or weak bases; their state of ionisation varies with pH according to the Henderson–Hasselbalch equation.
- With weak acids or bases, only the uncharged species (the protonated form for a weak acid, the unprotonated form for a weak base) can diffuse across lipid membranes; this gives rise to pH partition.
- pH partition means that weak acids tend to accumulate in compartments of relatively high pH, whereas weak bases do the reverse.
- Carrier-mediated transport involving solute carriers (SLCs) including organic cation transporters (OCTs) and organic anion transporters (OATs), and P-gps (ABC transporters) in the renal tubule, blood–brain barrier and gastrointestinal epithelium are important in determining the distribution of many drugs.

BINDING OF DRUGS TO PLASMA PROTEINS

At therapeutic concentrations in plasma, many drugs exist mainly in bound form. The fraction of drug that is free in aqueous solution can be less than 1%, the remainder being associated with plasma protein. It is the unbound drug that is pharmacologically active. Such seemingly small differences in protein binding (e.g. 99.5 versus 99.0%) can have large effects on free drug concentration and drug effect. Such differences are common between human plasma and plasma from species used in preclinical drug testing, and must be taken into account when estimating a suitable dose for 'first time in human' studies. The most important plasma protein in relation to drug binding is albumin, which binds many acidic drugs (e.g. **warfarin**, non-steroidal anti-inflammatory drugs, sulfonamides) and a smaller number of basic drugs (e.g. tricyclic antidepressants and **chlorpromazine**). Other plasma proteins, including β-globulin and an acid glycoprotein that increases in inflammatory disease, have also been implicated in the binding of certain basic drugs, such as **quinine**.

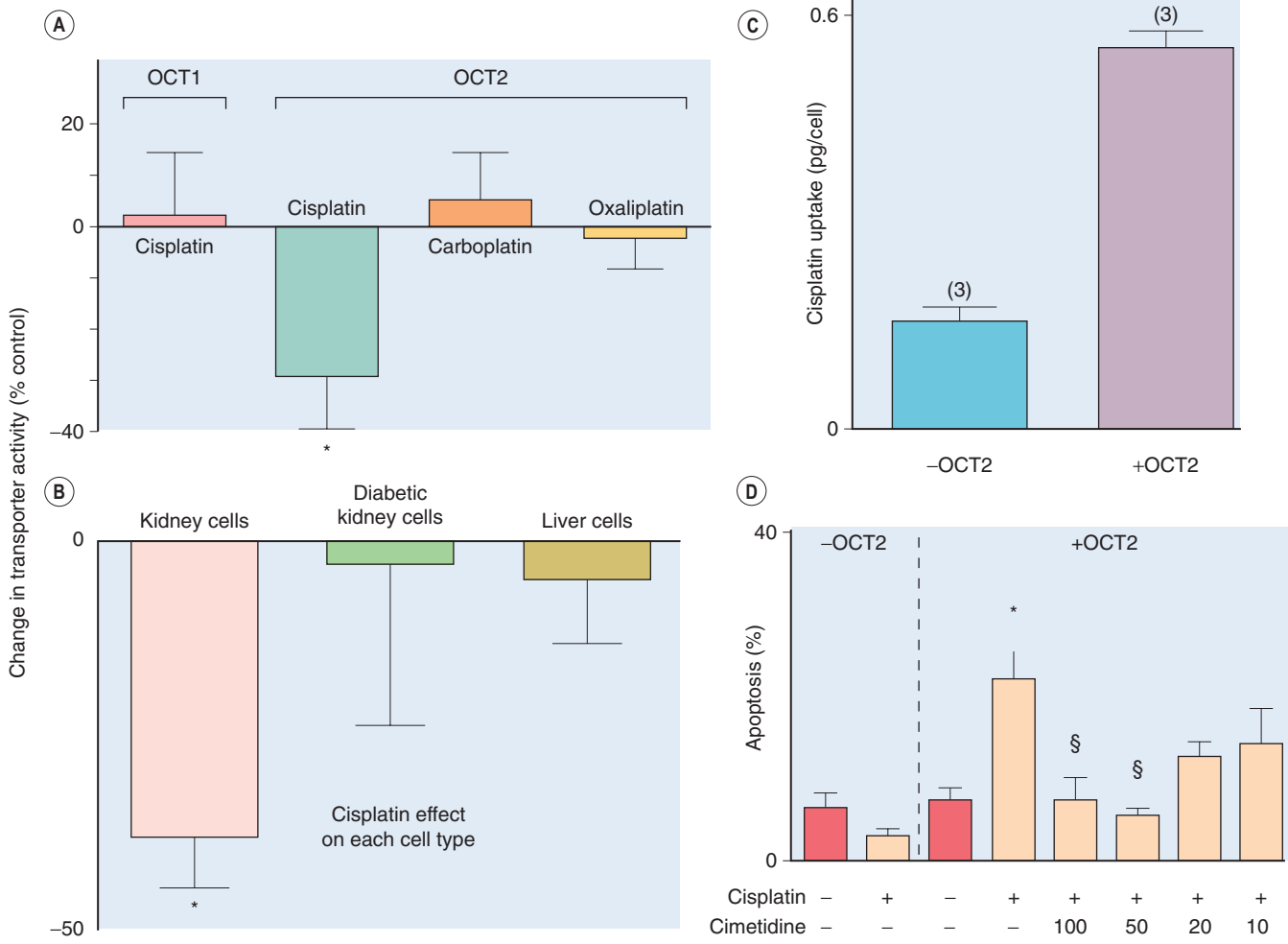

Fig. 8.5 Human organic cation transporter 2 (OCT2) mediates cisplatin nephrotoxicity. OCT2 is expressed in kidney whereas OCT1 is expressed in liver. Cisplatin (100 µmol/l) influences the activity of OCT2 but not of OCT1, each expressed in a cultured cell line [**A**], whereas the less nephrotoxic drugs carboplatin and oxaliplatin do not. Cisplatin similarly influences OCT2 activity in fresh human kidney tubule cells but not in fresh hepatocytes or kidney cells from diabetic patients who are less susceptible to ciplatin nephrotoxicity [**B**]. Cisplatin accumulates in cells that express OCT2 [**C**] and causes cell death [**D**]. Cimetidine competes with cisplatin for OCT2 and concentration dependently protects against cisplatin-induced apoptosis [**D**]—cimetidine concentrations are in µmol/l. (Data redrawn from Ciarimboli G et al. 2005 Am J Pathol 167: 1477–1484.)

The amount of a drug that is bound to protein depends on three factors:

- the concentration of free drug
- its affinity for the binding sites
- the concentration of protein.

As a first approximation, the binding reaction can be regarded as a simple association of the drug molecules with a finite population of binding sites, exactly analogous to drug–receptor binding (see Ch. 2):

$$\underset{\text{free drug}}{D} + \underset{\text{binding site}}{S} \rightleftharpoons \underset{\text{complex}}{DS}$$

The usual concentration of albumin in plasma is about 0.6 mmol/l (4 g/100 ml). With two sites per albumin molecule, the drug-binding capacity of plasma albumin would therefore be about 1.2 mmol/l. For most drugs, the total plasma concentration required for a clinical effect is much less than 1.2 mmol/l, so with usual therapeutic doses the binding sites are far from saturated, and the concentration bound [DS] varies nearly in direct proportion to the free concentration [D]. Under these conditions, the fraction bound, [DS]/([D] + [DS]), is independent of the drug concentration. However, some drugs, for example **tolbutamide** (Ch. 30), work at plasma concentrations at which the binding to protein is approaching saturation (i.e. on the flat part of the binding curve). This means that adding more drug to the plasma increases its free concentration disproportionately. Doubling the dose of such a drug can therefore more than double the free (pharmacologically active) concentration. This is illustrated in Figure 8.7.

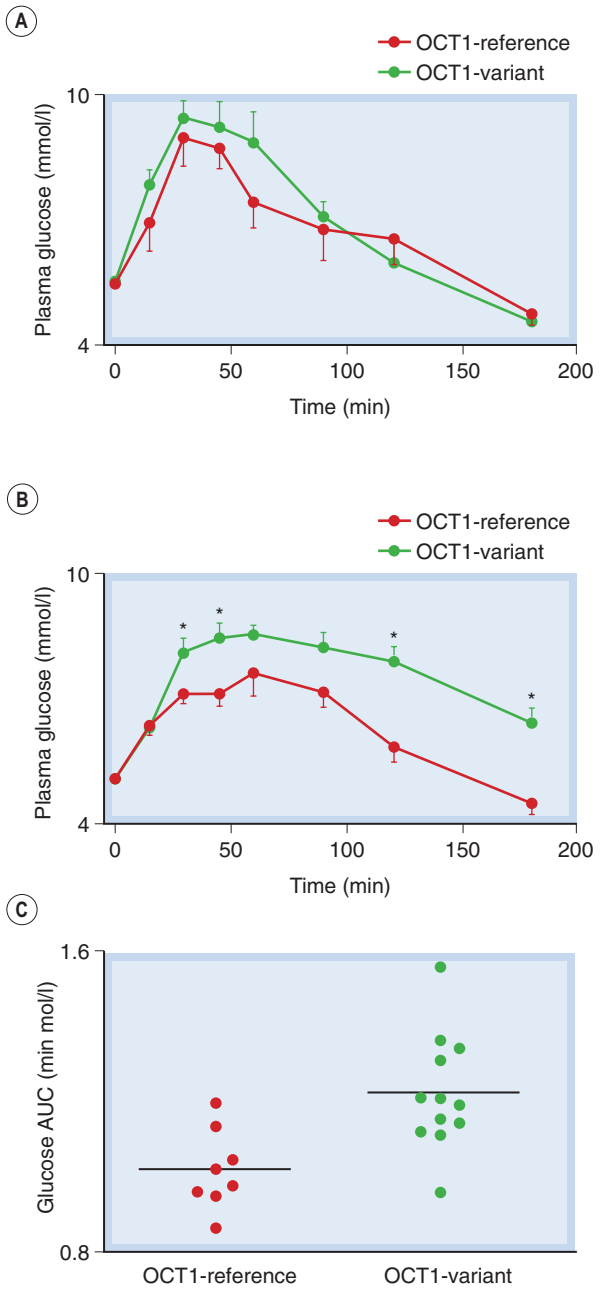

Fig. 8.6 Genetic variants of organic cation transporter 1 (OCT1) are associated with different responses to metformin in healthy humans. [**A**] An oral glucose tolerance test (OGTT) gave similar plasma glucose responses in control subjects with only reference *OCT1* alleles versus subjects with at least one reduced function *OCT1* allele. [**B**] In contrast, after metformin treatment the OGTT response was less in the same reference subjects than in those with reduced function *OCT1* alleles. [**C**] Glucose exposure estimated by area under the glucose time curves (AUC) was significantly lower in subjects with only reference OCT1 alleles, $P = 0.004$. (Data redrawn from Yan Shu et al. 2007 J Clin Invest 117: 1422–1431.)

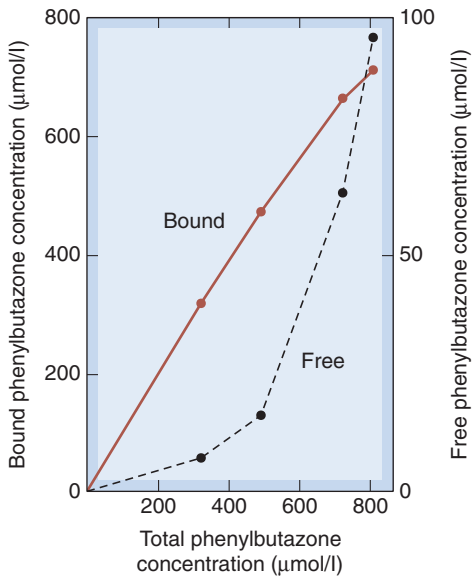

Fig. 8.7 Binding of phenylbutazone to plasma albumin. The graph shows the disproportionate increase in free concentration as the total concentration increases, owing to the binding sites approaching saturation. (Data from Brodie B, Hogben C A M 1957 J Pharm Pharmacol 9: 345.)

Binding sites on plasma albumin bind many different drugs, so competition can occur between them. If two drugs (A and B) compete in this way, administration of drug B can reduce the protein binding, and hence increase the free plasma concentration, of drug A. To do this, drug B needs to occupy an appreciable fraction of the binding sites. Few therapeutic drugs affect the binding of other drugs because they occupy, at therapeutic plasma concentrations, only a tiny fraction of the available sites. *Sulfonamides* (Ch. 50) are an exception, because they occupy about 50% of the binding sites at therapeutic concentrations and so can cause harmful effects by displacing other drugs or, in premature babies, bilirubin (Ch. 56). Much has been made of binding interactions of this kind as a source of untoward drug interactions in clinical medicine, but this type of competition is less important than was once thought (see Ch. 56).

PARTITION INTO BODY FAT AND OTHER TISSUES

Fat represents a large, non-polar compartment. In practice, this is important for only a few drugs, mainly because the effective fat:water partition coefficient is relatively low for most drugs. **Morphine**, for example, although quite lipid soluble enough to cross the blood–brain barrier, has a lipid:water partition coefficient of only 0.4, so sequestration of the drug by body fat is of little importance. **Thiopental**, by comparison (fat:water partition coefficient approximately 10), accumulates substantially in body fat. This has important consequences that limit its usefulness as an intravenous anaesthetic to short-term initiation ('induction') of anaesthesia (Ch. 40).

The second factor that limits the accumulation of drugs in body fat is its low blood supply—less than 2% of the

Binding of drugs to plasma proteins

- Plasma albumin is most important; β-globulin and acid glycoprotein also bind some drugs.
- Plasma albumin binds mainly acidic drugs (approximately two molecules per albumin molecule). Basic drugs may be bound by β-globulin and acid glycoprotein.
- Saturable binding sometimes leads to a non-linear relation between dose and free (active) drug concentration.
- Extensive protein binding slows drug elimination (metabolism and/or glomerular filtration).
- Competition between drugs for protein binding can lead, rarely, to clinically important drug interactions.

cardiac output. Consequently, drugs are delivered to body fat rather slowly, and the theoretical equilibrium distribution between fat and body water is approached slowly. For practical purposes, therefore, partition into body fat when drugs are given acutely is important only for a few highly lipid-soluble drugs (e.g. general anaesthetics; Ch. 40). When lipid-soluble drugs are given chronically, however, accumulation in body fat is often significant (e.g. benzodiazepines; Ch. 43). Some dugs and environmental contaminants, if ingested intermittently, accumulate slowly but progressively in body fat.

Body fat is not the only tissue in which drugs can accumulate. **Chloroquine**—an antimalarial drug (Ch. 53)—has a high affinity for melanin and is taken up by the retina, which is rich in melanin granules, accounting for its ocular toxicity. Tetracyclines (Ch. 50) accumulate slowly in bones and teeth, because they have a high affinity for calcium, and should not be used in children for this reason. Very high concentrations of **amiodarone** (an antidysrhythmic drug; Ch. 21) accumulate in liver and lung during chronic use, causing hepatitis and interstitial pulmonary fibrosis.

DRUG ABSORPTION AND ROUTES OF ADMINISTRATION

The main routes of drug administration and elimination are shown schematically in Figure 8.8. Absorption is defined as the passage of a drug from its site of administration into the plasma. It is important for all routes of administration except intravenous injection, where it is complete by definition. There are instances, such as topical administration of a steroid cream to skin or inhalation of a bronchodilator aerosol to treat asthma (Ch. 27), where absorption as just defined is not required for the drug to act, but in most cases the drug must enter plasma before reaching its site of action.

The main routes of administration are:

- oral
- sublingual
- rectal
- application to other epithelial surfaces (e.g. skin, cornea, vagina and nasal mucosa)
- inhalation
- injection
  - subcutaneous
  - intramuscular
  - intravenous
  - intrathecal
  - intravitreal.

ORAL ADMINISTRATION

Most drugs are taken by mouth and swallowed. Little absorption occurs until the drug enters the small intestine.

DRUG ABSORPTION FROM THE INTESTINE

For most drugs the mechanism of absorption is the same as for other epithelial barriers, namely passive transfer at a rate determined by the ionisation and lipid solubility of the drug molecules. Figure 8.9 shows the absorption of various weak acids and bases as a function of pK_a. As expected, strong bases of pK_a 10 or higher are poorly absorbed, as are strong acids of pK_a less than 3, because they are fully ionised. The arrow poison curare used by South American Indians contains quaternary ammonium compounds that block neuromuscular transmission (Ch. 13). These strong bases are poorly absorbed from the gastrointestinal tract, so the meat from animals killed in this way was safe to eat.

In a few instances, intestinal drug absorption depends on carrier-mediated transport rather than simple lipid diffusion. Examples include **levodopa**, used in treating Parkinson's disease (see Ch. 39), which is taken up by the carrier that normally transports phenylalanine, and **fluorouracil** (Ch. 55), a cytotoxic drug that is transported by the system that carries natural pyrimidines (thymine and uracil). Iron is absorbed via specific carriers in the epithelial cell membranes of jejunal mucosa, and calcium is absorbed by means of a vitamin D-dependent carrier system.

FACTORS AFFECTING GASTROINTESTINAL ABSORPTION

Typically, about 75% of a drug given orally is absorbed in 1–3 h, but numerous factors alter this, some physiological and some to do with the formulation of the drug. The main factors are:

- gastrointestinal motility
- splanchnic blood flow
- particle size and formulation
- physicochemical factors.

