8 Early Clinical Trials

Summary

Once the activity of a compound has been established in the laboratory (usually by use of experimental animals) the next stage of development is to bring this forward to man in what are often termed early phase clinical trials. A pharmacokinetic study aims to establish an effective dosing regimen for the compound in order to reach concentrations within the therapeutic window as quickly as possible, and to stay within the desired range by suitable choice of maintenance dose and dosing interval. The usual aim of Phase I trials is to determine a maximal safe dose with which more rigorous investigation of activity in a Phase II trial can be conducted. The Phase II trial, if demonstrating activity that suggests efficacy, may be the precursor to a randomised Phase III trial comparing the agent with a standard treatment for the disease or condition in question. This chapter deals with issues related to the design of such early human studies.

8.1 INTRODUCTION

As we have indicated earlier, the design of any study is the key component for obtaining a satisfactory answer to the question posed. We shall see later in Chapter 9 that careful consideration is given to the design (and sample size) aspects when considering randomised controlled trials. However, an important factor when considering the design of these Phase III trials is the information provided from earlier stage studies and trials. Consequently since the development of, for example, new therapies or medical devices, tends to progress at the clinical stage through pharmacokinetic (PK), Phase I to Phase II then to Phase III trials as in Figure 7.1 the sequential nature of this structure implies that reliable information from one step is important for the next. Poor ‘experimentation’ at the relevant stage can clearly jeopardise the design of the next stage and, at best, results in a waste of resource and, at worst, may compromise patient safety. Unfortunately, the evidence provided by published reports of early stage trials suggests that these have often not been well designed or well reported.

PK studies attempt to characterise the fate of a drug in the body while a Phase I trial aims to determine (often from a pre-selected range of potential doses) the dose that can be utilised at the next stage of development and so focuses on selecting the highest practicable dose. The presumption is that the greater the dose the greater the
anti-disease effect will be. However, safety considerations dictate that the dose chosen for the subsequent trials should have an acceptable toxicity profile. Early indications of activity against the disease may be noted at the Phase I stage, but this is often incidental to the main purpose of the trial.

The objective of a Phase II trial is to assess the activity of a new drug with a view to deciding whether or not the regimen has sufficient potential efficacy to warrant further study. Thus for patients with solid tumour cancers, patients are recruited to a Phase II trial and the proportion in whom there is complete or partial tumour shrinkage (these have to be carefully defined) is determined. If this proportion is sufficiently high a randomised Phase III trial comparing this (new) treatment with the current standard for the disease in question may be recommended as the next step in the development sequence.

The two primary outcomes of toxicity and sufficient activity can be considered simultaneously in certain Phase II designs. Although in most instances Phase II trials are single-arm studies, there may be several compounds available to study as potentials for a future Phase III trial in which case a randomised comparative design may be appropriate.

Although in this chapter we use the development of a new compound as the principal example, the methods do have wider applicability, for example, testing two established compounds but combined together in a single regimen.

### 8.2 PHARMACOKINETIC STUDIES

**AIMS**

The aims of a PK study are to establish tolerability of a new compound and determine the dose to use, the appropriate route and the associated schedule. They are often extended to investigating potential interactions with other drugs likely to be used in the target population and identification of patient characteristics, such as gender, weight, ethnic group or renal function, that exert an effect on the kinetics of the drug substantial enough to warrant dose adjustment.

**THERAPEUTIC WINDOW**

Underlying most dosage regimens is the idea of a ‘therapeutic window’, which is a range within which drug concentrations should be maintained to achieve clinical benefit. Concentrations that are too low may not achieve efficacy, whereas higher levels may result in undesirable side effects. For instance, most antibiotics require a certain minimum inhibitory concentration to be sustained to maintain efficacy against a particular target organism. An effective dosing regimen should aim to reach concentrations within the therapeutic window as quickly as possible, and to stay within the desired range by suitable choices of maintenance dose and dosing interval.

For some classes of agents, it may be unlikely that there is a common therapeutic window for all patients. Indeed subject-specific factors may alter the relationship between drug concentration and effect to such a degree that the desirable concentration range may differ substantially across patients.
Example – pharmacokinetic study – telithromycin for respiratory tract infection

Bhargava, Lenfant, Perret et al. (2002) state that in preliminary studies in humans telithromycin has been found to be well tolerated and to possess a pharmacokinetic profile supporting a once-daily 800 mg oral dose taken in the mornings with 240 ml water. They indicate that this dose has been selected for use in Phase III clinical trials against respiratory tract infections.