Gastrointestinal motility has a large effect. Many disorders (e.g. migraine, diabetic neuropathy) cause gastric stasis and slow drug absorption. Drug treatment can also affect motility, either reducing (e.g. drugs that block muscarinic receptors; see Ch. 13) or increasing it (e.g. **metoclopramide**, an antiemetic used in migraine to facilitate absorption of analgesic). Excessively rapid movement of gut contents (e.g. in some forms of diarrhoea) can impair absorption. Several drugs (e.g. **propranolol**) reach a higher plasma concentration if they are taken after a meal, probably because food increases splanchnic blood flow. Conversely, splanchnic blood flow is greatly reduced by hypovolaemia or heart failure, with a resultant reduction of drug absorption.

Particle size and formulation have major effects on absorption. In 1971, patients in a New York hospital were found to require unusually large maintenance doses of

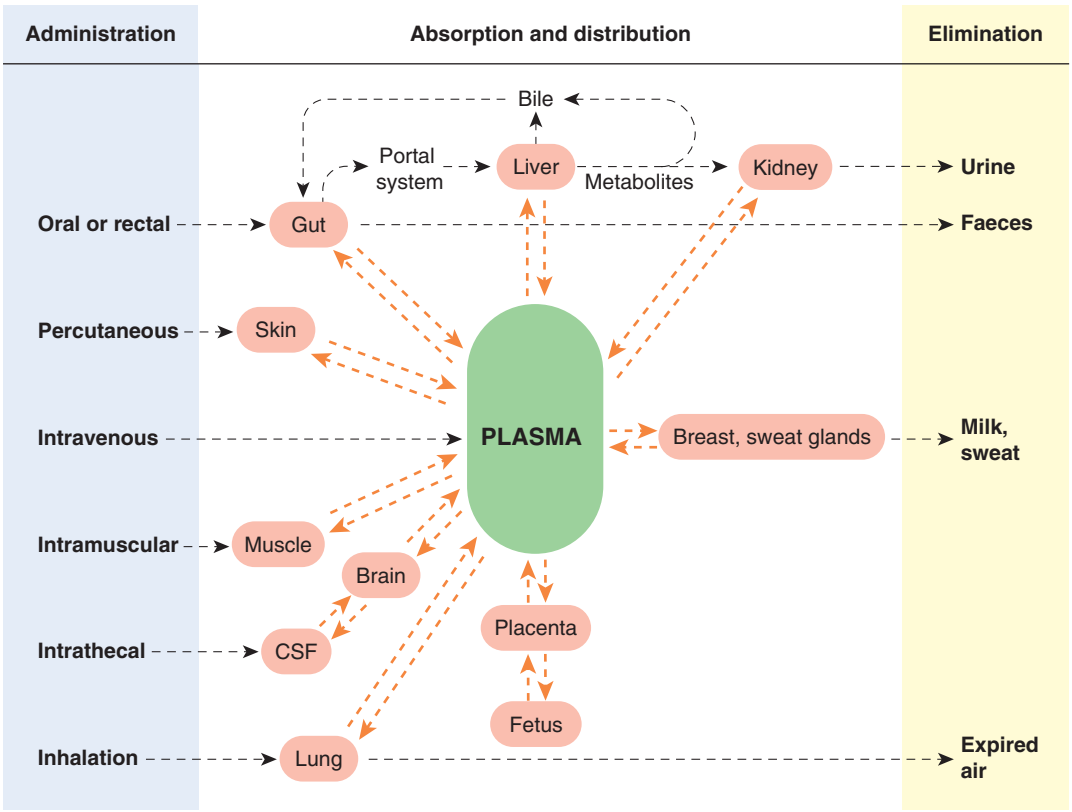

Fig. 8.8 The main routes of drug administration and elimination.

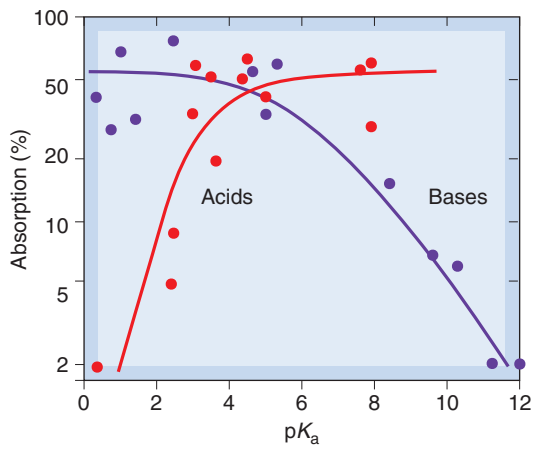

Fig. 8.9 Absorption of drugs from the intestine, as a function of pK_a, for acids and bases. Weak acids and bases are well absorbed; strong acids and bases are poorly absorbed. (Redrawn from Schanker L S et al. 1957 J Pharmacol 120: 528.)

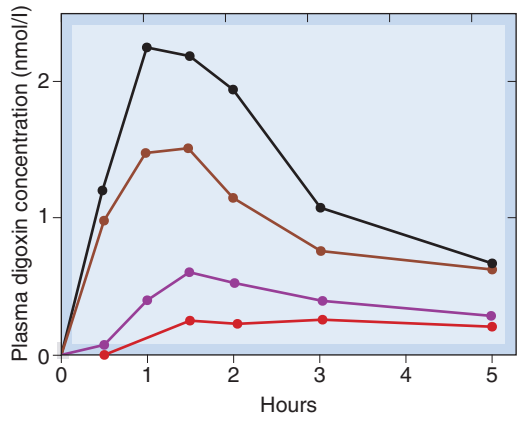

Fig. 8.10 Variation in oral absorption among different formulations of digoxin. The four curves show the mean plasma concentrations attained for the four preparations, each of which was given on separate occasions to four subjects. The large variation has caused the formulation of digoxin tablets to be standardised since this study was published. (From Lindenbaum J et al. 1971 N Engl J Med 285: 1344.)

digoxin (Ch. 21). In a study on normal volunteers, it was found that standard digoxin tablets from different manufacturers resulted in grossly different plasma concentrations (Fig. 8.10), even though the digoxin content of the tablets was the same, because of differences in particle size. Because digoxin is rather poorly absorbed, small differences in the pharmaceutical formulation can make a large difference to the extent of absorption.

Therapeutic drugs are formulated pharmaceutically to produce desired absorption characteristics. Capsules may be designed to remain intact for some hours after ingestion in order to delay absorption, or tablets may have a resistant

coating to give the same effect. In some cases, a mixture of slow- and fast-release particles is included in a capsule to produce rapid but sustained absorption. More elaborate pharmaceutical systems include modified-release preparations that permit less frequent dosing. Such preparations not only increase the dose interval but also reduce adverse effects related to high peak plasma concentrations following administration of a conventional formulation. Osmotically driven 'minipumps' can be implanted experimentally, and some oral extended-release preparations that are used clinically use the same principle, the tablet containing an osmotically active core and being bound by an impermeable membrane with a precisely engineered pore to allow drug to exit in solution, delivering drug at an approximately constant rate into the bowel lumen. Such preparations may, however, cause problems related to high local concentrations of drug in the intestine (an osmotically released preparation of the anti-inflammatory drug **indometacin**, Ch. 26, had to be withdrawn because it caused small bowel perforation), and are sensitive to variations in small bowel transit time that occur during ageing and with disease.

Physicochemical factors (including some drug interactions; Ch. 56) affect drug absorption. **Tetracycline** binds strongly to Ca^{2+}, and calcium-rich foods (especially milk) prevent its absorption (Ch. 50). Bile acid-binding resins such as **colestyramine** (used to treat diarrhoea caused by bile acids) bind several drugs, for example **warfarin** (Ch. 24) and **thyroxine** (Ch. 33).

When drugs are administered by mouth, the intention is usually that they should be absorbed and cause a systemic effect, but there are exceptions. **Vancomycin** is very poorly absorbed, and is administered orally to eradicate toxin-forming *Clostridium difficile* from the gut lumen in patients with pseudomembranous colitis (an adverse effect of broad-spectrum antibiotics caused by appearance of this organism in the bowel). **Mesalazine** is a formulation of 5-aminosalicylic acid in a pH-dependent acrylic coat that degrades in the terminal ileum and proximal colon, and is used to treat inflammatory bowel disease affecting this part of the gut. **Olsalazine** is a prodrug (see below) consisting of a dimer of two molecules of 5-aminosalicylic acid that is cleaved by colonic bacteria in the distal bowel and is used to treat patients with distal colitis.

Bioavailability and bioequivalence

To get from the lumen of the small intestine into the systemic circulation, a drug must not only penetrate the intestinal mucosa, it must also run the gauntlet of enzymes that may inactivate it in gut wall and liver, referred to as 'presystemic' or 'first-pass' metabolism or clearance. The term *bioavailability* is used to indicate the fraction (F) of an orally administered dose that reaches the systemic circulation as intact drug, taking into account both absorption and local metabolic degradation. F is measured by determining the plasma drug concentration versus time curves in a group of subjects following oral and (on a separate occasion) intravenous administration (the fraction absorbed following an intravenous dose is 1 by definition). The areas under the plasma concentration time curves (AUC) are used to estimate F as $AUC_{oral}/AUC_{intravenous}$. Bioavailability is not a characteristic solely of the drug preparation: variations in enzyme activity of gut wall or liver, in gastric pH or intestinal motility all affect it. Because of this, one cannot speak strictly of the bioavailability of a particular preparation, but only of that preparation in a given individual on a particular occasion, and F determined in a group of healthy volunteer subjects may differ substantially from the value determined in patients with diseases of gastrointestinal or circulatory systems.

Bioavailability relates only to the total proportion of the drug that reaches the systemic circulation and neglects the rate of absorption. If a drug is completely absorbed in 30 min, it will reach a much higher peak plasma concentration (and have a more dramatic effect) than if it were absorbed more slowly. Regulatory authorities—which have to make decisions about the licensing of products that are 'generic equivalents' of patented products—require evidence of 'bioequivalence' based on the maximum concentration achieved (C_{max}) and time between dosing and C_{max} (t_{max}) as well as $AUC_{(0-\infty)}$. For most drugs, each of these parameters ($AUC_{(0-\infty)}$, C_{max}, t_{max}) must lie between 80% and 125% of the lead product for the new generic product to be accepted as bioequivalent.

SUBLINGUAL ADMINISTRATION

Absorption directly from the oral cavity is sometimes useful (provided the drug does not taste too horrible) when a rapid response is required, particularly when the drug is either unstable at gastric pH or rapidly metabolised by the liver. **Glyceryl trinitrate** and **buprenorphine** are examples of drugs that are often given sublingually (Chs 21 and 41, respectively). Drugs absorbed from the mouth pass directly into the systemic circulation without entering the portal system, and so escape first-pass metabolism by enzymes in the gut wall and liver.

RECTAL ADMINISTRATION

Rectal administration is used for drugs that are required either to produce a local effect (e.g. anti-inflammatory drugs for use in ulcerative colitis) or to produce systemic effects. Absorption following rectal administration is often unreliable, but this route can be useful in patients who are vomiting or are unable to take medication by mouth (e.g. postoperatively). It is used to administer **diazepam** to children who are in *status epilepticus* (Ch. 44), in whom it is difficult to establish intravenous access.

APPLICATION TO EPITHELIAL SURFACES

CUTANEOUS ADMINISTRATION

Cutaneous administration is used when a local effect on the skin is required (e.g. topically applied steroids). Appreciable absorption may nonetheless occur and lead to systemic effects.

Most drugs are absorbed very poorly through unbroken skin. However, a number of organophosphate insecticides (see Ch. 13), which need to penetrate an insect's cuticle in order to work, are absorbed through skin, and accidental poisoning occurs in farm workers.

▼ A case is recounted of a 35-year-old florist in 1932. 'While engaged in doing a light electrical repair job at a work bench he sat down in a chair on the seat of which some "Nico-Fume liquid" (a 40% solution of free nicotine) had been spilled. He felt the solution wet through his clothes to the skin over the left buttock, an area about the size of the palm of his hand. He thought nothing further of it and continued at his work for about 15 minutes, when he was suddenly seized with nausea and faintness … and found himself in a drenching sweat. On the way to hospital he lost consciousness.' He survived, just, and then

4 days later: 'On discharge from the hospital he was given the same clothes that he had worn when he was brought in. The clothes had been kept in a paper bag and were still damp where they had been wet with the nicotine solution.' The sequel was predictable. He survived again but felt thereafter 'unable to enter a greenhouse where nicotine was being sprayed'. Transdermal dosage forms of nicotine are now used to reduce the withdrawal symptoms that accompany stopping smoking (Ch. 48).

Transdermal dosage forms, in which the drug is incorporated in a stick-on patch applied to the skin, are used increasingly, and several drugs—for example **oestrogen** and **testosterone** for hormone replacement (Ch. 34)—are available in this form. Such patches produce a steady rate of drug delivery and avoid presystemic metabolism. **Fentanyl** is available in a patch to treat intermittent breakthrough pain (Ch. 41). However, the method is suitable only for lipid-soluble drugs and is relatively expensive.

NASAL SPRAYS

Some peptide hormone analogues, for example of **antidiuretic hormone** (Ch. 32) and of **gonadotrophin-releasing hormone** (see Ch. 34), are given as nasal sprays, as is **calcitonin** (Ch. 35). Absorption is believed to take place through mucosa overlying nasal-associated lymphoid tissue. This is similar to mucosa overlying Peyer's patches in the small intestine, which is also unusually permeable.

EYE DROPS

Many drugs are applied as eye drops, relying on absorption through the epithelium of the conjunctival sac to produce their effects. Desirable local effects within the eye can be achieved without causing systemic side effects; for example, **dorzolamide** is a carbonic anhydrase inhibitor that is given as eye drops to lower ocular pressure in patients with glaucoma. It achieves this without affecting the kidney (see Ch. 28), thus avoiding the acidosis that is caused by oral administration of **acetazolamide**. Some systemic absorption from the eye occurs, however, and can result in unwanted effects (e.g. bronchospasm in asthmatic patients using **timolol** eye drops for glaucoma).

ADMINISTRATION BY INHALATION

Inhalation is the route used for volatile and gaseous anaesthetics (see Ch. 40), the lung serving as the route of both administration and elimination. The rapid exchange resulting from the large surface area and blood flow makes it possible to achieve rapid adjustments of plasma concentration. The pharmacokinetic behaviour of inhalation anaesthetics is discussed more fully in Chapter 40.

Drugs used for their effects on the lung are also given by inhalation, usually as an aerosol. Glucocorticoids (e.g. **beclometasone dipropionate**) and bronchodilators (e.g. **salbutamol**; Ch. 27) are given in this way to achieve high local concentrations in the lung while minimising systemic side effects. However, drugs given by inhalation in this way are usually partly absorbed into the circulation, and systemic side effects (e.g. tremor following salbutamol) can occur. Chemical modification of a drug may minimise such absorption. For example, **ipratropium**, a muscarinic receptor antagonist (Chs 13 and 27), is a quaternary ammonium ion analogue of atropine. It is used as an inhaled bronchodilator because its poor absorption minimises systemic adverse effects.

ADMINISTRATION BY INJECTION

Intravenous injection is the fastest and most certain route of drug administration. Bolus injection rapidly produces a high concentration of drug, first in the right heart and lungs and then in the systemic circulation. The peak concentration reaching the tissues depends critically on the rate of injection. Administration by steady intravenous infusion avoids the uncertainties of absorption from other sites, while avoiding high peak plasma concentrations caused by bolus injection.

Subcutaneous or intramuscular injection of drugs usually produces a faster effect than oral administration, but the rate of absorption depends greatly on the site of injection and on local blood flow. The rate-limiting factors in absorption from the injection site are:

- diffusion through the tissue
- removal by local blood flow.

Absorption from a site of injection (sometimes but not always desirable, see below) is increased by increased blood flow. *Hyaluronidase* (an enzyme that breaks down the intercellular matrix, thereby increasing diffusion) also increases drug absorption from the site of injection. Conversely, absorption is reduced in patients with circulatory failure ('shock') in whom tissue perfusion is reduced (Ch. 22).

METHODS FOR DELAYING ABSORPTION

It may be desirable to delay absorption, either to produce a local effect or to prolong systemic action. For example, addition of **adrenaline** (**epinephrine**) to a local anaesthetic reduces absorption of the anaesthetic into the general circulation, usefully prolonging the anaesthetic effect (Ch. 42). Formulation of insulin with protamine or zinc produces a long-acting form (see Ch. 30). **Procaine penicillin** (Ch. 50) is a poorly soluble salt of penicillin; when injected as an aqueous suspension, it is slowly absorbed and exerts a prolonged action. Esterification of steroid hormones (e.g. **medroxyprogesterone acetate**, **testosterone propionate**; Ch. 34) and antipsychotic drugs (e.g. **fluphenazine decanoate**; Ch. 45) increases their solubility in oil and slows their rate of absorption when they are injected in an oily solution.

Another method used to achieve slow and continuous absorption of certain steroid hormones (e.g. **estradiol**; Ch. 34) is the subcutaneous implantation of solid pellets. The rate of absorption is proportional to the surface area of the implant.