COMPARTMENTAL MODELS

Underlying all PK studies is the concept of a compartmental model. Essentially these represent the body as a system of compartments that communicate reversibly with each other. A compartment is not so much a particular anatomical or physiological region, but rather a tissue or tissues with similar blood flow and drug affinity. For example, the liver and kidneys, being highly perfused organs, are often considered as being in the same compartment as the circulation. The compartments are assumed well mixed with a uniform distribution of the administered drug throughout. Figure 8.1 shows a two-compartment model to describe drug kinetics following a single intravenous (IV) bolus injection of a drug in which, following essentially instantaneous absorption of the drug into the circulation, it is assumed to distribute into the two compartments. The central compartment ($V_1$) represents the blood, extracellular fluid and highly perfused organs and tissues. The second (peripheral) compartment ($V_2$) may be thought of as other, poorly perfused, tissues. For the model of Figure 8.1, elimination of the drug is assumed to occur from the central (plasma) compartment only. The rate constants, $k_{1,0}$, $k_{1,2}$ and $k_{2,1}$, govern the kinetic transfers into and out of the relevant compartments.

![Figure 8.1](image)

**Figure 8.1** Two-compartment model to describe drug kinetics following a single intravenous bolus injection (from Mant and Allen, 2001. Early phase studies, pharmacokinetics and adverse drug interactions. In I Di Giovanna and G Hayes, *Principles of Clinical Research*. Wrightson Biomedical Publishing, Petersfield, pp 117–160. [8])
Sampling Design

A full population PK sampling design requires that blood samples should be drawn from subjects at various times (typically one to six time-points) following drug administration. The objective is to obtain multiple drug levels per patient at different times to describe the individual PK profiles. These are then averaged over all subjects studied. The number of repeated measures taken, and their location in time, will depend on the shape of the PK profile and this will depend on the type of drug under investigation.

The basic features of a PK profile are given in Figure 8.2, which plots the concentration of the drug in the plasma against the time from the dose being given. The area under this concentration/time curve, $AUC$, serves to measure the extent of absorption, whereas, in the case of fast-releasing formulations, the maximum concentration, $C_{\text{Maximum}}$, and the time of occurrence, $t_{\text{Maximum}}$, characterise the rate of absorption.

Example – repeated measures design – telithromycin for respiratory tract infections

Venous blood samples were taken by Bhargava, Lenfant, Perret et al. (2002) for telithromycin plasma level determination immediately before and at 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, 8, 12, 24, 34, 48 and 58 h after medication. In this example, measures extended over a 2.5-day period but were most frequent in Day 1. Clearly the early observation points will focus on estimating $C_{\text{Maximum}}$ and $t_{\text{Maximum}}$, while the later observations are crucial to establish the $AUC$. 
Study Size

Since in these single-group studies each subject provides an estimate of $C_{\text{Maximum}}$, $t_{\text{Maximum}}$, and $AUC$, these can be averaged over all subjects studied for summary purposes. Julious and Debarnot (2000) state categorically that these summaries should be on the logarithmic scale, which is equivalent to quoting their geometric rather than arithmetic mean values. Sample sizes can then be estimated using equation (3.3).

Example – sample size – mean log $C_{\text{Maximum}}$

Wooding (1994) gives an example of mean log $C_{\text{Maximum}}$ for a drug as 3.36 and, on the same scale, the corresponding within-subject as $SD_{\text{Within}}=0.40$. For a 95% CI of width $\alpha=0.25$, equation (3.3) gives, $N = 4[0.40^2/0.25^2] \times 1.96^2 = 39.4 \approx 40$ subjects.

BIOEQUIVALENCE

In certain situations, one may wish to compare the PK profiles of different formulations of the same compound or the same compound used in different circumstances, perhaps in a paediatric as opposed to an adult population. In either case the studies may be seeking equivalence rather than superiority. Bioequivalence of different formulations of the same drug is usually taken to mean equivalence with respect to rate and extent of drug absorption. For many drug substances, a large between-subject variation is known to exist and so crossover designs are recommended for bioequivalence studies. It is usual to employ a balanced two-period design. In such a crossover trial, if a test ($T$) formulation is to be compared with the reference ($R$) formulation, then the subjects will usually be randomised equally between the sequences $TR$ and $RT$.

Example – crossover bioequivalence study – absorption of telithromycin in healthy volunteers

Bhargava, Lenfant, Perret et al. (2002) describe a two-period design in which 18 healthy volunteers took a single dose, 800 mg in two tablets, of telithromycin on one occasion following fasting ($F$) and on another occasion following a meal ($M$). Their summary results at each observation time are given in Figure 8.3. The object of the study was to determine whether or not the rate and extent of absorption of this antibacterial remained unaffected by food intake. Each period comprised 58 hours of post-drug observation requiring venous blood samples at 16 different time points. The wash-out period was between 6 and 8 days.
As we have indicated, measures of drug absorption are plausibly log-Normally distributed and studies should focus on the ratio of the two means, \( \theta = \frac{\mu_{\text{Test}}}{\mu_{\text{Reference}}} \), with corresponding null hypothesis \( H_0: \theta = 1 \), rather than their difference. Thus lower and upper bounds of bioequivalence are expressed as above and below the null hypothesis ratio of unity. Commonly used values for these are \( \theta_L = 0.8 \) and \( \theta_U = 1.25 \). On a logarithmic scale these are equidistant from \( \log 1 = 0 \), since \( \log 0.8 = -0.22 \) and \( \log 1.25 = +0.22 \). Bioequivalence is conceded if the two-sided \( 100(1-\alpha)\% \) CI for the ratio \( \frac{\mu_{\text{Test}}}{\mu_{\text{Reference}}} \) is completely contained within the interval \( (\theta_L, \theta_U) \).

Although many bioequivalence studies continue to be planned and reported with respect to a difference between means rather than by a ratio, Julious and Debarnot (2000) do not advocate this approach and so it is omitted here.

### Study Size

Since bioequivalence studies are usually small, the sample-size equations require that (essentially) the Normal distribution values \( z_{1-\alpha} \) and \( z_{1-\beta/2} \) are replaced by those of the Student’s \( t \)-distribution \( t_{f,1-\alpha} \) and \( t_{f,1-\beta/2} \) respectively, where \( f \) is the number of degrees of freedom.

Thus adapting equation (3.22) for the equivalence of two means for larger sample sizes and specifying equal group sizes, that is \( \lambda = 1 \), the required number of subjects, half to receive the sequence \( TR \) and half \( RT \), is

\[
N_{\text{Bioequivalence}} = \frac{2\sigma^2(t_{f,1-\alpha} + t_{f,1-\beta/2})^2}{\varepsilon^2}.
\]

(8.1)
Here, $\varepsilon$ is the limit for equivalence and it is assumed that $\mu_{\text{Test}} = \mu_{\text{Reference}} = \mu$ hence their ratio is equal to 1.

Equation (8.1) can be expressed in terms of the coefficient of variation, $CV = \sigma/\mu$, and if we also define $\Omega = \varepsilon/\mu$, then

$$N = \frac{2CV^2(t_{f,1-\varepsilon} + t_{f,1-\beta/2})^2}{\Omega^2}. \quad (8.2)$$

Now, for example, $t_{f,1-\varepsilon}$ besides depending on $\varepsilon$ also depends on the number of degrees of freedom, $f$, utilised to estimate $\sigma$ in the final analysis of the design. For a two-period crossover design in which there is no period effect, if the analysis is by means of a paired $t$-test of the $N$ differences observed, then there are $f = N - 1$ degrees of freedom. Thus $t_{f,1-\varepsilon}$ depends on the sample size, $N$, whereas $z_{1-\varepsilon}$ does not.

To obtain the sample size from equation (8.1), or (8.2), an iterative process is required. This starts by assuming infinite degrees of freedom, that is, using $z_{1-\varepsilon}$ and $z_{1-\beta/2}$ in place of $t_{f,1-\varepsilon}$ and $t_{f,1-\beta/2}$ to obtain a starting value for the sample size, denoted $N_I$. From this a provisional value for the degrees of freedom is $f_I = N_I - 1$. This can then be used in Table T9 of the Student’s $t$-distribution to give $t_{f_I,1-\varepsilon}$ and $t_{f_I,1-\beta/2}$, which are then substituted in equation (8.1). This then provides a revised estimate of the sample size, $N_{II}$ and so $f_{II} = N_{II} - 1$. The whole process is then repeated as often as necessary until convergence.

**Example – sample size – ratio of two means**

Wooding (1994) gives an example of defining bioequivalence on a ratio scale and gives the planning mean log $C_{\text{Maximum}}$ as $\mu_{\text{Plan}} = 3.45$ and within-subject $SD$, $\sigma_{\text{Plan}} = 0.40$. Then with $\varepsilon = 0.1$, $\beta = 0.1$ and if we assume $\varepsilon = 0.22$, Table T1 and equation (8.1) gives

$$N_I = 2 \times \frac{0.12^2(1.6449 + 0.8416)^2}{0.2^2} = 4.45 \approx 5.$$  

From this $f_I = 5 - 1 = 4$ and from Table T9 of Student’s $t$-distribution, $t_{4,0.95} = 2.132$ and $t_{4,0.8} = 0.941$. Substituting these in equation (8.1) gives

$$N_{II} = 2 \times \frac{0.12^2(2.132 + 0.941)^2}{0.2^2} = 6.8 \approx 8$$

when rounded up to the next even integer. Now for $f_{II} = 8 - 1 = 7$, Table T9 gives $t_{7,0.95} = 1.895$ and $t_{7,0.8} = 0.896$. Substituting these in equation (8.1) gives

$$N_{III} = 2 \times \frac{0.12^2(1.895 + 0.896)^2}{0.2^2} = 5.6 \approx 6.$$  

Repeating this process once more finally gives $N_{\text{Bioequivalence}} = 8$. This implies that $m = 4$ subjects will be randomised to one sequence and 4 to the other in the planned crossover trial.
### Design features – PK studies