INTRATHECAL INJECTION

Injection of a drug into the subarachnoid space via a lumbar puncture needle is used for some specialised purposes. **Methotrexate** (Ch. 55) is administered in this way in the treatment of certain childhood leukaemias to prevent relapse in the CNS. Regional anaesthesia can be produced by intrathecal administration of a local anaesthetic such as **bupivacaine** (see Ch. 42); opioid analgesics can also be used in this way (Ch. 41). **Baclofen** (a GABA analogue; Ch. 37) is used to treat disabling muscle spasms. It has been administered intrathecally to minimise its adverse effects. Some antibiotics (e.g. aminoglycosides) cross the blood–brain barrier very slowly, and in rare clinical situations where they are essential (e.g. nervous system infections with bacteria resistant to other antibiotics) can be given

Drug absorption and bioavailability

- Drugs of very low lipid solubility, including those that are strong acids or bases, are generally poorly absorbed from the gut.
- A few drugs (e.g. **levodopa**) are absorbed by carrier-mediated transfer.
- Absorption from the gut depends on many factors, including:
  - gastrointestinal motility
  - gastrointestinal pH
  - particle size
  - physicochemical interaction with gut contents (e.g. chemical interaction between calcium and tetracycline antibiotics).
- Bioavailability is the fraction of an ingested dose of a drug that gains access to the systemic circulation. It may be low because absorption is incomplete, or because the drug is metabolised in the gut wall or liver before reaching the systemic circulation.
- Bioequivalence implies that if one formulation of a drug is substituted for another, no clinically untoward consequences will ensue.

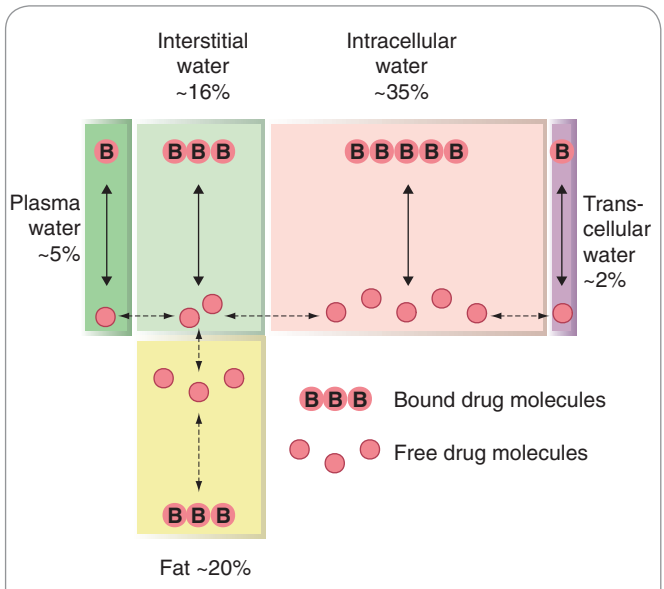

Fig. 8.11 The main body fluid compartments, expressed as a percentage of body weight. Drug molecules exist in bound or free form in each compartment, but only the free drug is able to move between the compartments.

intrathecally or directly into the cerebral ventricles via a reservoir.

INTRAVITREAL INJECTION

Ranibizumab (monoclonal antibody fragment that binds to vascular endothelial growth factor; Ch. 22) is given by intravitreal injection by ophthalmologists treating patients with wet age-related macular degeneration.

DISTRIBUTION OF DRUGS IN THE BODY

BODY FLUID COMPARTMENTS

Body water is distributed into four main compartments, as shown in Figure 8.11. The total body water as a percentage of body weight varies from 50% to 70%, being rather less in women than in men.

Extracellular fluid comprises the blood plasma (about 4.5% of body weight), interstitial fluid (16%) and lymph (1.2%). Intracellular fluid (30–40%) is the sum of the fluid contents of all cells in the body. Transcellular fluid (2.5%) includes the cerebrospinal, intraocular, peritoneal, pleural and synovial fluids, and digestive secretions. The fetus may also be regarded as a special type of transcellular compartment. Within each of these aqueous compartments, drug molecules usually exist both in free solution and in bound form; furthermore, drugs that are weak acids or bases will exist as an equilibrium mixture of the charged and uncharged forms, the position of the equilibrium depending on the pH.

The equilibrium pattern of distribution between the various compartments will therefore depend on:

- permeability across tissue barriers
- binding within compartments
- pH partition
- fat:water partition.

To enter the transcellular compartments from the extracellular compartment, a drug must cross a cellular barrier, a particularly important example in the context of pharmacokinetics being the blood–brain barrier.

THE BLOOD–BRAIN BARRIER

The concept of the blood–brain barrier was introduced by Paul Ehrlich to explain his observation that intravenously injected dye stained most tissues yet the brain remained unstained. The barrier consists of a continuous layer of endothelial cells joined by tight junctions and surrounded by pericytes. The brain is consequently inaccessible to many drugs with a lipid solubility that is insufficient to allow penetration of the blood–brain barrier. However, inflammation can disrupt the integrity of the blood–brain barrier, allowing normally impermeant substances to enter the brain (Fig. 8.12); consequently, **penicillin** (Ch. 50) can be given intravenously (rather than intrathecally) to treat bacterial meningitis (which is accompanied by intense inflammation).

Furthermore, in some parts of the CNS, including the *chemoreceptor trigger zone*, the barrier is leaky. This enables **domperidone**, an antiemetic dopamine receptor antagonist (Ch. 29 & 39) that does not penetrate the blood–brain barrier but does access the chemoreceptor trigger zone, to be used to prevent the nausea caused by dopamine agonists such as **apomorphine** when these are used to treat advanced Parkinson's disease. This is achieved without loss of efficacy, because dopamine receptors in the basal ganglia are accessible only to drugs that have traversed the blood–brain barrier.

Methylnaltrexone bromide is a peripherally acting μ-opioid receptor antagonist used in treating opioid-induced constipation in patients requiring opioids as part

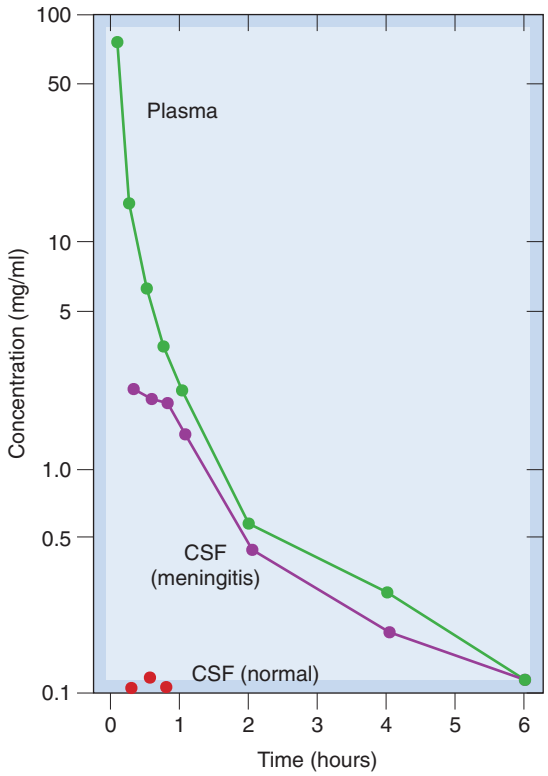

Fig. 8.12 Plasma and cerebrospinal fluid concentrations of an antibiotic (thienamycin) following an intravenous dose (25 mg/kg). In normal rabbits, no drug reaches the cerebrospinal fluid (CSF), but in animals with experimental *Escherichia coli* meningitis the concentration of drug in CSF approaches that in the plasma. (From Patamasucon & McCracken 1973 Antimicrob Agents Chemother 3: 270.)

of palliative care. It has limited gastrointestinal absorption and does not cross the blood–brain barrier, so does not block the desired CNS opioid effects. Several peptides, including bradykinin and enkephalins, increase blood–brain barrier permeability. There is interest in exploiting this to improve penetration of chemotherapy during treatment of brain tumours. In addition, extreme stress renders the blood–brain barrier permeable to drugs such as **pyridostigmine** (Ch. 13), which normally act peripherally.[2]

VOLUME OF DISTRIBUTION

The apparent volume of distribution, V_d, (see Ch. 10) is defined as the volume of fluid required to contain the total amount, Q, of drug in the body at the same concentration as that present in the plasma, C_p:

$$V_d = \frac{Q}{C_p}$$

Values of V_d have been measured for many drugs (see Table 8.1).[3] It is important to avoid identifying a given range of V_d too closely with a particular anatomical compartment. For example, insulin has a measured V_d similar to the volume of plasma water but exerts its effects on muscle, fat and liver via receptors that are exposed to interstitial fluid but not to plasma (Ch. 30).

DRUGS CONFINED TO THE PLASMA COMPARTMENT

The plasma volume is about 0.05 l/kg body weight. A few drugs, such as **heparin** (Ch. 24), are confined to plasma because the molecule is too large to cross the capillary wall easily. More often, retention of a drug in the plasma following a single dose reflects strong binding to plasma protein. It is, nevertheless, the free drug in the interstitial fluid that exerts a pharmacological effect. Following repeated dosing, equilibration occurs and measured V_d increases. Some dyes, such as Evans blue, bind so strongly to plasma albumin that its V_d is used experimentally to measure plasma volume.

DRUGS DISTRIBUTED IN THE EXTRACELLULAR COMPARTMENT

The total extracellular volume is about 0.2 l/kg, and this is the approximate V_d for many polar compounds, such as **vecuronium** (Ch. 13), **gentamicin** and **carbenicillin** (Ch. 50). These drugs cannot easily enter cells because of their low lipid solubility, and they do not traverse the blood–brain or placental barriers freely.

DISTRIBUTION THROUGHOUT THE BODY WATER

Total body water represents about 0.55 l/kg. This approximates the distribution of relatively lipid-soluble drugs that readily cross cell membranes, such as **phenytoin** (Ch. 44) and **ethanol** (Ch. 48). Binding of drug outside the plasma compartment, or partitioning into body fat, increases V_d beyond total body water. Consequently, there are many drugs with V_d greater than the total body volume, such as **morphine** (Ch. 41), tricyclic antidepressants (Ch. 46) and **haloperidol** (Ch. 45). Such drugs are not efficiently removed from the body by haemodialysis, which is therefore unhelpful in managing overdose with such agents.

SPECIAL DRUG DELIVERY SYSTEMS

Several approaches are used or in development to improve drug delivery and localise the drug to the target tissue. They include:

- biologically erodible nanoparticles
- prodrugs
- antibody–drug conjugates
- packaging in liposomes
- coated implantable devices.

[2]This has been invoked to explain the central symptoms of cholinesterase inhibition experienced by some soldiers during the Gulf War. These soldiers may have been exposed to cholinesterase inhibitors (developed as chemical weapons and also, somewhat bizarrely, used externally during the conflict to prevent insect infestation) in the context of the stress of warfare.

[3]The experimental measurement of V_d is complicated by the fact that Q does not stay constant (because of metabolism and excretion of the drug) during the time that it takes for it to be distributed among the various body compartments that contribute to the overall V_d. It therefore has to be calculated indirectly from a series of measurements of plasma concentrations as a function of time (see Fig. 10.1).

Table 8.1 Distribution volumes for some drugs compared with volume of body fluid compartments

Volume (l/kg body weight)	Compartment	Volume of distribution (V_d; l/kg body weight)	Drug(s)
0.05	Plasma	0.05–0.1	Heparin Insulin
		0.1–0.2	Warfarin Sulfamethoxazole Glibenclamide Atenolol
0.2	Extracellular fluid	0.2–0.4	Tubocurarine
		0.4–0.7	Theophylline
0.55	Total body water	1–2	Ethanol Neostigmine Phenytoin Methotrexate Indometacin Paracetamol Diazepam Lidocaine (lignocaine)
		2–5	Glyceryl trinitrate Morphine Propranolol Digoxin Chlorpromazine
		>10	Nortriptyline Imipramine

Drug distribution

- The major compartments are:
  - plasma (5% of body weight)
  - interstitial fluid (16%)
  - intracellular fluid (35%)
  - transcellular fluid (2%)
  - fat (20%).
- Volume of distribution (V_d) is defined as the volume of plasma that would contain the total body content of the drug at a concentration equal to that in the plasma.
- Lipid-insoluble drugs are mainly confined to plasma and interstitial fluids; most do not enter the brain following acute dosing.
- Lipid-soluble drugs reach all compartments and may accumulate in fat.
- For drugs that accumulate outside the plasma compartment (e.g. in fat or by being bound to tissues), V_d may exceed total body volume.

BIOLOGICALLY ERODIBLE NANOPARTICLES

Microspheres of biologically erodible polymers (see Varde & Pack, 2004) can be engineered to adhere to mucosal epithelium in the gut. Such particles can be loaded with drugs, including high-molecular-weight substances, as a means of improving absorption, which occurs both through mucosal absorptive epithelium and also through epithelium overlying Peyer's patches. This approach has yet to be used clinically, but microspheres made from polyanhydride co-polymers of fumaric and sebacic acids by a technique known as phase inversion nanoencapsulation have been used to produce systemic absorption of insulin and of plasmid DNA following oral administration in rats, potentially enabling gene therapy (Ch. 59) to be administered orally. Various polymer nanoparticles, that can be loaded with drug molecules and targeted to specific tissues, are in development for many therapeutic applications (see Singh & Lillard, 2008), particularly as a means of delivering cytotoxic drugs specifically to cancer cells (see Ch. 55).

PRODRUGS

Prodrugs are inactive precursors that are metabolised to active metabolites; they are described in Chapter 9. Some of the examples in clinical use confer no obvious benefits and have been found to be prodrugs only retrospectively, not having been designed with this in mind. However, some do have advantages. For example, the cytotoxic drug **cyclophosphamide** (see Ch. 55) becomes active only after it has been metabolised in the liver; it can therefore be taken orally without causing serious damage to the gastrointestinal epithelium. **Levodopa** is absorbed from the gastrointestinal tract and crosses the blood–brain barrier via an amino acid transport mechanism before conversion to active dopamine in nerve terminals in the basal ganglia (Ch. 39). **Zidovudine** is phosphorylated to its active triphosphate metabolite only in cells containing the appropriate reverse transcriptase, hence conferring selective

toxicity towards cells infected with HIV (Ch. 51). **Valaciclovir** and **famciclovir** are each ester prodrugs, respectively of **aciclovir** and of **penciclovir**. Their bioavailability is greater than that of aciclovir and penciclovir, which are themselves prodrugs that are converted into active metabolites in virally infected cells (Ch. 51).

Other problems could theoretically be overcome by the use of suitable prodrugs; for example, instability of drugs at gastric pH, direct gastric irritation (aspirin was synthesised in the 19th century in a deliberate attempt to produce a prodrug of salicylic acid that would be tolerable when taken by mouth), failure of drug to cross the blood–brain barrier and so on. Progress with this approach remains slow, however, and the optimistic prodrug designer 'will have to bear in mind that an organism's normal reaction to a foreign substance is to burn it up for food'.

ANTIBODY–DRUG CONJUGATES

One of the aims of cancer chemotherapy is to improve the selectivity of cytotoxic drugs (see Ch. 55). One interesting possibility is to attach the drug to an antibody directed against a tumour-specific antigen, which will bind selectively to tumour cells.

PACKAGING IN LIPOSOMES

Liposomes are minute vesicles produced by sonication of an aqueous suspension of phospholipids. They can be filled with non-lipid-soluble drugs, which are retained until the liposome is disrupted. Liposomes are taken up by reticuloendothelial cells, especially in the liver. They are also concentrated in malignant tumours, and there is a possibility of achieving selective delivery of drugs in this way. **Amphotericin**, an antifungal drug used to treat systemic mycoses (Ch. 52), is available in a liposomal formulation that is less nephrotoxic and better tolerated than the conventional form, albeit considerably more expensive. In the future, it may be possible to direct drugs or genes selectively to a specific target by incorporating antibody molecules into liposomal membrane surfaces.

COATED IMPLANTABLE DEVICES

Impregnated coatings have been developed that permit localised drug delivery from implants. Examples include hormonal delivery to the endometrium from intrauterine devices, and delivery of antithrombotic and antiproliferative agents (drugs or radiopharmaceuticals) to the coronary arteries from *stents* (devices inserted via a catheter after a diseased coronary artery has been dilated with a balloon). Stents reduce the occurrence of re-stenosis, but this can still occur at the margin of the device. Coating stents with drugs such as **sirolimus** (a potent immunosuppressant; see Ch. 26) embedded in a surface polymer prevents this important clinical problem.

REFERENCES AND FURTHER READING

Drug distribution (including blood–brain barrier)

Bauer, B., Hartz, A.M.S., Fricker, G., Miller, D.S., 2005. Modulation of P-glycoprotein transport function at the blood–brain barrier. Exp. Biol. Med. 230, 118–127. (*Reviews mechanisms by which P-glycoprotein activity can be modulated, including direct inhibition by specific competitors, and functional and transcriptional modulation*)

Ciarimboli, G., 2008. Organic cation transporters. Xenobiotica 38, 936–971. (*Discusses species- and tissue-specific distribution of different OCT isoforms and polymorphisms in OCTs as a source of variation in drug response*)

de Boer, A.G., van der Sandt, I.C.J., Gaillard, P.J., 2003. The role of drug transporters at the blood–brain barrier. Ann. Rev. Pharmacol. Toxicol. 43, 629–656. (*Reviews the role of carrier- and receptor-mediated transport systems in the blood–brain barrier; these include P-glycoprotein, multidrug-resistance proteins 17, nucleoside transporters, organic anion transporters and large amino acid transporters, the transferrin-1 and -2 receptors, and the scavenger receptors SB-AI and SB-BI*)

Eraly, S.A., Bush, K.T., Sampogna, R.V., et al., 2004. The molecular pharmacology of organic anion transporters: from DNA to FDA? Mol. Pharmacol. 65, 479–487. (*Reviews aspects of the molecular biology and pharmacology of the organic anion transporters, and discusses their structural biology, paired genomic organisation, developmental regulation, toxicology and pharmacogenetics*)

Hediger, M.A., Romero, M.F., Peng, J.-B., et al., 2004. The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins. Pflug. Arch. 447, 465–468.