- Identify the type of subjects or patients required
- Decide on the summary characteristics of the PK profile required – $C_{\text{Maximum}}$, $t_{\text{Maximum}}$, $AUC$
- Consider carefully, the number and location of observations in time
- Ensure balance between experimental rigour and subject discomfort
- For bioequivalence – choose the value for equivalence on the ratio scale
- Anticipate the action to take if the full profile is not obtained in some subjects

---

#### 8.3 PHASE I TRIALS

In broad terms the aim of a Phase I trial is to establish the maximum tolerated dose (MTD) of a particular compound or treatment modality that can then be used in a subsequent Phase II trial to assess the corresponding activity. In some circumstances, the treatment under test may prove to be too toxic and so no MTD is established. In this case a Phase II trial would not be initiated for subsequent further testing. Underestimation of the MTD may lead to an apparent lack of efficacy at the later stages. Overestimation may lead to unacceptable toxicity (even death) in some patients. In either situation, a potentially useful compound may be shelved and opportunities for a therapeutic advance stalled.

---

#### THE MAXIMUM TOLERATED DOSE

For patients with a specific disease, one objective of treatment may be to reduce (or eradicate) the disease burden. However, it is recognised that any attack on the disease itself by a chemotherapeutic or other agent may bring collateral damage to normal tissue and vital organs. The usual strategy is to attempt to balance the two by establishing the concept of dose-limiting toxicity (DLT) by means of a Phase I trial. Such trials establish the dose at which DLT occurs and then step down from this dose by one step to define the MTD. The purpose is to establish the dose for use in any subsequent Phase II trial.

The level and type of DLT may be very specific to the clinical situation under investigation but should be defined before the Phase I trial commences. Once so determined, the presence or absence of such toxicity is recorded carefully when a patient receives the compound under study. However, this presumes that a first dose (say $d_{\text{Start}}$) has been identified for the design and that this has been given to the first patient.
Example – DLT and MTD – paediatric cancers

Shepherd, Burkes, Cormier et al. (1996) define the DLT as an absolute neutrophil count less than $0.5 \times 10^9$/L, or a platelet count less than $25 \times 10^9$/L, or any grade 3 non-haematological toxicity on the WHO scale. Following their Phase I trial, Estlin, Pinkerton, Lewis et al. (2001) recommended a MTD of 640 mg/m$^2$/day following DLT at 768 mg/m$^2$/day for nolatrexed dihydrochloride in children with advanced cancer.

ASSESSING TOXICITY

Since the assessment of toxicity is the key measure in these designs this must be defined very carefully indeed. This may be defined in general terms, for example, any WHO grade 3 toxicity of whatever type, or may very specific. It is very important that standard criteria are used to define the corresponding grades of toxicity. In oncology, a frequently adopted standard is set by the Common Terminology Criteria for Adverse Events (CTCAE) of the National Cancer Institute (2003).

CHOOSING THE DOES TO INVESTIGATE

In advance of the first patient being recruited in a Phase I trial, the investigators first identify the range of doses to use and all the specific dose levels to test. Thus $d_{\text{START}}$ will be one of these options and the ultimately identified MTD will also be one of these predefined doses. There are some difficulties with such an approach, since one is likely to start at low dose and then proceed dose-step by dose-step to successively higher doses. The choice of doses to investigate in humans will often depend on what has been observed in animal studies. These animal studies may have determined, for example, the MTD for a certain species of animal and experience has suggested that a reasonable starting dose for human studies may be one-tenth of this value. Caution (for safety reasons) may then dictate an even lower dose should be considered, but then there is concern that such low doses may be entirely innocuous and so could never be of benefit. The chances therefore of treating patients at potentially ineffective doses are clearly high. So, even with Phase I trials, there is concern that too low a dose may bring no potential benefit to the patient yet expose them to some (possibly high) risk.

Example – initial dose – cancer studies

Smith, Lee, Kantarjian et al. (1996) describe examples of initial doses for Phase I studies in patients with cancer as: 1/50 safe dose in mouse, 1/3 low-toxic dose in dog, 1/20 lethal dose (LD) in rat and less than 1/10 LD in mouse. It is clear that the design and conduct of the experiments leading to these recommendations are crucial.
Once the minimum dose to investigate, $d_{\text{MINIMUM}}$, is determined, then attention naturally turns to establishing what might be considered the therapeutic range and the setting of the maximum dose, $d_{\text{MAXIMUM}}$, for the study. Once these are established then the remaining doses to study will then also be determined.