Koepsell, H., 2004. Polyspecific organic cation transporters: their functions and interactions with drugs. Trends Pharmacol. Sci. 25, 375–381. (*Reviews organic cation transporters [OCT]1–3, which are expressed in gut, liver, kidney, heart, placenta, lung and brain, and facilitate diffusion of structurally diverse organic cations including monoamine neurotransmitters and many drugs; studies in knockout mice implicate OCT1 in the hepatic uptake and biliary excretion of cationic drugs, and OCT1 and 2 in renal proximal tubules participate in secreting cationic drugs into urine*)

McNamara, P.J., Abbassi, M., 2004. Neonatal exposure to drugs in breast milk. Pharm. Res. 21, 555–566. (*Review*)

Neff, M.W., Robertson, K.R., Wong, A.K., et al., 2004. Breed distribution and history of canine mdr1-1 delta, a pharmacogenetic mutation that marks the emergence of breeds from the collie lineage. Proc. Natl. Acad. Sci. USA 101, 11725–11730. (*The breed distribution and frequency of mdr1-1 delta have applications in veterinary medicine, whereas the allele's history recounts the emergence of formally recognised breeds from an admixed population of working sheepdogs*)

Petzinger, E., Geyer, J., 2006. Drug transporters in pharmacokinetics. Naunyn-Schmiederberg's Arch. Pharmacol. 372, 465–475. (*Emphasises the interplay between drug metabolism and drug transport, especially in liver*)

Ritter, C.A., Jedlitschky, G., Schwabedissen, H.M.Z., et al., 2005. Cellular export of drugs and signaling molecules by the ATP-binding cassette transporters MRP4 (ABCC4) and MRP5 (ABCC5). Drug Metab. Rev. 37, 253–278. (*Members of the multidrug resistance-associated protein [MRP] subfamily of ATP-binding cassette transporters, MRP4 and 5 are organic anion transporters; they transport nucleotides and nucleotide analogues, and also cyclic nucleotides, so are implicated in signal transduction. MRP4 also transports conjugated steroids, prostaglandins and glutathione*)

Sasaki, M., Suzuki, H., Aoki, J., et al., 2004. Prediction of in vivo biliary clearance from the in vitro transcellular transport of organic anions across a double-transfected Madin–Darby canine kidney II monolayer expressing both rat organic anion transporting polypeptide 4 and multidrug resistance associated protein 2. Mol. Pharmacol. 66, 450–459. (*Double-transfected Madin–Darby canine kidney cell monolayer may be useful in analysing hepatic transport of organic anions and in predicting in vivo biliary clearance*)

Drug delivery

Cornford, E.M., Cornford, M.E., 2002. New systems for delivery of drugs to the brain in neurological disease. Lancet Neurol. 1, 306–315. (*Reviews augmentation of pinocytosis to deliver drugs to the brain. Macromolecules can be conjugated to peptidomimetic ligands that bind peptide receptors, and are then internalised and transported in small vesicles across the cytoplasmic brain–capillary barrier. Such conjugates can remain effective in animal models of neurological disease)*

Mahato, R.I., Narang, A.S., Thoma, L., Miller, D.D., 2003. Emerging trends in oral delivery of peptide and protein drugs. Crit. Rev. Ther. Drug Carrier Syst. 20, 153–214. (*Various strategies currently under investigation include amino acid backbone modifications; formulation approaches; chemical conjugation of hydrophobic or targeting ligand; and use of enzyme inhibitors, mucoadhesive polymers and absorption enhancers*)

Mizuno, N., Niwa, T., Yotsumoto, Y., Sugiyama, Y., 2003. Impact of drug transporter studies on drug discovery and development. Pharmacol. Rev. 55, 425–461. (*Reviews drug transport in intestine, liver, kidney and brain, and its roles in absorption, distribution and excretion*)

Singh, R., Lillard, J.W., 2008. Nanoparticle-based targeted drug delivery. Exp. Mol. Pathol. 86, 215–223.

Taguchi, A., Sharma, N., Saleem, R.M., 2001. Selective postoperative inhibition of gastrointestinal opioid receptors. N. Engl. J. Med. 345, 935–940. (*Speeds recovery of bowel function and shortens hospitalisation: notionally 'poor' absorption is used to advantage by providing a selective action on the gut*)

Varde, N.K., Pack, D.W., 2004. Microspheres for controlled release drug delivery. Exp. Opin. Biol. Ther. 4, 35–51. (*Describes methods of microparticle fabrication and factors controlling the release rates of encapsulated drugs; recent advances for delivery of single-shot vaccines, plasmid DNA and therapeutic proteins are discussed*)

9 Drug metabolism and elimination

OVERVIEW

We describe phases I and II of drug metabolism, emphasising the importance of the cytochrome P450 monooxygenase system. We then cover the processes of biliary excretion and enterohepatic recirculation of drugs, and of drug and drug metabolite elimination by the kidney.

INTRODUCTION

Drug elimination is the irreversible loss of drug from the body. It occurs by two processes: *metabolism* and *excretion*. Metabolism consists of anabolism and catabolism, i.e. respectively the build-up and breakdown of substances by enzymic conversion of one chemical entity to another within the body, whereas excretion consists of elimination from the body of chemically unchanged drug or its metabolites. The main routes by which drugs and their metabolites leave the body are:

- the kidneys
- the hepatobiliary system
- the lungs (important for volatile/gaseous anaesthetics).

Most drugs leave the body in the urine, either unchanged or as polar metabolites. Some drugs are secreted into bile via the liver, but most of these are then reabsorbed from the intestine. There are, however, instances (e.g. **rifampicin**; Ch. 50) where faecal loss accounts for the elimination of a substantial fraction of unchanged drug in healthy individuals, and faecal elimination of drugs such as **digoxin** that are normally excreted in urine (Ch. 21) becomes progressively more important in patients with advancing renal failure. Excretion via the lungs occurs only with highly volatile or gaseous agents (e.g. general anaesthetics; Ch. 40). Small amounts of some drugs are also excreted in secretions such as milk or sweat. Elimination by these routes is quantitatively negligible compared with renal excretion, although excretion into milk can sometimes be important because of effects on the baby (e.g. see McNamara & Abbassi, 2004; Ito, 2000).

Lipophilic substances are not eliminated efficiently by the kidney. Consequently, most lipophilic drugs are metabolised to more polar products, which are then excreted in urine. Drug metabolism occurs predominantly in the liver, especially by the cytochrome P450 (CYP) system. Some P450 enzymes are extrahepatic and play an important part in the biosynthesis of steroid hormones (Ch. 32) and eicosanoids (Ch. 17), but here we are concerned with catabolism of drugs by the hepatic P450 system.

DRUG METABOLISM

Animals have evolved complex systems that detoxify foreign chemicals ('xenobiotics'), including carcinogens and toxins present in poisonous plants. Drugs are a special case of such xenobiotics and, like plant alkaloids, they often exhibit *chirality* (i.e. there is more than one stereoisomer), which affects their overall metabolism. Drug metabolism involves two kinds of reaction, known as phase 1 and phase 2. These often, although not invariably, occur sequentially. Both phases decrease lipid solubility, thus increasing renal elimination.

PHASE 1 REACTIONS

Phase 1 reactions are catabolic (e.g. oxidation, reduction or hydrolysis), and the products are often more chemically reactive and hence, paradoxically, sometimes more toxic or carcinogenic than the parent drug. Phase 1 reactions often introduce a reactive group, such as hydroxyl, into the molecule, a process known as 'functionalisation'. This group then serves as the point of attack for the conjugating system to attach a substituent such as glucuronide (Fig. 9.1), explaining why phase 1 reactions so often precede phase 2 reactions (see below). Phase 1 reactions take place mainly in the liver. Many hepatic drug-metabolising enzymes, including CYP enzymes, are embedded in the smooth endoplasmic reticulum. They are often called 'microsomal' enzymes because, on homogenisation and differential centrifugation, the endoplasmic reticulum is broken into very small fragments that sediment only after prolonged high-speed centrifugation in the microsomal fraction. To reach these metabolising enzymes in life, a drug must cross the plasma membrane. Polar molecules do this less readily than non-polar molecules except where there are specific transport mechanisms (Ch. 8), so intracellular metabolism is important for lipid-soluble drugs, while polar drugs are at least partly excreted unchanged in the urine.

THE P450 MONOOXYGENASE SYSTEM

Nature, classification and mechanism of P450 enzymes

Cytochrome P450 enzymes are haem proteins, comprising a large family ('superfamily') of related but distinct enzymes, each referred to as CYP followed by a defining set of numbers and a letter. These enzymes differ from one another in amino acid sequence, in sensitivity to inhibitors and inducing agents (see below), and in the specificity of the reactions that they catalyse (see Anzenbacher, 2007 for reviews). Different members of the family have distinct, but often overlapping, substrate specificities, and may act on the same substrates but at different rates. Purification of P450 enzymes and complementary DNA cloning form the basis of the current classification, which is based on amino acid sequence similarities. Seventy-four CYP gene families have been described, of which three main ones (CYP1, CYP2 and CYP3) are involved in drug metabolism in human liver. Examples of therapeutic drugs that are substrates for some important P450 isoenzymes are shown in Table 9.1. Drug oxidation by the monooxygenase P450

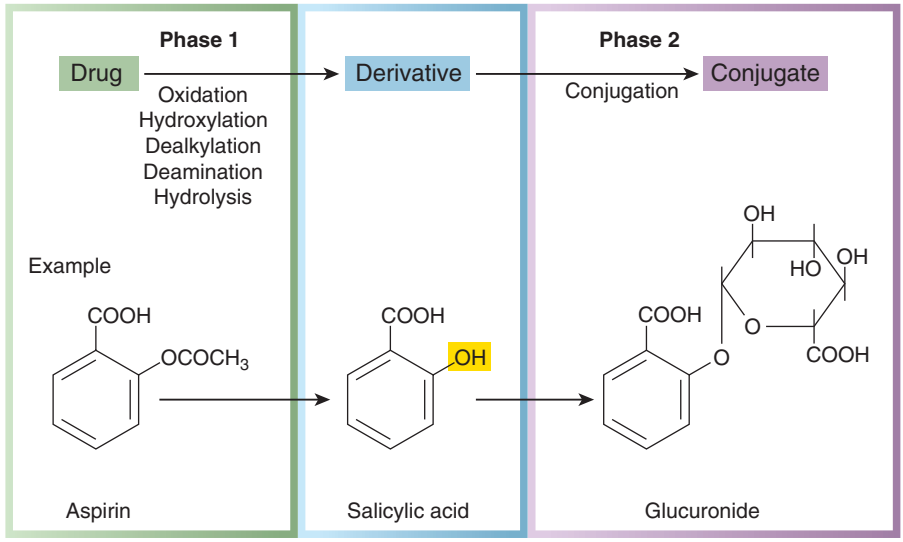

Fig. 9.1 The two phases of drug metabolism.

Table 9.1 Examples of drugs that are substrates of P450 isoenzymes

Isoenzyme P450	Drug(s)
CYP1A2	Caffeine, paracetamol (→NAPQI), tacrine, theophylline
CYP2B6	Cyclophosphamide, methadone
CYP2C8	Paclitaxel, repaglinide
CYP2C19	Omeprazole, phenytoin
CYP2C9	Ibuprofen, tolbutamide, warfarin
CYP2D6	Codeine, debrisoquine, S-metoprolol
CYP2E1	Alcohol, paracetamol
CYP3A4, 5, 7	Ciclosporin, nifedipine, indinavir, simvastatin

(Adapted from http://medicine.iupui.edu/flockhart/table.htm.)

system requires drug (substrate, 'DH'), P450 enzyme, molecular oxygen, NADPH and a flavoprotein (NADPH–P450 reductase). The mechanism involves a complex cycle (Fig. 9.2), but the overall net effect of the reaction is quite simple, namely the addition of one atom of oxygen (from molecular oxygen) to the drug to form a hydroxyl group (product, 'DOH'), the other atom of oxygen being converted to water.

▼ P450 enzymes have unique spectral properties, and the reduced forms combine with carbon monoxide to form a pink compound (hence 'P') with absorption peaks near 450 nm (range 447–452 nm). The first clue that there is more than one form of CYP came from the observation that treatment of rats with 3-methylcholanthrene (3-MC), an inducing agent (see below), causes a shift in the absorption maximum from 450 to 448 nm—the 3-MC-induced isoform of the enzyme absorbs light maximally at a slightly shorter wavelength than the un-induced enzyme.

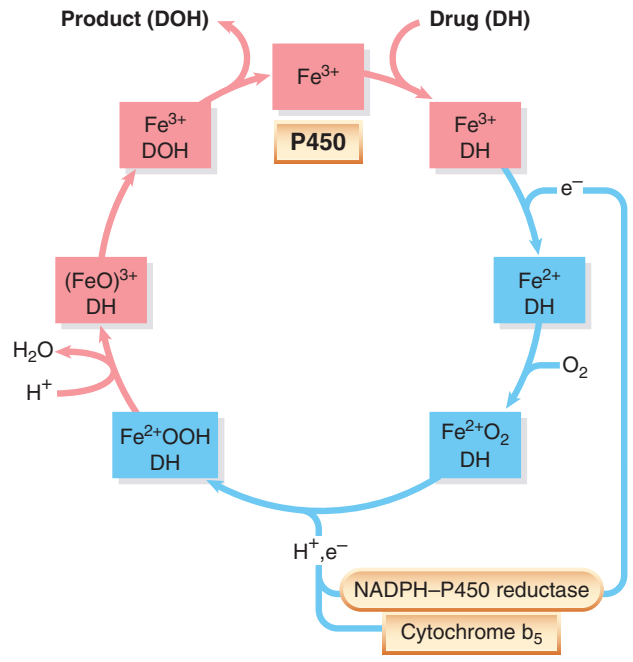

Fig. 9.2 The monooxygenase P450 cycle. Each of the pink or blue rectangles represents one single molecule of cytochrome P450 (P450) undergoing a catalytic cycle. Iron in P450 is in either the ferric (pink rectangles) or ferrous (blue rectangles) state. P450 containing ferric iron (Fe^{3+}) combines with a molecule of drug ('DH'); receives an electron from NADPH–P450 reductase, which reduces the iron to Fe^{2+}; combines with molecular oxygen, a proton and a second electron (either from NADPH–P450 reductase or from cytochrome b_5) to form an $Fe^{2+}OOH$–DH complex. This combines with another proton to yield water and a ferric oxene $(FeO)^{3+}$–DH complex. $(FeO)^{3+}$ extracts a hydrogen atom from DH, with the formation of a pair of short-lived free radicals (see text), liberation from the complex of oxidised drug ('DOH'), and regeneration of P450 enzyme.

P450 and biological variation

There are important variations in the expression and regulation of P450 enzymes between species. For instance, the pathways by which certain dietary heterocyclic amines (formed when meat is cooked) generate genotoxic products involves one member of the P450 superfamily (CYP1A2) that is constitutively present in humans and rats (which develop colon tumours after treatment with such amines) but not in cynomolgus monkeys (which do not). Such species differences have crucial implications for the choice of species to be used for toxicity and carcinogenicity testing during the development of new drugs for use in humans.

Within human populations, there are major sources of interindividual variation in P450 enzymes that are of great importance in therapeutics. These include genetic polymorphisms (alternative sequences at a locus within the DNA strand—alleles—that persist in a population through several generations; Ch. 11). Environmental factors (Ch. 56) are also important, since enzyme inhibitors and inducers are present in the diet and environment. For example, a component of grapefruit juice inhibits drug metabolism (leading to potentially disastrous consequences, including cardiac dysrhythmias; Ch. 56), whereas Brussels sprouts and cigarette smoke induce P450 enzymes. Components of St John's wort (used to treat depression in 'alternative' medicine; Ch. 46) induce CYP450 isoenzymes as well as P-glycoprotein (P-gp) (see Ch. 8 and below, and Henderson et al., 2002).

Not all drug oxidation reactions involve the P450 system: some drugs are metabolised in plasma (e.g. hydrolysis of **suxamethonium** by plasma cholinesterase; Ch. 13), lung (e.g. various prostanoids; Ch. 17) or gut (e.g. **tyramine**, **salbutamol**; Chs 14 and 27). **Ethanol** (Ch. 48) is metabolised by a soluble cytoplasmic enzyme, alcohol dehydrogenase, in addition to CYP2E1. Other P450-independent enzymes involved in drug oxidation include xanthine oxidase, which inactivates **6-mercaptopurine** (Ch. 55), and monoamine oxidase, which inactivates many biologically active amines (e.g. **noradrenaline** [norepinephrine], tyramine, 5-hydroxytryptamine; Chs 14 and 15).