For convenience we label the $k$ doses finally chosen as $d_1 = d_{\text{MINIMUM}}$, $d_2$, $d_3$, \ldots, $d_k = d_{\text{MAXIMUM}}$. However, we still need to choose $k$ and the specific values for each of the intermediate doses between the minimum and maximum values already defined. Statistical design considerations may suggest that these should be chosen equally spaced between $d_{\text{MINIMUM}}$ and $d_{\text{MAXIMUM}}$ on either a linear or a logarithmic scale. The doses may depend on how the drug is ‘packaged’ – perhaps in tablet form or vial of a certain volume where dose choice may be limited, or in a powder or liquid more easily constituted into any dose.

However, practice has often recognised that as the dose increases in equal steps it may become sequentially more and more toxic and hence possibly dangerous for the wellbeing of the patient. This caution has then led many investigators to decrease the step sizes as the dose increases. One method uses the Fibonacci series. Fibonacci (c. 1180 to c. 1250) was an Italian mathematician who first studied the following mathematical series: $a_0 = a_1 = 1$, then from $a_2$ onwards $a_{n+1} = a_n + a_{n-1}$. This gives the series: 1, 1, 2, 3, 5, 8, 13, 21, 34, etc. The corresponding Fibonacci ratios of successive terms are: $1/1 = 1$, $2/1 = 2$, $3/2 = 1.5$, $5/3 = 1.667$, $8/5 = 1.600$, $13/8 = 1.625$, $21/13 = 1.615$, $34/21 = 1.619$, \ldots, and eventually as $n$ gets larger and larger this approaches $1.33 = 2/(\sqrt{5} - 1)$. These ratios are shown in Table 8.1 and, for relatively small $n$ appropriate to the number of dose levels in a Phase I study, the ratio oscillates up and down. In mathematical terminology the series of ratios is not monotonically decreasing and so in fact does not provide successively decreasing step sizes. There is no theoretical reason why this or any other mathematical series should be chosen – they are merely empirical devices.

Nevertheless, it is usually regarded as desirable that successive doses are a decreasing multiplier of the preceding dose and thus (often without a clear explanation provided)
‘modified’ Fibonacci multipliers like those of Table 8.1 are substituted in practice. However, it is usually pragmatic considerations that determine the modifications and no systematic rationale underlies the changes.

**Example – Phase I trial design – nolatrexed dihydrochloride in advanced paediatric cancer**

In the Phase I study of nolatrexed dihydrochloride in children with advanced cancers conducted by Estlin, Pinkerton, Lewis *et al.* (2001) the corresponding protocol states: ‘The study is designed to incorporate a minimal number of patients in order to achieve the primary aim of a Phase I study, i.e. estimation of the MTD (Korn, Midthune, Chen *et al.*, 1994). At most dose levels, three to a maximum of six patients are to be included, so formal statistical analysis is not planned’.

In fact the doses actually recommended in this protocol, which are given in Table 8.1, are not entirely uniformly decreasing in terms of the ratio of successive doses, and only the first escalation corresponds to the limiting Fibonacci ratio of 1.33. No reason for the sequence chosen is explicit in the protocol itself.

**C33D**

A common, sequential, design is to choose a ‘low’ starting dose, perhaps with $d_{\text{START}} = d_{\text{MINIMUM}}$, and a fixed number of replicates (often 3). The choice of the next dose, $d_{\text{NEXT}}$, then depends on the number of patients (0, 1, 2 or 3) experiencing DLT. Clearly if no patients experience DLT then the subsequent dose to investigate will be higher than that just tested. This process continues until either the stopping level of DLT is attained in the successive groups of three patients or $d_{\text{MAXIMUM}}$ has been tested.

Commencing with the lowest dose:

(a) if zero of three experience DLT, then escalate to the next higher dose level.

(b) if one of three patients experience DLT, then add three more patients at that dose level:

(i) if zero of these three patients experience DLT (i.e. only one of six patients at the dose level), then escalate to the next higher dose level.

(ii) if one (or more) of these three patients experience DLT, then the MTD has been exceeded; three more are then added at the previous dose level (if only three patients had been treated previously at the prior level).

**Figure 8.4** Establishing the MTD in a C33D for a Phase I trial (after Smith, Bernstein, Bleyer *et al.*, 1998. Conduct of phase I trials in children with cancer. *Journal of Clinical Oncology*, 16, 966–978. [8]
In circumstances where the first two patients both experience DLT at a particular dose, it is not usual to give the third patient this same dose but to change the dose chosen to a lower one from the pre-specified dose range. Using this type of strategy Smith, Bernstein, Bleyer et al. (1998) state that the MTD from a Phase I design is established by adding cohorts of three patients at each dose level, and using the rules of Figure 8.4 to determine whether dose escalation should occur. This is known as a Cumulative ‘3+3’ Dose (C33D) approach and is one that is used by the cancer chemotherapy programme of the USA National Cancer Institute.