Hydrolytic reactions (e.g. of **aspirin**; Fig. 9.1) do not involve hepatic microsomal enzymes but occur in plasma and in many tissues. Both ester and (less readily) amide bonds are susceptible to hydrolysis. Reductive reactions are much less common than oxidations, but some are important. For example, **warfarin** (Ch. 24) is inactivated by conversion of a ketone to a hydroxyl group by CYP2A6.

PHASE 2 REACTIONS

Phase 2 reactions are synthetic ('anabolic') and involve conjugation (i.e. attachment of a substituent group), which usually results in inactive products, although there are exceptions (e.g. the active sulfate metabolite of **minoxidil**, a potassium channel activator used to treat severe hypertension, Ch. 22; **morphine-6-glucuronide** is an active metabolite of morphine that is being developed as an analgesic agent [Ch. 41]—on acute administration it induces less nausea and vomiting than the parent drug perhaps because, being more polar, it fails to access the vomiting centres). Phase 2 reactions also take place mainly in the liver. If a drug molecule has a suitable 'handle' (e.g. a hydroxyl, thiol or amino group), either in the parent molecule or in a product resulting from phase 1 metabolism, it is susceptible to conjugation. The groups most often involved are glucuronyl (Fig. 9.3), sulfate, methyl and acetyl. The tripeptide glutathione can conjugate drugs or their phase 1 metabolites via its sulfhydryl group, as in the detoxification of **paracetamol** (see Fig. 57.1, p. 701). Glucuronide formation involves the formation of a high-energy phosphate compound, uridine diphosphate glucuronic acid (UDPGA), from which glucuronic acid is transferred to an electron-rich atom (N, O or S) on the substrate, forming an amide, ester or thiol bond. UDP-glucuronyl transferase, which catalyses these reactions, has very broad substrate specificity embracing many drugs and other foreign molecules. Several important endogenous substances, including bilirubin and adrenal corticosteroids, are conjugated by the same system.

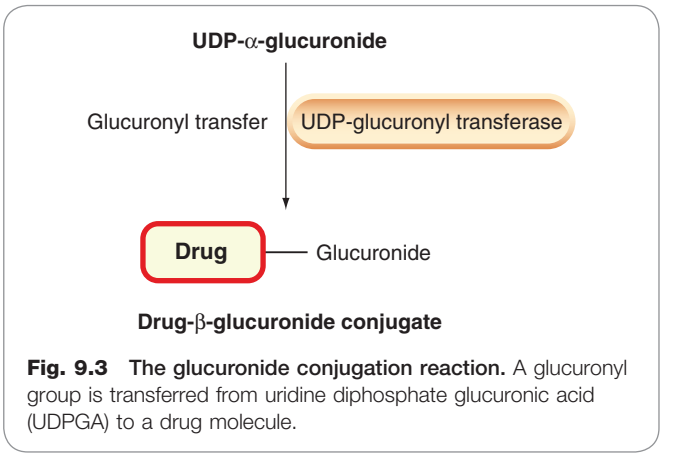

Fig. 9.3 The glucuronide conjugation reaction. A glucuronyl group is transferred from uridine diphosphate glucuronic acid (UDPGA) to a drug molecule.

Acetylation and methylation reactions occur with acetyl-CoA and S-adenosyl methionine, respectively, acting as the donor compounds. Many of these conjugation reactions occur in the liver, but other tissues, such as lung and kidney, are also involved.

STEREOSELECTIVITY

Many clinically important drugs, such as **sotalol** (Ch. 21), **warfarin** (Ch. 24) and **cyclophosphamide** (Ch. 55), are mixtures of stereoisomers, the components of which differ not only in their pharmacological effects but also in their metabolism, which may follow completely distinct pathways (see Campo et al., 2009 for a recent review). Several clinically important drug interactions involve stereospecific inhibition of metabolism of one drug by another (Ch. 56). In some cases, drug toxicity is mainly linked to one of the stereoisomers, not necessarily the pharmacologically active one. Where practicable, regulatory authorities urge that new drugs should consist of single isomers to avoid these complications.[1]

INHIBITION OF P450

Inhibitors of P450 differ in their selectivity towards different isoforms of the enzyme, and are classified by their mechanism of action. Some drugs compete for the active

[1]No doubt a good idea though the usefulness of effort directed towards developing 'novel' entities that are actually just the active isomers of well-established and safe racemates has been questioned.

site but are not themselves substrates (e.g. **quinidine** is a potent competitive inhibitor of CYP2D6 but is not a substrate for it). Non-competitive inhibitors include drugs such as **ketoconazole**, which forms a tight complex with the Fe^{3+} form of the haem iron of CYP3A4, causing reversible non-competitive inhibition. So-called mechanism-based inhibitors require oxidation by a P450 enzyme. Examples include the oral contraceptive **gestodene** (CYP3A4) and the anthelminthic drug **diethylcarbamazine** (CYP2E1). An oxidation product (e.g. a postulated epoxide intermediate of gestodene) binds covalently to the enzyme, which then destroys itself ('suicide inhibition'; see Pelkonen et al., 2008 for a fuller review). Many clinically important interactions between drugs are the result of inhibition of P450 enzymes (see Ch. 56).

INDUCTION OF MICROSOMAL ENZYMES

A number of drugs, such as **rifampicin** (Ch. 50), **ethanol** (Ch. 48) and **carbamazepine** (Ch. 44), increase the activity of microsomal oxidase and conjugating systems when administered repeatedly. Many carcinogenic chemicals (e.g. benzpyrene, 3-MC) also have this effect, which can be substantial; Figure 9.4 shows a nearly 10-fold increase in the rate of benzpyrene metabolism 2 days after a single dose. The effect is referred to as *induction*, and is the result of increased synthesis and/or reduced breakdown of microsomal enzymes—see Park et al. (1996), Dickins (2004) and Pelkonen et al. (2008) for more detail.

Enzyme induction can increase drug toxicity and carcinogenicity (Park et al., 2005), because several phase 1 metabolites are toxic or carcinogenic: paracetamol is an important example of a drug with a highly toxic metabolite (see Ch. 57).

The mechanism of induction is incompletely understood but is similar to that involved in the action of steroid and other hormones that bind to nuclear receptors (see Ch. 3). The most thoroughly studied inducing agents are polycyclic aromatic hydrocarbons (e.g. 3-MC). These bind to the ligand-binding domain of a soluble protein, termed the aromatic hydrocarbon (Ah) receptor. This complex is transported to the nucleus by an Ah receptor nuclear translocator and binds Ah receptor response elements in the DNA, thereby promoting transcription of the gene CYP1A1. In addition to enhanced transcription, some inducing agents (e.g. ethanol, which induces CYP2E1 in humans) also stabilise mRNA or P450 protein.

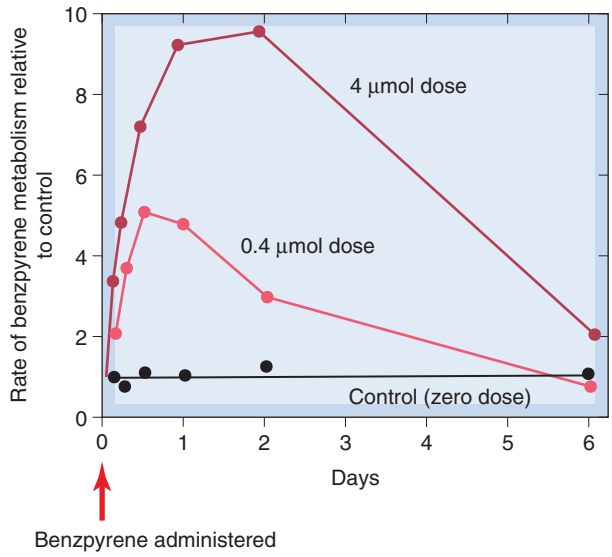

Fig. 9.4 Stimulation of hepatic metabolism of benzpyrene. Young rats were given benzpyrene (intraperitoneally) in the doses shown, and the benzpyrene-metabolising activity of liver homogenates was measured at times up to 6 days. (From Conney A H et al. 1957 J Biol Chem 228: 753.)

FIRST-PASS (PRESYSTEMIC) METABOLISM

Some drugs are extracted so efficiently by the liver or gut wall that the amount reaching the systemic circulation is considerably less than the amount absorbed. This is known as first-pass or presystemic metabolism and reduces bioavailability (Ch. 8) even when a drug is well absorbed. Presystemic metabolism is important for many therapeutic drugs (Table 9.2 shows some examples), and is a problem because:

- a much larger dose of the drug is needed when it is given orally than when it is given parenterally
- marked individual variations occur in the extent of first-pass metabolism (see Ch. 56).

PHARMACOLOGICALLY ACTIVE DRUG METABOLITES

In some cases (see Table 9.3), a drug becomes pharmacologically active only after it has been metabolised. For example, **azathioprine**, an immunosuppressant drug (Ch. 26), is metabolised to **mercaptopurine**; and **enalapril**, an angiotensin-converting enzyme inhibitor (Ch. 22), is hydrolysed to its active form **enalaprilat**. Such drugs, in which the parent compound lacks activity of its own, are known as *prodrugs*. These are sometimes designed deliberately to overcome problems of drug delivery (Ch. 8). Metabolism can alter the pharmacological actions of a drug qualitatively. **Aspirin** inhibits some platelet functions and has anti-inflammatory activity (Chs 24 and 26). It is hydrolysed to salicylic acid (Fig. 9.1), which has anti-inflammatory but not antiplatelet activity. In other instances, metabolites have pharmacological actions similar to those of the parent compound (e.g. benzodiazepines, many of which form long-lived active metabolites that cause sedation to persist after the parent drug has disappeared; Ch. 43). There are also cases in which metabolites are responsible for toxicity.

Table 9.2 Examples of drugs that undergo substantial first-pass elimination

Aspirin	Metoprolol
Glyceryl trinitrate	Morphine
Isosorbide dinitrate	Propranolol
Levodopa	Salbutamol
Lidocaine	Verapamil

Table 9.3 Some drugs that produce active or toxic metabolites

Inactive (prodrugs)	Active drug	Active metabolite	Toxic metabolite	See Chapter
Azathioprine →		Mercaptopurine		26
Cortisone →		Hydrocortisone		32
Prednisone →		Prednisolone		32
Enalapril →		Enalaprilat		22
Zidovudine →		Zidovudine trisphosphate		51
Cyclophosphamide →		Phosphoramide mustard →	Acrolein	55
	Diazepam →	Nordiazepam → Oxazepam		43
	Morphine →	Morphine 6-glucuronide		41
	Halothane →		Trifluoroacetic acid	40
	Methoxyflurane →		Fluoride	40
	Paracetamol →		*N*-Acetyl-*p*-benzoquinone imine	26, 57

Hepatotoxicity of **paracetamol** is one example (see Ch. 57), and bladder toxicity of **cyclophosphamide**, which is caused by its toxic metabolite acrolein (Ch. 55), is another. Methanol and ethylene glycol both exert their toxic effects via metabolites formed by alcohol dehydrogenase. Poisoning with these agents is treated with ethanol (or with a more potent inhibitor), which competes for the active site of the enzyme. **Disulfiram** inhibits CYP2E1 and reduces substantially the formation of trifluoroacetic acid during halothane anaesthesia, raising the intriguing possibility that it could prevent halothane hepatitis (see Kharasch, 2008).

Drug metabolism

- Phase 1 reactions involve oxidation, reduction and hydrolysis. They:
  - usually form more chemically reactive products, which can be pharmacologically active, toxic or carcinogenic
  - often involve a monooxygenase system in which cytochrome P450 plays a key role.
- Phase 2 reactions involve conjugation (e.g. glucuronidation) of a reactive group (often inserted during phase 1 reaction) and usually lead to inactive and polar products that are readily excreted.
- Some conjugated products are excreted via bile, are reactivated in the intestine and then reabsorbed ('enterohepatic circulation').
- Induction of P450 enzymes can greatly accelerate hepatic drug metabolism. It can increase the toxicity of drugs with toxic metabolites.
- Presystemic metabolism in liver or gut wall reduces the bioavailability of several drugs when they are administered by mouth.

DRUG AND METABOLITE EXCRETION

BILIARY EXCRETION AND ENTEROHEPATIC CIRCULATION

Liver cells transfer various substances, including drugs, from plasma to bile by means of transport systems similar to those of the renal tubule including organic cation transporters (OCTs), organic anion transporters (OATs) and P-glycoproteins (P-gp) (see Ch. 8). Various hydrophilic drug conjugates (particularly glucuronides) are concentrated in bile and delivered to the intestine, where the glucuronide is usually hydrolysed, releasing active drug once more; free drug can then be reabsorbed and the cycle repeated (*enterohepatic circulation*). The effect of this is to create a 'reservoir' of recirculating drug that can amount to about 20% of total drug in the body and prolongs drug action. Examples where this is important include **morphine** (Ch. 41) and **ethinylestradiol** (Ch. 34). Several drugs are excreted to an appreciable extent in bile. **Vecuronium** (a non-depolarising muscle relaxant; Ch. 13) is an example of a drug that is excreted mainly unchanged in bile. **Rifampicin** (Ch. 50) is absorbed from the gut and slowly deacetylated, retaining its biological activity. Both forms are secreted in the bile, but the deacetylated form is not reabsorbed, so eventually most of the drug leaves the body in this form in the faeces.

RENAL EXCRETION OF DRUGS AND METABOLITES

Drugs differ greatly in the rate at which they are excreted by the kidney, ranging from **penicillin** (Ch. 50), which is cleared from the blood almost completely on a single transit through the kidney, to **diazepam** (Ch. 43), which is cleared extremely slowly. Most drugs fall between these extremes, and metabolites are nearly always cleared more quickly than the parent drug. Three fundamental processes account for renal drug excretion:

1. glomerular filtration
2. active tubular secretion
3. passive diffusion across tubular epithelium.

GLOMERULAR FILTRATION

Glomerular capillaries allow drug molecules of molecular weight below about 20 000 to pass into the glomerular filtrate. Plasma albumin (molecular weight approximately 68 000) is almost completely impermeant, but most drugs—with the exception of macromolecules such as **heparin** (Ch. 24) or biological products (Ch. 59)—cross the barrier freely. If a drug binds to plasma albumin, only free drug is filtered. If, like **warfarin** (Ch. 24), a drug is approximately 98% bound to albumin, the concentration in the filtrate is only 2% of that in plasma, and clearance by filtration is correspondingly reduced.

TUBULAR SECRETION

Up to 20% of renal plasma flow is filtered through the glomerulus, leaving at least 80% of delivered drug to pass on to the peritubular capillaries of the proximal tubule. Here, drug molecules are transferred to the tubular lumen by two independent and relatively non-selective carrier systems (see Ch. 8). One of these, the OAT, transports acidic drugs (as well as various endogenous acids, such as uric acid), while an OCT handles organic bases. Some important drugs that are transported by these two carrier systems are shown in Table 9.4. The OAT carrier can transport drug molecules against an electrochemical gradient, and can therefore reduce the plasma concentration nearly to zero, whereas OCT facilitates transport down an electrochemical gradient. Because at least 80% of the drug delivered to the kidney is presented to the carrier, tubular secretion is potentially the most effective mechanism of renal drug elimination. Unlike glomerular filtration, carrier-mediated transport can achieve maximal drug clearance even when most of the drug is bound to plasma protein.[2] **Penicillin** (Ch. 50), for example, although about 80% protein bound and therefore cleared only slowly by filtration, is almost completely removed by proximal tubular secretion, and is therefore rapidly eliminated.

Many drugs compete for the same transport system (Table 9.4), leading to drug interactions. For example, **probenecid** was developed originally to prolong the action of penicillin by retarding its tubular secretion.

DIFFUSION ACROSS THE RENAL TUBULE

Water is reabsorbed as fluid traverses the tubule, the volume of urine emerging being only about 1% of that of the glomerular filtrate. Consequently, if the tubule is freely permeable to drug molecules, some 99% of the filtered drug will be reabsorbed passively down the resulting concentration gradient. Lipid-soluble drugs are therefore excreted poorly, whereas polar drugs of low tubular permeability remain in the lumen and become progressively concentrated as water is reabsorbed. Polar drugs handled in this way include **digoxin** and *aminoglycoside antibiotics*. These exemplify a relatively small but important group of drugs (Table 9.5) that are not inactivated by metabolism, the rate of renal elimination being the main factor that determines their duration of action. These drugs have to be used with special care in individuals whose renal function may be impaired, including the elderly and patients with renal disease or any severe acute illness (Ch. 56).

The degree of ionization of many drugs—weak acids or weak bases—is pH dependent, and this markedly influences their renal excretion. The ion-trapping effect means that a basic drug is more rapidly excreted in an acid urine which favours the charged form and thus inhibits reabsorption. Conversely, acidic drugs are most rapidly excreted if the urine is alkaline (Fig. 9.5). Urinary alkalinisation is used to accelerate the excretion of salicylate in treating selected cases of aspirin overdose.

Table 9.4 Important drugs and related substances secreted into the proximal renal tubule by OAT or OCT transporters

OAT	OCT
p-Aminohippuric acid	Amiloride
Furosemide	Dopamine
Glucuronic acid conjugates	Histamine
Glycine conjugates	Mepacrine
Indometacin	Morphine
Methotrexate	Pethidine
Penicillin	Quaternary ammonium compounds
Probenecid	Quinine
Sulfate conjugates	5-Hydroxytryptamine (serotonin)
Thiazide diuretics	Triamterene
Uric acid	

Table 9.5 Examples of drugs that are excreted largely unchanged in the urine

Percentage	Drugs excreted
100–75	Furosemide, gentamicin, methotrexate, atenolol, digoxin
75–50	Benzylpenicillin, cimetidine, oxytetracycline, neostigmine
~50	Propantheline, tubocurarine

[2]Because filtration involves isosmotic movement of both water and solutes, it does not affect the free concentration of drug in the plasma. Thus the equilibrium between free and bound drug is not disturbed, and there is no tendency for bound drug to dissociate as blood traverses the glomerular capillary. The rate of clearance of a drug by filtration is therefore reduced directly in proportion to the fraction that is bound. In the case of active tubular secretion, this is not so; secretion may be retarded very little even though the drug is mostly bound. This is because the carrier transports drug molecules unaccompanied by water. As free drug molecules are taken from the plasma, therefore, the free plasma concentration falls, causing dissociation of bound drug from plasma albumin. Consequently, effectively 100% of the drug, bound and free, is available to the carrier.

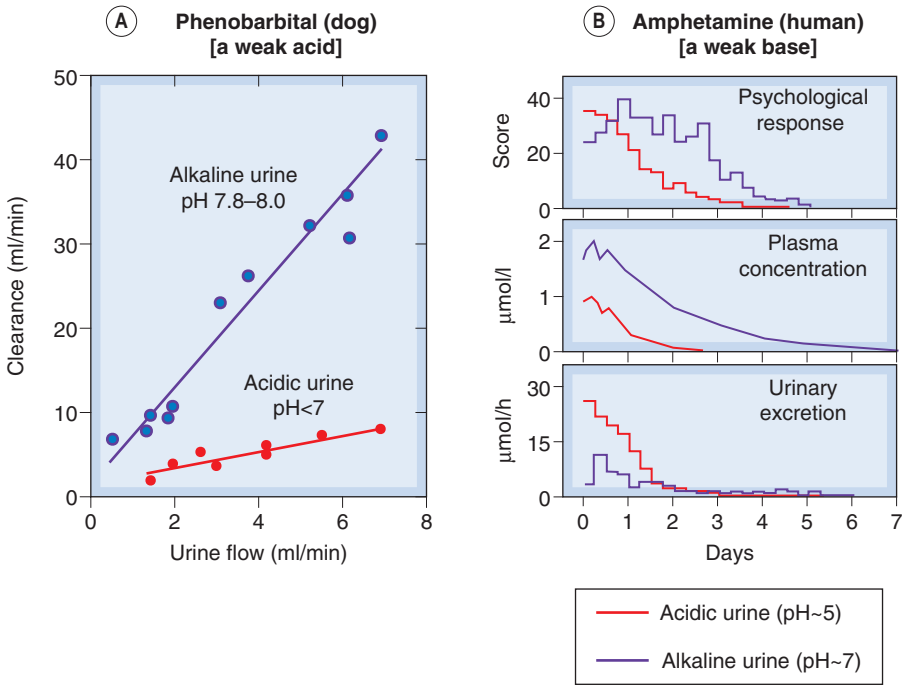

Fig. 9.5 The effect of urinary pH on drug excretion. [**A**] Phenobarbital clearance in the dog as a function of urine flow. Because phenobarbital is acidic, alkalinising the urine increases clearance about five-fold. [**B**] Amphetamine excretion in humans. Acidifying the urine increases the rate of renal elimination of amphetamine, reducing its plasma concentration and its effect on the subject's mental state. (Data from Gunne & Anggard 1974. In: Torrell T et al. (eds) Pharmacology and pharmacokinetics. Plenum, New York.)

RENAL CLEARANCE

Elimination of drugs by the kidneys is best quantified by the renal clearance (CL_r). This is defined as the volume of plasma containing the amount of substance that is removed from the body by the kidneys in unit time. It is calculated from the plasma concentration, C_p, the urinary concentration, C_u, and the rate of flow of urine, V_u, by the equation:

$$CL_r = \frac{C_u \times V_u}{C_p}$$

CL_r varies greatly for different drugs, from less than 1 ml/min to the theoretical maximum set by the renal plasma flow, which is approximately 700 ml/min, measured by *p*-aminohippuric acid (PAH) clearance (renal extraction of PAH approaches 100%).

Elimination of drugs by the kidney

- Most drugs, unless highly bound to plasma protein, cross the glomerular filter freely.
- Many drugs, especially weak acids and weak bases, are actively secreted into the renal tubule and thus more rapidly excreted.
- Lipid-soluble drugs are passively reabsorbed by diffusion across the tubule, so are not efficiently excreted in the urine.
- Because of pH partition, weak acids are more rapidly excreted in alkaline urine, and vice versa.
- Several important drugs are removed predominantly by renal excretion, and are liable to cause toxicity in elderly persons and patients with renal disease.

REFERENCES AND FURTHER READING

General further reading

Nassar, A.F., 2009. Drug Metabolism Handbook: Concepts and Applications. Wiley-Blackwell, Hoboken, NJ. (*Multi-authored handbook aimed at bench scientists; will be invaluable for pharmaceutical industry scientists*)

Testa, B., Krämer, S.D., 2009. The biochemistry of drug metabolism. Wiley-VCH, Weinheim. (*Two-volume reference work*)

Drug metabolism

Anzenbacher, P. (Ed.), 2007. Special issue: cytochrome P450. BBA general subjects 1770 (3), 313–494. (*Contains sections on mechanisms and principles; structural insights and biophysics; substrates, tissue specificities and regulation; clinical implications*)

Campo, V.L., Bernardes, L.S.C., Carvalho, I., 2009. Stereoselectivity in drug metabolism: molecular mechansisms and analytical methods. Curr. Drug Metab. 10, 188–205.

Coon, M.J., 2005. Cytochrome P450: nature's most versatile biological catalyst. Annu. Rev. Pharmacol. Toxicol. 45, 1–25. (*Summarises the individual steps in the P450 and reductase reaction cycles*)

Kharasch, E.D., 2008. Adverse drugs reactions with halogenated anesthetics. Clin. Pharmacol. Ther. 84, 158–162. (*This review focuses on adverse effects that are attributable to anesthetic metabolism*)

Kinirons, M.T., O'Mahony, M.S., 2004. Drug metabolism and ageing. Br. J. Clin. Pharmacol. 57, 540–544. (*Reviews age-related changes in drug metabolism*)

P450 enzyme induction and inhibition

Dickins, M., 2004. Induction of cytochromes P450. Curr. Top. Med. Chem. 4, 1745–1766. (*Recent advances*)

Henderson, L., Yue, Q.Y., Bergquist, C., et al., 2002. St John's wort (*Hypericum perforatum*): drug interactions and clinical outcomes. Br. J. Clin. Phar-

macol. 54, 349–356. (*Reviews the induction of CYP450 isoenzymes and of P-glycoprotein by constituents in this herbal remedy*)

Pelkonen, O., Turpeinen, M., Hakkola, J. et al., 2008. Inhibition and induction of human cytochrome P450 enzymes: current status. Arch. Toxicol. 82, 667–715. (*Review*)

Drug elimination

Ito, S., 2000. Drug therapy: drug therapy for breast-feeding women. N. Engl. J. Med. 343, 118–126

Keppler, D., Konig, J., 2000. Hepatic secretion of conjugated drugs and endogenous substances. Semin. Liver Dis. 20, 265–272. (*'Conjugate export pumps of the multidrug resistance protein-MRP-family mediate ATP-dependent secretion of anionic conjugates across the canalicular and the basolateral hepatocyte membrane into bile and sinusoidal blood, respectively. Xenobiotic and endogenous lipophilic substances may be conjugated with glutathione, glucuronate, sulfate, or other negatively charged groups and thus become substrates for export pumps of the MRP family'*)

Kusuhara, H., Sugiyama, Y., 2009. In vitro–in vivo extrapolation of transporter-mediated clearance in the liver and kidney. Drug Metab. Pharmacokinet. 24, 37–52. (*Review*)

10 Pharmacokinetics

OVERVIEW

We explain the importance of pharmacokinetic analysis and present a simple approach to this. We explain how drug clearance determines the steady-state plasma concentration during constant-rate drug administration and how the characteristics of absorption and distribution (considered in Ch. 8) plus metabolism and excretion (considered in Ch. 9) determine the time course of drug concentration in blood plasma during and following drug administration. The effect of different dosing regimens on the time course of drug concentration in plasma is explained. Population pharmacokinetics is mentioned briefly, and a final section considers limitations to the pharmacokinetic approach.

INTRODUCTION: DEFINITION AND USES OF PHARMACOKINETICS

Pharmacokinetics may be defined as the measurement and formal interpretation of changes with time of drug concentrations in one or more different regions of the body in relation to dosing ('what the body does to the drug'). This distinguishes it from pharmacodynamics ('what the drug does to the body', i.e. events consequent on interaction of the drug with its receptor or other primary site of action). The distinction is useful, although the words cause dismay to etymological purists. 'Pharmacodynamic' received an entry in a dictionary of 1890 ('relating to the powers or effects of drugs') whereas pharmacokinetic studies only became possible with the development of sensitive, specific and accurate physicochemical analytical techniques, especially chromatography and mass spectrometry, for measuring drug concentrations in biological fluids in the latter part of the 20th century. The time course of drug concentration following dosing depends on the processes of absorption, distribution, metabolism and excretion that we have considered qualitatively in Chapters 8 and 9.

In practice, pharmacokinetics usually focuses on concentrations of drug in *blood plasma*, which is easily sampled via venepuncture, since plasma concentrations are assumed usually to bear a clear relation to the concentration of drug in extracellular fluid surrounding cells that express the receptors or other targets with which drug molecules combine. This underpins what is termed the *target concentration strategy*. Individual variation (Ch. 56) in response to a given dose of a drug is often greater than variability in the plasma concentration at that dose. Plasma concentrations (C_p) are therefore useful in the early stages of drug development (see below), and in the case of a few drugs plasma drug concentrations are also used in routine clinical practice to individualise dosage so as to achieve the desired therapeutic effect while minimising adverse effects in each individual patient, an approach known as *therapeutic drug monitoring* (often abbreviated TDM—see Table 10.1 for examples of some drugs where a therapeutic range of plasma concentrations has been established). Concentrations of drug in other body fluids (e.g. urine,[1] saliva, cerebrospinal fluid, milk) may add useful information in some special situations.

Formal interpretation of pharmacokinetic data consists of fitting concentration versus time data to a theoretical model and determining parameters that describe the observed behaviour. The parameters can then be used to adjust the dose regimen to achieve a desired target plasma concentration estimated initially from pharmacological experiments on cells, tissues or laboratory animals, and modified in light of the human pharmacology if necessary. Some descriptive pharmacokinetic characteristics can be observed directly by inspecting the time course of drug concentration in plasma following dosing—important examples[2] are the *maximum plasma concentration* following a given dose of a drug administered in a defined dosing form (C_{max}) and the *time* (T_{max}) between drug administration and achieving C_{max}. Other pharmacokinetic parameters are estimated mathematically from experimental data; examples include *volume of distribution* (V_d) and *clearance* (CL), concepts that have been introduced in Chapters 8 and 9 respectively and to which we return below.

USES OF PHARMACOKINETICS

Knowledge of pharmacokinetics is crucial in drug development, both to make sense of preclinical toxicity testing and of whole animal pharmacology,[3] and to decide on an appropriate dosing regimen for clinical studies of efficacy (see Ch. 60). Drug regulators need detailed pharmacokinetic information for the same reasons, and must understand principles of *bioavailability* and *bioequivalence* (Ch. 8) to make decisions about licensing generic versions of drugs as these lose their patent protection. An understanding of the general principles of pharmacokinetics is important for clinicians, who need to understand how dosage recommendations in the product information provided with licensed drugs have been arrived at if they are to use the drug optimally. Clinicians also need to understand the principles of pharmacokinetics if they are to identify and evaluate possible drug interactions (see Ch. 56). They also need to be able to interpret drug concentrations for TDM and to adjust dose regimens rationally. In particular, clinicians dealing with a severely ill patient often need to

[1] *Clinical* pharmacology became at one time so associated with the measurement of drugs in urine that the canard had it that clinical pharmacologists were the new alchemists—they turned urine into airline tickets …

[2] Important because dose-related adverse effects often occur around C_{max}.

[3] For example, doses used in experimental animals often need to be much greater than those in humans (on a 'per unit body weight' basis), because drug metabolism is commonly much more rapid in rodents.

Table 10.1 Examples of drugs where therapeutic drug monitoring (TDM) of plasma concentrations is used clinically

Category	Example(s)	See Chapter
Immunosuppressants	Ciclosporine, tacrolimus	26
Cardiovascular	Digoxin	21
Respiratory	Theophylline	16, 27
CNS	Lithium, several antiepileptic drugs	46, 44
Antibacterials	Aminoglycosides	50
Antineoplastics	Methotrexate	55

individualise the dose regimen depending on the urgency of achieving a therapeutic plasma concentration, and whether the clearance of the drug is impaired because of renal or liver disease.

SCOPE OF THIS CHAPTER

The objectives of this chapter are to familiarise the reader with the meanings of important pharmacokinetic parameters; to explain how the total clearance of a drug determines its steady-state plasma concentration during continuous administration; to present a simple model in which the body is represented as a single well-stirred compartment, of volume V_d, that describes the situation before steady state is reached in terms of elimination half-life ($t_{1/2}$); to consider some situations where the simple model is inadequate, and either a two-compartment model or a model where clearance varies with drug concentration ('non-linear kinetics') is needed; to mention briefly the field of population kinetics; and finally to consider some of the limitations inherent in the pharmacokinetic approach. More detailed accounts are provided by Atkinson et al. (2002), Birkett (2002), Jambhekar & Breen (2009) and Rowland & Tozer (2010).

DRUG ELIMINATION EXPRESSED AS CLEARANCE

The concept of *clearance* was introduced in 1929 as a means of expressing the rate of urea excretion in adult humans, in terms of the volume of blood cleared of urea in 1 minute. Clearance of a drug can be defined analogously as the volume of plasma from which all the drug molecules would need to be removed per unit time to achieve the overall rate of elimination of drug from the body. Subsequently, as mentioned in Chapter 9, creatinine rather than urea clearance has become the routine clinical measure of renal functional status because it more closely reflects the glomerular filtration rate. Van Slyke introduced the equation given in Chapter 9 for estimating renal clearance (CL_{ren}). This follows from the law of conservation of mass, and is written:

$$CL_{ren} = \frac{C_u V_u}{C_p} \quad (10.1)$$

where C_u is the urine concentration of the substance of interest (whether endogenous such as urea or creatinine, or exogenous as in the case of an administered drug), C_p its concentration in plasma and V_u the urine flow rate in units of volume/time. C_u and C_p are expressed in the same units of mass/unit volume (e.g. mg/l) so their units cancel out and CL_{ren} has the same units as V, namely volume/unit time—e.g. ml/min or l/h.

The overall clearance of a drug (CL_{tot}) is the fundamental pharmacokinetic parameter describing drug elimination. It is defined as the volume of plasma containing the total amount of drug that is removed from the body in unit time by all routes. Overall clearance is the sum of clearance rates for each mechanism involved in eliminating the drug, usually renal clearance (CL_{ren}) and metabolic clearance (CL_{met}) plus any additional appreciable routes of elimination (faeces, breath, etc.). It relates the rate of elimination of a drug (in units of mass/unit time) to C_p:

$$\text{Rate of drug elimination} = C_p \times CL_{tot} \quad (10.2)$$

Drug clearance can be determined in an individual subject by measuring the plasma concentration of the drug (in units of, say, mg/l) at intervals during a constant-rate intravenous infusion (delivering, say, X mg of drug per h), until a steady state is approximated (Fig. 10.1A). At steady state, the rate of input to the body is equal to the rate of elimination, so:

$$X = C_{SS} \times CL_{tot} \quad (10.3)$$

Rearranging this,

$$CL_{tot} = \frac{X}{C_{SS}} \quad (10.4)$$

where C_{SS} is the plasma concentration at steady state, and CL_{tot} is in units of volume/time (l/h in the example given).

For many drugs, clearance in an individual subject is the same at different doses (at least within the range of doses used therapeutically—but see the section on saturation kinetics below for exceptions), so knowing the clearance enables one to calculate the dose rate needed to achieve a desired steady-state ('target') plasma concentration from equation 10.3.

CL_{tot} can also be estimated by measuring plasma concentrations at intervals following a single intravenous bolus dose of, say, Q mg (Fig. 10.1B):

$$CL_{tot} = \frac{Q}{AUC_{0-\infty}} \quad (10.5)$$

where $AUC_{0-\infty}$ is the area under the full curve[4] relating C_p to time following a bolus dose given at time $t = 0$. (See Ch. 8, and Birkett, 2002, for a fuller account of $AUC_{0-\infty}$.)

Note that these estimates of CL_{tot}, unlike estimates based on the rate constant or half-life (see below), do not depend on any particular compartmental model.