Although this process will (in general) establish the MTD it is only a pragmatic consideration that dictates that the Phase I trial should have tested at least six patients at $d_{MTD}$. This usually implies (as indicated in Figure 8.4) that once first identified, extra patients are then recruited and tested at this provisional $d_{MTD}$ until six patients in total have experienced this dose.

**Example – DLT and MTD – nolatrexed dehydrochloride in childhood cancer**

Estlin, Pinkerton, Lewis et al. (2001) report on the Phase I study conducted in children with advanced cancer the design of which was described earlier in Table 8.1. The three doses actually tested are given in Table 8.2 and were not those specified in the design. At the conclusion of this study, Estlin, Pinkerton, Lewis et al. (2001) recommended a MTD of 640 mg/m$^2$/day of nolatrexed dihydrochloride. However, it is clear from their report that DLT was observed with a dose of 768 mg/m$^2$/day, although only four rather than six patients required of the C33D design were accumulated at the recommended MTD of 640 mg/m$^2$/day.

However, practical (and ethical) issues usually constrain the size of Phase I trials and a maximum size in the region of 24 ($8 \times 3$) is often chosen. This multiple of 3 arises from the use of the C33D design. This implies that if predetermined doses are to be used, and the final dose chosen will have three extra patients tested, then $k=7$ dose options are the maximum that can be chosen for the design as $(k \times 3)+3=24$ patients.


<table>
<thead>
<tr>
<th>Patient</th>
<th>Dose (mg/m$^2$/day)</th>
<th>Dose escalation</th>
<th>DLT (0=No, 1=Yes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3</td>
<td>480</td>
<td></td>
<td>0/3</td>
</tr>
<tr>
<td>4, 5, 6, 7</td>
<td>640</td>
<td>1.33</td>
<td>0/4</td>
</tr>
<tr>
<td>8, 9, 10, 113</td>
<td>768</td>
<td>1.20</td>
<td>3/4</td>
</tr>
</tbody>
</table>
**Storer Design**

Storer (2001) describes what is essentially a modification to the C33D design by adding a stage before that design is implemented. The strategy is essentially to start the C33D process at a more informative dose than $d_{\text{MINIMUM}}$. This adjunct design suggests recruiting single individuals (rather than three) to successive doses and moving up and down the dose escalation scale according to whether or not a DLT is observed. The design moves into C33D once the current patient has not experienced a DLT and one previous patient has experienced DLT and one has not.

**Limitations**

The C33D design, with or without the Storer (2001) modification, has no real statistical basis, and more efficient alternatives have been sought. Efficiency here can be thought of as achieving the right MTD and with as few patients as possible. However, the design is easy to implement and requires little (statistical) manipulation – only keeping a count of the number of patients experiencing DLT at each dose tested. However, published studies appear to suggest that many variations from the basic C33D occur in practice. Indeed, Smith, Lee, Kantarjian et al. (1996) comment, following a review of Phase I studies conducted at the MD Anderson Cancer Center, Houston, USA, that: ‘investigators sometimes entered cohorts of patients at a dose intermediate between two previously tested levels’. This clearly makes designing Phase I trials somewhat problematic but perhaps unavoidable since critically ill patients are often involved. Nevertheless, such difficulties imply that the results need to be interpreted with due caution and carefully reviewed before taking the next step in the development process.

**CONTINUAL REASSESSMENT METHOD**

O’Quigley, Pepe and Fisher (1990) and O’Quigley (2001) have proposed the continual reassessment method (CRM) as an alternative to C33D. This design recruits the first patient to a dose closer to the centre of the range of pre-specified doses than the $d_{\text{MINIMUM}}$ of C33D. Essentially, if DLT is observed in this first patient then the next patient (Patient 2) is given the dose below $d_{\text{START}}$, whereas if no DLT is observed he or she receives the dose above $d_{\text{START}}$. Once this second patient receives the corresponding dose, and presence or absence of DLT is observed, the subsequent dose to utilise (which may be below, at or above the dose last used) is determined. However, at any stage of this process, the results from all individual patients so far recruited are utilised to provide the basis for the choice of the dose to be tested in the next patient recruited.