SINGLE-COMPARTMENT MODEL

Consider a highly simplified model of a human being, which consists of a single well-stirred compartment, of

[4]The area is obtained by integrating from time = 0 to time = ∞, and is designated $AUC_{0-\infty}$. The area under the curve has units of time—on the abscissa—multiplied by concentration (mass/volume)—on the ordinate; so $CL = Q/AUC_{0-\infty}$ has units of volume/time as it should.

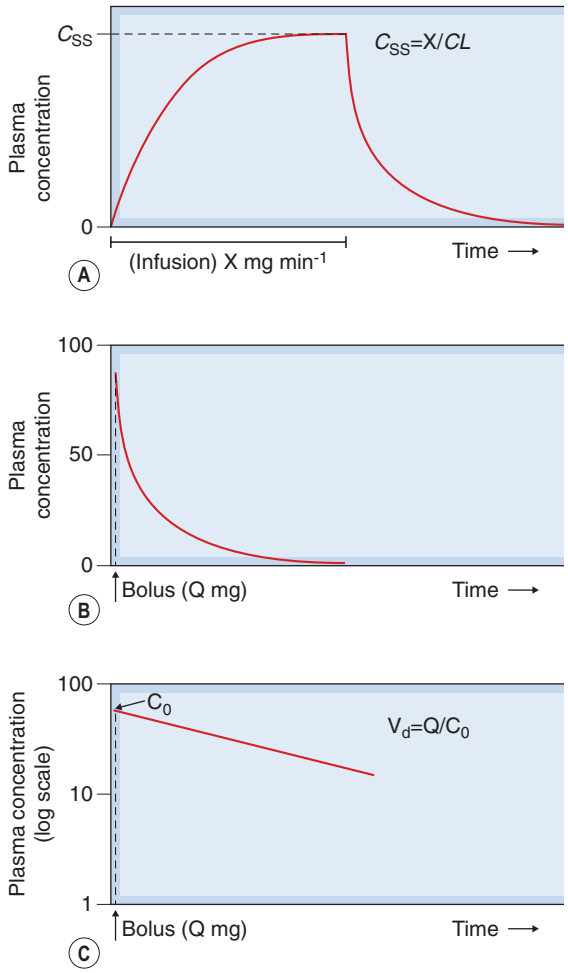

Fig. 10.1 Plasma drug concentration-time curves. [**A**] During a constant intravenous infusion at rate X mg/min, indicated by the horizontal bar, the plasma concentration (C) increases from zero to a steady-state value (C_{SS}); when the infusion is stopped, C declines to zero. [**B**] Following an intravenous bolus dose (Q mg), the plasma concentration rises abruptly and then declines towards zero. [**C**] Data from panel B plotted with plasma concentrations on a logarithmic scale. The straight line shows that concentration declines exponentially. Extrapolation back to the ordinate at zero time gives an estimate of C_0, the concentration at zero time, and hence of V_d, the volume of distribution.

volume V_d (distribution volume), into which a quantity of drug Q is introduced rapidly by intravenous injection, and from which it can escape either by being metabolised or by being excreted (Fig. 10.2). For most drugs, V_d is an apparent volume rather than the volume of an anatomical compartment. It links the total amount of drug in the body to its concentration in plasma (see Ch. 8). The quantity of drug in the body when it is administered as a single bolus is equal to the administered dose Q. The initial concentration, C_0, will therefore be given by:

$$C_0 = \frac{Q}{V_d} \qquad (10.6)$$

In practice, C_0 is estimated by extrapolating the linear portion of a semilogarithmic plot of C_p against time back

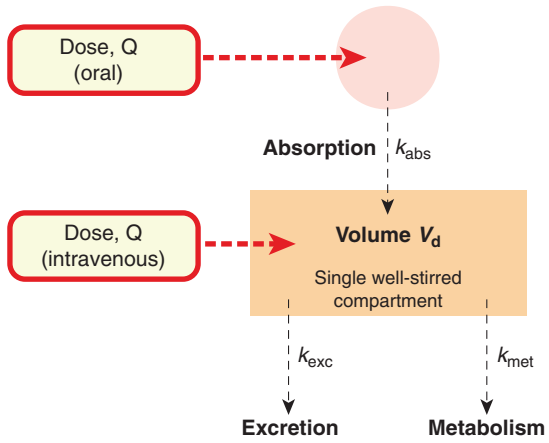

Fig. 10.2 Single-compartment pharmacokinetic model. This model is applicable if the plasma concentration falls exponentially after drug administration (as in Fig. 10.1).

to its intercept at time 0 (Fig. 10.1C). C_p at any time depends on the rate of elimination of the drug (i.e. on its total clearance, CL_{tot}) as well as on the dose and V_d. Many drugs exhibit *first-order kinetics* where the rate of elimination is directly proportional to drug concentration. Drug concentration then decays exponentially (Fig. 10.3), being described by the equation:

$$C_{(t)} = C_{(0)} \exp \frac{-CL_{tot}}{V_d} t \qquad (10.7)$$

Taking logarithms:

$$\ln C_{(t)} = \ln C_{(0)} - \frac{-CL_{tot}}{V_d} t \qquad (10.8)$$

Plotting C_t on a logarithmic scale against t (on a linear scale) yields a straight line with slope $-CL_{tot}/V_d$. The inverse of this slope (CL_{tot}/V_d) is the *elimination rate constant* k_{el}, which has units of (time)$^{-1}$. It represents the *fraction* of drug in the body eliminated per unit of time. For example, if the rate constant is 0.1 h^{-1} this implies that one-tenth of the drug remaining in the body is eliminated each hour.

The *elimination half-life*, $t_{1/2}$, is an easily conceptualised parameter inversely related to k_{el}. It is the time taken for C_p to decrease by 50%, and is equal to $\ln 2/k_{el}$ (= 0.693/k_{el}). The plasma half-life is therefore determined by V_d as well as by CL_{tot}. It enables one to predict what will happen after drug administration is initiated before steady state is reached, and after drug administration has been stopped while C_p declines toward zero.

When the single-compartment model is applicable, the drug concentration in plasma approaches the steady-state value approximately exponentially during a constant infusion (Fig. 10.1A). When the infusion is discontinued, the concentration falls exponentially towards zero: after one half-life, the concentration will have fallen to half the initial concentration; after two half-lives, it will have fallen to one-quarter the initial concentration; after three half-lives, to one-eighth; and so on. It is intuitively obvious that the longer the half-life, the longer the drug will persist in the body after dosing is discontinued. It is less obvious, but nonetheless true, that during chronic drug administration the longer the half-life, the longer it will take for the drug

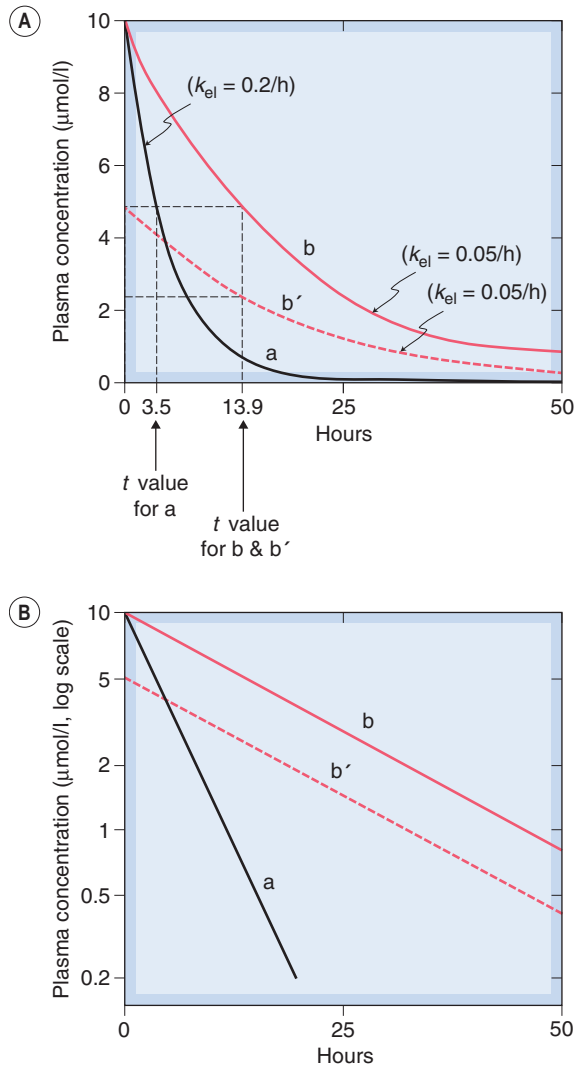

Fig. 10.3 Predicted behaviour of single-compartment model following intravenous drug administration at time 0. Drugs a and b differ only in their elimination rate constant, k_{el}. Curve b′ shows the plasma concentration time course for a smaller dose of b. Note that the half-life ($t_{1/2}$) (indicated by broken lines) does not depend on the dose. [**A**] Linear concentration scale. [**B**] Logarithmic concentration scale.

to accumulate to its steady-state level: one half-life to reach 50% of the steady-state value, two to reach 75%, three to reach 87.5% and so on. This is extremely helpful to a clinician deciding how to start treatment. If the drug in question has a half-life of approximately 24 h, for example, it will take 3–5 days to approximate the steady-state concentration during a constant-rate infusion. If this is too slow in the face of the prevailing clinical situation, a *loading dose* may be used in order to achieve a therapeutic concentration of drug in the plasma more rapidly (see below). The size of such a dose is determined by the volume of distribution (equation 10.6).

EFFECT OF REPEATED DOSING

Drugs are usually given as repeated doses rather than single injections or a constant infusion. Repeated injections (each of dose Q) give a more complicated pattern than the smooth exponential rise during intravenous infusion, but the principle is the same (Fig. 10.4). The concentration will rise to a mean steady-state concentration with an approximately exponential time course, but will oscillate (through a range Q/V_d). The smaller and more frequent the doses, the more closely the situation approaches that of a continuous infusion, and the smaller the swings in concentration. The exact dosage schedule, however, does not affect the mean steady-state concentration, or the rate at which it is approached. In practice, a steady state is effectively achieved after three to five half-lives. Speedier attainment of the steady state can be achieved by starting with a larger dose, as mentioned above. Such a loading dose is sometimes used when starting treatment with a drug with a half-life that is long in the context of the urgency of the clinical situation, as may be the case when treating cardiac dysrhythmias with drugs such as **amiodarone** or **digoxin** (Ch. 21) or initiating anticoagulation with **heparin** (Ch. 24).

EFFECT OF VARIATION IN RATE OF ABSORPTION

If a drug is absorbed slowly from the gut or from an injection site into the plasma, it is (in terms of a compartmental model) as though it were being slowly infused at a variable rate into the bloodstream. For the purpose of kinetic modelling, the transfer of drug from the site of administration to the central compartment can be represented approximately by a rate constant, k_{abs} (see Fig. 10.2). This assumes that the rate of absorption is directly proportional, at any moment, to the amount of drug still unabsorbed, which is at best a rough approximation to reality. The effect of slow absorption on the time course of the rise and fall of the plasma concentration is shown in Figure 10.5. The curves show the effect of spreading out the absorption of the same total amount of drug over different times. In each case, the drug is absorbed completely, but the peak concentration appears later and is lower and less sharp if absorption is slow. In the limiting case, a dosage form that releases drug at a constant rate as it traverses the ileum (Ch. 8) approximates a constant-rate infusion. Once absorption is complete, the plasma concentration declines with the same half-time, irrespective of the rate of absorption.

▼ For the kind of pharmacokinetic model discussed here, the area under the plasma concentration–time curve (AUC) is directly proportional to the total amount of drug introduced into the plasma compartment, irrespective of the rate at which it enters. Incomplete absorption, or destruction by first-pass metabolism before the drug reaches the plasma compartment, reduces AUC after oral administration (see Ch. 8). Changes in the rate of absorption, however, do not affect AUC. Again, it is worth noting that provided absorption is complete, the relation between the rate of administration and the steady-state plasma concentration (equation 10.4) is unaffected by k_{abs}, although the size of the oscillation of plasma concentration with each dose is reduced if absorption is slowed.

MORE COMPLICATED KINETIC MODELS

So far, we have considered a single-compartment pharmacokinetic model in which the rates of absorption, metabolism and excretion are all assumed to be directly proportional to the concentration of drug in the compartment from which transfer is occurring. This is a useful way to illustrate some basic principles but is clearly a physio-

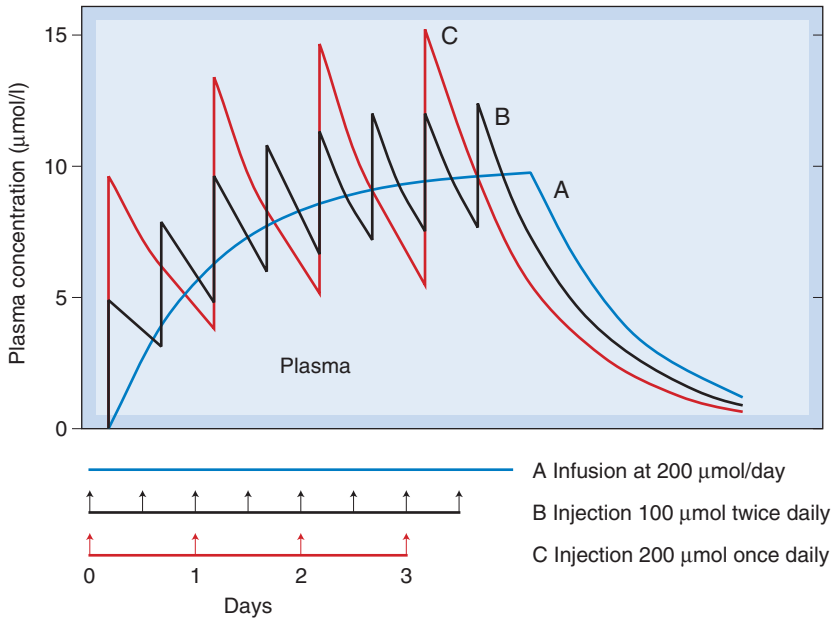

Fig. 10.4 Predicted behaviour of single-compartment model with continuous or intermittent drug administration. Smooth curve A shows the effect of continuous infusion for 4 days; curve B the same total amount of drug given in eight equal doses; and curve C the same total amount of drug given in four equal doses. The drug has a half-life of 17 h and a volume of distribution of 20 l. Note that in each case a steady state is effectively reached after about 2 days (about three half-lives), and that the mean concentration reached in the steady state is the same for all three schedules.

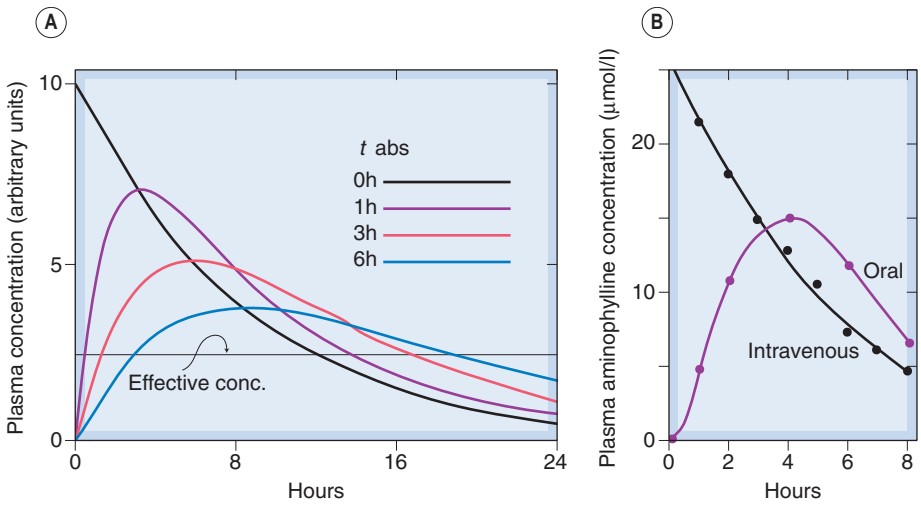

Fig. 10.5 The effect of slow drug absorption on plasma drug concentration. [**A**] Predicted behaviour of single-compartment model with drug absorbed at different rates from the gut or an injection site. The elimination half-time is 6 h. The absorption half-times ($t_{1/2}$ abs) are marked on the diagram. (Zero indicates instantaneous absorption, corresponding to intravenous administration.) Note that the peak plasma concentration is reduced and delayed by slow absorption, and the duration of action is somewhat increased. [**B**] Measurements of plasma aminophylline concentration in humans following equal oral and intravenous doses. (Data from Swintowsky J V 1956 J Am Pharm Assoc 49: 395.)

logical oversimplification. The characteristics of different parts of the body, such as brain, body fat and muscle, are quite different in terms of their blood supply, partition coefficient for drugs and the permeability of their capillaries to drugs. These differences, which the single-compartment model ignores, can markedly affect the time courses of drug distribution and action, and much theoretical work has gone into the mathematical analysis of more complex models (see Atkinson et al., 2002; Rowland & Tozer, 2010). They are beyond the scope of this book, and perhaps also beyond the limit of what is actually useful, for the experimental data on pharmacokinetic properties of drugs are seldom accurate or reproducible enough to enable complex models to be tested critically.

The two-compartment model, which introduces a separate 'peripheral' compartment to represent the tissues, in communication with the 'central' plasma compartment, more closely resembles the real situation without involving excessive complications.

TWO-COMPARTMENT MODEL

The two-compartment model is a widely used approximation in which the tissues are lumped together as a peripheral compartment. Drug molecules can enter and leave the peripheral compartment only via the central compartment (Fig. 10.6), which usually represents the plasma (or plasma plus some extravascular space in the case of a few drugs that distribute especially rapidly). The effect of adding a second compartment to the model is to introduce a second exponential component into the predicted time course of the plasma concentration, so that it comprises a fast and a

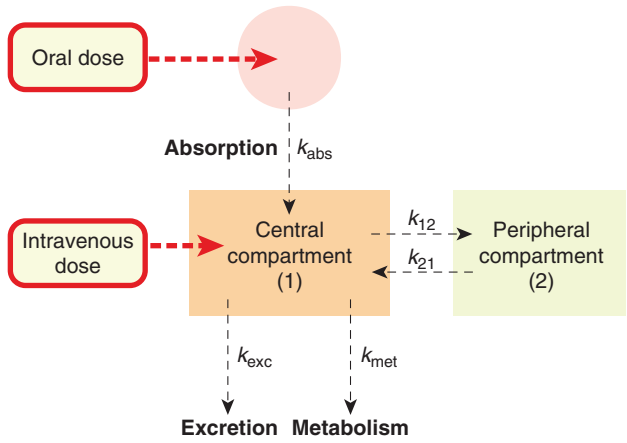

Fig. 10.6 Two-compartment pharmacokinetic model.

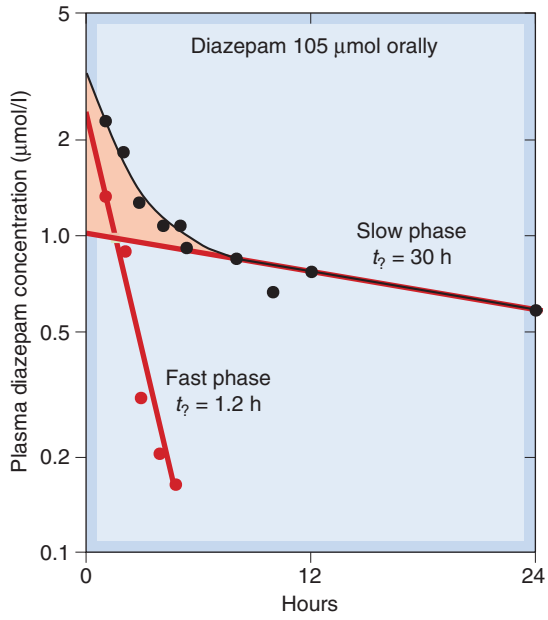

Fig. 10.7 Kinetics of diazepam elimination in humans following a single oral dose. The graph shows a semilogarithmic plot of plasma concentration versus time. The experimental data (black symbols) follow a curve that becomes linear after about 8 h (slow phase). Plotting the deviation of the early points (pink shaded area) from this line on the same coordinates (red symbols) reveals the fast phase. This type of two-component decay is consistent with the two-compartment model (Fig. 10.6) and is obtained with many drugs. (Data from Curry S H 1980 Drug disposition and pharmacokinetics. Blackwell, Oxford.)

slow phase. This pattern is often found experimentally, and is most clearly revealed when the concentration data are plotted semilogarithmically (Fig. 10.7). If, as is often the case, the transfer of drug between the central and peripheral compartments is relatively fast compared with the rate of elimination, then the fast phase (often called the *α phase*) can be taken to represent the redistribution of the drug (i.e. drug molecules passing from plasma to tissues, thereby

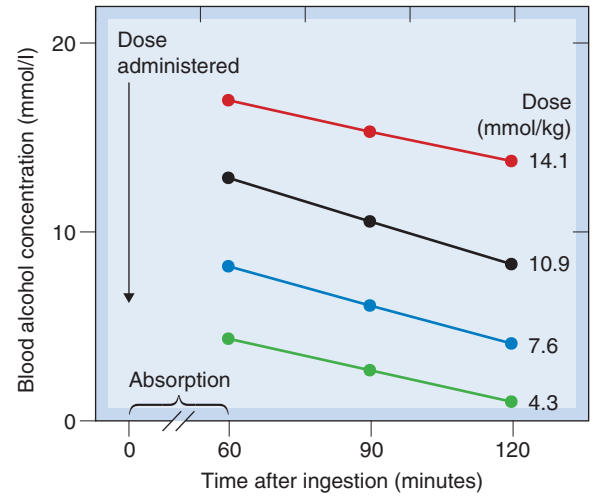

Fig. 10.8 Saturating kinetics of alcohol elimination in humans. The blood alcohol concentration falls linearly rather than exponentially, and the rate of fall does not vary with dose. (From Drew G C et al. 1958 Br Med J 2: 5103.)

rapidly lowering the plasma concentration). The plasma concentration reached when the fast phase is complete, but before appreciable elimination has occurred, allows a measure of the combined distribution volumes of the two compartments; the half-time for the slow phase (the *β phase*) provides an estimate of k_{el}. If a drug is rapidly metabolised, the α and β phases are not well separated, and the calculation of V_d and k_{el} is not straightforward. Problems also arise with drugs (e.g. very fat-soluble drugs) for which it is unrealistic to lump all the peripheral tissues together.

SATURATION KINETICS

In a few cases, such as **ethanol**, **phenytoin** and **salicylate**, the time course of disappearance of drug from the plasma does not follow the exponential or biexponential patterns shown in Figures 10.3 and 10.7 but is initially linear (i.e. drug is removed at a constant rate that is independent of plasma concentration). This is often called *zero-order kinetics* to distinguish it from the usual first-order kinetics that we have considered so far (these terms have their origin in chemical kinetic theory). *Saturation kinetics* is a better term. Figure 10.8 shows the example of ethanol. It can be seen that the rate of disappearance of ethanol from the plasma is constant at approximately 4 mmol/l per h, irrespective of dose or of the plasma concentration of ethanol. The explanation for this is that the rate of oxidation by the enzyme alcohol dehydrogenase reaches a maximum at low ethanol concentrations, because of limited availability of the cofactor NAD^+ (see Ch. 48, Fig. 48.5).

Saturation kinetics has several important consequences (see Fig. 10.9). One is that the duration of action is more strongly dependent on dose than is the case with drugs that do not show metabolic saturation. Another consequence is that the relationship between dose and steady-state plasma concentration is steep and unpredictable, and it does not obey the proportionality rule implicit in equation 10.4 for non-saturating drugs (see Fig. 48.6 for another example related to ethanol). The maximum rate of metabolism sets a limit to the rate at which the drug can

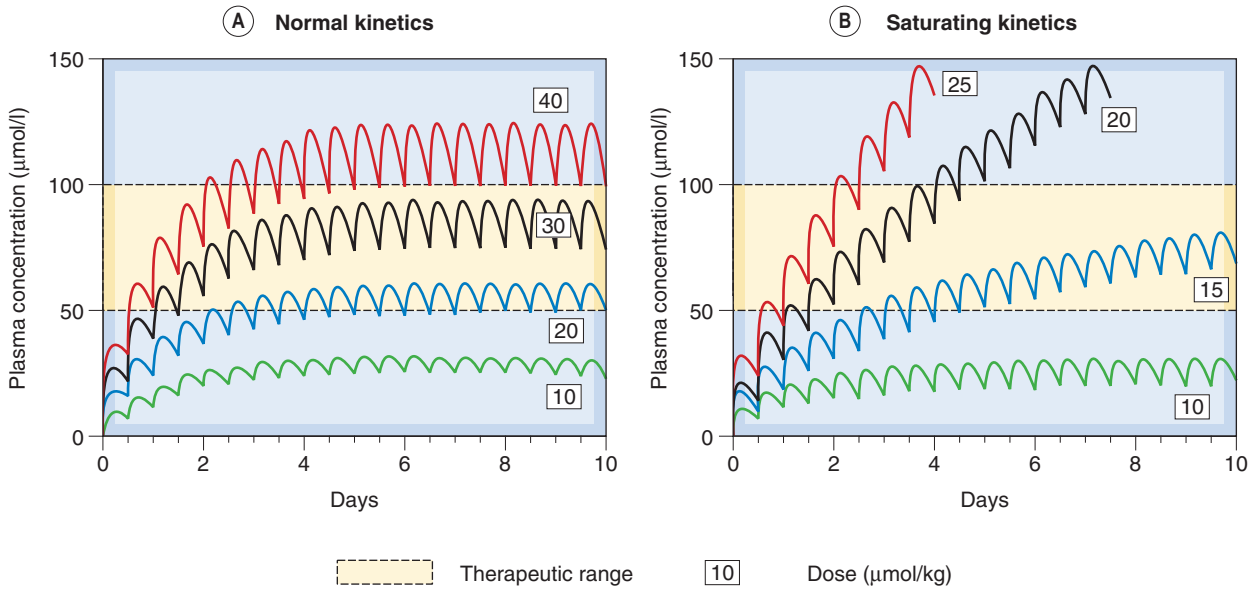

Fig. 10.9 Comparison of non-saturating and saturating kinetics for drugs given orally every 12 h. [**A**] The curves showing an imaginary drug, similar to the antiepileptic drug phenytoin at the lowest dose, but with linear kinetics. The steady-state plasma concentration is reached within a few days, and is directly proportional to dose. [**B**] Curves for saturating kinetics calculated from the known pharmacokinetic parameters of phenytoin (see Ch. 44). Note that no steady state is reached with higher doses of phenytoin, and that a small increment in dose results after a time in a disproportionately large effect on plasma concentration. (Curves were calculated with the Sympak pharmacokinetic modelling program written by Dr J G Blackman, University of Otago.)

be administered; if this rate is exceeded, the amount of drug in the body will, in principle, increase indefinitely and never reach a steady state (Fig. 10.9). This does not actually happen, because there is always some dependence of the rate of elimination on the plasma concentration (usually because other, non-saturating metabolic pathways or renal excretion contribute significantly at high concentrations). Nevertheless, steady-state plasma concentrations of drugs of this kind vary widely and unpredictably with dose. Similarly, variations in the rate of metabolism (e.g. through enzyme induction) cause disproportionately large changes in the plasma concentration. These problems are well recognised for drugs such as phenytoin, an anticonvulsant for which plasma concentration needs to be closely controlled to achieve an optimal clinical effect (see Ch. 44, Fig. 44.4). Drugs showing saturation kinetics are less predictable in clinical use than ones with linear kinetics, so may be rejected during drug development if a pharmacologically similar candidate with linear kinetics is available (Ch. 60).

Clinical applications of pharmacokinetics are summarised in the clinical box.

POPULATION PHARMACOKINETICS

▼ In some situations, for example when the drug is intended for use in chronically ill children, it is desirable to obtain pharmacokinetic data in a patient population rather than in healthy adult volunteers. Such studies are inevitably constrained and samples for drug analysis are often obtained opportunistically during clinical care, with limitations as to quality of the data and only sparse data collected from each patient. Population pharmacokinetics addresses how best to analyse such data. Various approaches that have been used, including fitting data from all subjects as if there were no kinetic differences between individuals, and fitting each individual's data separately and then combining the individual parameter estimates, have obvious shortcomings. A better method is to use non-linear mixed effects modelling (NONMEM). The statistical technicalities are considerable and beyond the scope of this chapter: the interested reader is referred to Sheiner et al. (1997); and, for NONMEM software user guides, to Beale & Sheiner (1989).

LIMITATIONS OF PHARMACOKINETICS

Some limitations of the pharmacokinetic approach will be obvious from the above account, such as the proliferation of parameters in even quite conceptually simple models. Here we comment on two assumptions that underpin the idea that by relating response to a drug to its plasma concentration we reduce variability by accounting for pharmacokinetic variation—that is, variation in absorption, distribution, metabolism and excretion:

1. That plasma concentration of a drug bears a precise relation to the concentration of drug in the immediate environment of its target (receptor, enzyme, etc.).
2. That drug response depends only on the concentration of the drug in the immediate environment of its target.

While the first of these assumptions is very plausible in the case of a drug working on a target in the circulating blood (e.g. a fibrinolytic drug working on fibrinogen) and reasonably plausible for a drug working on an enzyme, ion channel or G-protein-coupled or kinase-linked receptor located in the cell membrane, it is less likely in the case of a nuclear receptor or when the target cells are protected by the blood–brain barrier. In the latter case it is not perhaps surprising that, despite considerable efforts, it has

Uses of pharmacokinetics

- Pharmacokinetic studies performed during drug development underpin the standard dose regimens approved by regulatory agencies.
- Clinicians sometimes need to individualise dose regimens to account for individual variation in a particular patient (e.g. a neonate, a patient with impaired and changing renal function, or a patient taking drugs that interfere with drug metabolism; see Ch. 56).
- Drug effect (pharmacodynamics) is often used for such individualisation, but there are drugs (including some anticonvulsants, immunosuppressants and antineoplastics) where a therapeutic range of plasma concentrations has been defined, and for which it is useful to adjust the dose to achieve a concentration in this range.
- Knowledge of kinetics enables rational dose adjustment. For example:
  - the dose interval of a drug such as **gentamicin** eliminated by renal excretion may need to be markedly increased in a patient with renal impairment (Ch. 50)
  - the dose increment needed to achieve a target plasma concentration range of a drug such as **phenytoin** with saturation kinetics (Ch. 44, Fig.44.4) is much less than for a drug with linear kinetics.
- Knowing the approximate $t_{1/2}$ of a drug can be very useful, even if a therapeutic concentration is not known:
  - in correctly interpreting adverse events that occur some considerable time after starting regular treatment (e.g. benzodiazepines; see Ch. 43)
  - in deciding on the need or otherwise for an initial loading dose when starting treatment with drugs such as **digoxin** and **amiodarone** (Ch. 21).
- The volume of distribution (V_d) of a drug determines the size of loading dose needed. If V_d is large (as for many tricyclic antidepressants), haemodialysis will not be an effective way of increasing the rate of elimination in treating overdose.

never proved clinically useful to measure plasma concentrations of antidepressant or antipsychotic drugs, where there are, in addition, complex metabolic pathways with numerous active metabolites. It is, if anything, surprising that the approach does as well as it does in the case of some other centrally acting drugs, notably antiepileptics and **lithium**.

The second assumption is untrue in the case of drugs that form a stable covalent attachment with their target, and so produce an effect that outlives their presence in solution. Examples include the antiplatelet effects of **aspirin** and **clopidogrel** (Ch. 24) and the effect of some monoamine oxidase inhibitors (Ch. 46). In other cases, drugs in therapeutic use act only after delay (e.g. antidepressants, Ch. 46), or gradually induce tolerance (e.g. opioids, Ch. 41) or physiological adaptations (e.g. corticosteroids, Ch. 32) which alter the relation between concentration and drug effect in a time-dependent manner.

Pharmacokinetics

- Total clearance (CL_{tot}) of a drug is the fundamental parameter describing its elimination: the rate of elimination equals CL_{tot} multiplied by plasma concentration.
- CL_{tot} determines steady-state plasma concentration (C_{SS}): C_{SS} = rate of drug administration/CL_{tot}.
- For many drugs, disappearance from the plasma follows an approximately exponential time course. Such drugs can be described by a model where the body is treated as a single well-stirred compartment of volume V_d. V_d is an apparent volume linking the amount of drug in the body at any time to the plasma concentration.
- Elimination half-life ($t_{1/2}$) is directly proportional to V_d and inversely proportional to CL_{tot}.
- With repeated dosage or sustained delivery of a drug, the plasma concentration approaches a steady value within three to five plasma half-lives.
- In urgent situations, a loading dose may be needed to achieve therapeutic concentration rapidly.
- The loading dose (L) needed to achieve a desired initial plasma concentration C_{target} is determined by V_d: $L = C_{target} \times V_d$.
- A two-compartment model is often needed. In this case, the kinetics are biexponential. The two components roughly represent the processes of transfer between plasma and tissues (α phase) and elimination from the plasma (β phase).
- Some drugs show non-exponential 'saturation' kinetics, with important clinical consequences, especially a disproportionate increase in steady-state plasma concentration when daily dose is increased.

REFERENCES AND FURTHER READING

Atkinson, A.J., Daniels, C.E., Dedrick, R.L., et al. (Eds.), 2002. Principles of clinical pharmacology. Academic Press, London. (*Section on pharmacokinetics includes the application of Laplace transformations, effects of disease, compartmental versus non-compartmental approaches, population pharmacokinetics, drug metabolism and transport*)

Birkett, D.J., 2002. Pharmacokinetics made easy (revised), 2nd edn. McGraw-Hill Australia, Sydney. (*Excellent slim volume that lives up to the promise of its title*)

Jambhekar, S.S., Breen, P.J., 2009. Basic pharmacokinetics. Pharmaceutical Press, London. (*Basic textbook*)

Population pharmacokinetics

Beale, S.L., Sheiner, L.B., 1989. NONMEM user's guides. NONMEM Project Group, University of California, San Francisco.

Rowland, M., Tozer, T.N., 2010. Clinical pharmacokinetics and pharmacodynamics. Concepts and applications. Wolters Kluwer/Lippincott Williams & Wilkins, Baltimore. Online simulations by H Derendorf and G Hochhaus. (*Excellent text; emphasises clinical applications*)

Sheiner, L.B., Rosenberg, B., Marethe, V.V., 1997. Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. J. Pharmacokinet. Biopharm. 5, 445–479.