**Selecting the Doses**

The same process of selecting the range and actual dose in the C33D design is necessary for the CRM design. In addition, however, to implement CRM it is necessary to attach to each of these doses (based on investigator opinion) the probability of patients experiencing DLT at that dose. We label these probabilities $\theta_1, \theta_2, \theta_3, \ldots, \theta_k$. This prior elicitation of investigator opinion about toxicity leads to CRM being termed a Bayesian design.
It is implicit in the method of selecting these probabilities that, once they are assigned, then a ‘reasonable’ starting dose, $d_{\text{START}}$, would correspond to the dose that gives a value of $\theta_{\text{START}}$ close to some ‘acceptable’ value. This probability is often chosen as less than 0.3 – the 0.3 arising as a less than 1 in 3 chance, the ‘3’ coming from history associated with the use of C33D. The chosen $d_{\text{START}}$ would not usually correspond to the extremes $d_{\text{MINIMUM}}$ or $d_{\text{MAXIMUM}}$ of the dose range cited.

**Example – selecting the doses for CRM – non-Hodgkin’s lymphoma**

In the Phase I study of Flinn, Goodman, Post et al. (2000) summarised in Table 8.3, a dose-escalation strategy was utilised with decreasing multiples of the previous dose used. They defined minimum, $d_{\text{MINIMUM}}=40$, and maximum, $d_{\text{MAXIMUM}}=100$, doses with six 10 mg/m$^2$ steps. A CRM-based design was used and the investigator prior probabilities attached to each dose are given in Table 8.3. As might be expected, as the dose is increased the anticipated probability of DLT increases, so that with dose 40 mg/m$^2$, $\theta$ is only 0.05 (or anticipated to be seen in 1 in every 20 patients with this dose), whereas at dose 100 mg/m$^2$ $\theta$ is 0.8 (four in every five patients).


<table>
<thead>
<tr>
<th>Liposomal daunorubicin (mg/m$^2$)</th>
<th>Dose escalation</th>
<th>Prior probability of DLT, $\theta$</th>
<th>Number of patients recruited</th>
<th>Number of patients with DLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>—</td>
<td>0.05</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>50 (start)</strong></td>
<td>1.25</td>
<td>0.10</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>1.20</td>
<td>0.20</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>70</td>
<td>1.17</td>
<td>0.30</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>80</td>
<td>1.14</td>
<td>0.50</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>90</td>
<td>1.13</td>
<td>0.65</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>100</td>
<td>1.11</td>
<td>0.80</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

The $d_{\text{START}}=50$ mg/m$^2$ chosen corresponding to the prior probability of toxicity $\theta$ close to 0.1 and not the 0.3 we indicated as a common value to be used. A total of 20 patients were eventually included in total. Their final conclusion was that in patients with advanced non-Hodgkin’s lymphoma (NHL) the MTD for liposomal daunorubicin was 70–80 mg/m$^2$.

**Implementation**

Although the CRM method is more efficient than the C33D design it is considerably more difficult to implement, as the (statistical) manipulation required to determine the
next dose to use is technically complex and requires specialist computer statistical software such as that of Vernier, Brown and Thall (1999). The design reduces the number of patients receiving the (very) low dose options and thereby avoids patients receiving doses at which there is little prospect of them deriving benefit. Nevertheless, the design has been criticised by Korn, Midthune, Chen et al. (1994) for exposing patients to the risk of receiving potentially very toxic doses. However, modifications to the original design have been proposed to overcome both these difficulties (too low or too high) by Goodman, Zahurak and Piantadosi (1995) who suggest assigning more than one patient to each dose level chosen, and only allowing escalation/de-escalation by one dose level at a time.

**PRACTICALITIES**

**C33D or CRM**

The comparative features of the C33D and CRM designs are summarised in Figure 8.5.

**Small Size**

It has to be recognised that Phase I trials, however carefully designed, will include relatively few patients and so the corresponding level of uncertainty with respect to the

<table>
<thead>
<tr>
<th>C33D</th>
<th>CRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requires establishing the specific doses to be used at the design stage</td>
<td>Requires establishing the specific doses to be used at the design stage</td>
</tr>
<tr>
<td>Requires clinical opinion of the associated probability of toxicity at each of the chosen doses</td>
<td>Requires clinical opinion of the associated probability of toxicity at each of the chosen doses</td>
</tr>
<tr>
<td>For each patient, requires the presence of DLT to be determined</td>
<td>For each patient, requires presence of DLT to be determined</td>
</tr>
<tr>
<td>Dose for the next patient easily established</td>
<td>Dose for the next patient requires detailed calculation</td>
</tr>
<tr>
<td>Dose for the next patient uses information on all those so far included in the study</td>
<td>Usually requires fewer patients</td>
</tr>
<tr>
<td>Easy to explain</td>
<td>Difficult to explain</td>
</tr>
<tr>
<td>Requires no specialist statistical software</td>
<td>Requires specialist statistical software</td>
</tr>
</tbody>
</table>

*Figure 8.5  Basic features of the C33D and CRM designs*
true MTD will be high. It is also recognised that the designs do not (in one sense) estimate the MTD but rather choose one of the options presented by the investigators. This implies that very careful consideration needs to be given to the dose options available within the design. Further patients are not randomised to the doses chosen for investigation.

<table>
<thead>
<tr>
<th>Phase I – design and conduct issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clearly define patient eligibility</td>
</tr>
<tr>
<td>Clearly define the DLT</td>
</tr>
<tr>
<td>Establish the dose levels to be investigated</td>
</tr>
<tr>
<td>Choose the design, C33D or CRM</td>
</tr>
<tr>
<td>Consider the Storer option</td>
</tr>
<tr>
<td>If CRM, elicit the prior probabilities of DLT for each dose</td>
</tr>
<tr>
<td>Ensure that all patients are registered</td>
</tr>
<tr>
<td>Ensure all evaluations are made</td>
</tr>
<tr>
<td>Ensure the final report details information on all patients</td>
</tr>
</tbody>
</table>

### 8.4 PHASE II TRIALS

In contrast to Phase I trials, there are a relatively large number of alternative designs for Phase II trials. These include single-stage designs, in which a predetermined number of patients are recruited and two-stage designs, in which patients are recruited in two stages and the move to Stage 2 is consequential on the results observed in Stage 1. Multi-stage designs have been proposed, but the practicalities of having several decision points have limited their use because of the inherent further delays involved with each extra stage. Most Phase II trials are of a single-arm, non-comparative design. However, randomised Phase II selection designs, in which the objective is to select only one, the ‘best’, of several agents tested simultaneously, are strongly recommended in some situations.

Since most Phase II trials are single-arm experiments, Estey and Thall (2003) point out difficulties if the results are used for comparative purposes. Thus when different treatments are studied in separate single-arm trials, actual differences between response rates associated with the treatments (treatment effects) are confounded, as there is no randomisation to treatment, with differences between the trials (trial effects). Consequently an apparent treatment effect may in reality only be a trial effect.

In considering the design of a Phase II trial of a new drug, the investigators will usually have some knowledge of the activity of other drugs for the same disease. The anticipated response to the new drug is therefore compared, at the planning stage, with the observed responses to other therapies. This may lead to the investigators pre-specifying a response probability that, if the new drug does not achieve it, results in no further investigation. They might also have some idea of a response probability that, if
achieved or exceeded, would certainly imply that the new drug has activity worthy of further investigation, perhaps in a Phase III randomised trial to determine efficacy. As with Phase I designs, if a Phase II trial either fails to identify efficacy or overestimates the potential efficacy, there will be adverse consequences for the next stage of the development process.

ASSESSING RESPONSE

For a Phase I trial, the key endpoint measure is DLT as determined by pre-specified levels of toxicity and so an integral part of the design process is to define these precisely and, during the course of the trial, to carefully record their presence or absence. For Phase II studies, the endpoint is usually some measure of anti-disease activity and this translates into a measure of response. However, it is first essential to define what is meant exactly by response.

SINGLE-STAGE DESIGN

Fleming–A’Hern

Fleming (1982) proposed a single-stage procedure for Phase II trials in which a predetermined number of patients are recruited to the study and a decision about activity obtained from the number of responses observed amongst these patients alone. In constructing the design, the investigators are asked to determine the largest response proportion as $p_0$ which, if true, would clearly imply that the treatment does not warrant further investigation. For example, for a new anti-tumour drug this may be set at 0.1 because the investigators are aware of ‘many’ alternative agents for the same tumour that have at least that level of efficacy. The investigators are then asked to judge what is the smallest response proportion, $p_{New}$, that would, if demonstrated, imply the treatment warrants further investigation. For a new anti-tumour drug this may be set at 0.2, but will vary from circumstance to circumstance. This choice implies that, should the response rate turn out to be larger than this, then the agent under test would be worthy of future investigation, perhaps in a Phase III comparative trial. Obviously, between these limits, one is left in a position of not knowing quite how to proceed.

Study Size

The structure implied by defining $p_0$ and $p_{New}$, means that, in the Phase II trial itself, two, one-sided hypotheses, are to be tested. These are that the true response rate $\pi$ is either $\leq p_0$ or $\geq p_{New}$. It is then necessary to specify $\alpha$, the probability of rejecting the hypothesis $\pi \leq p_0$ when it is, in fact, true. Further, one specifies $\beta$, the probability of rejecting the hypothesis $\pi \geq p_{New}$ when that is true.

For this design, Fleming (1982) gives an approximation to the sample size required but A’Hern (2001) has used more exact methods for the calculation and his table of sample sizes should be used in all circumstances.
**Example – sample size using Fleming–A’Hern Phase II design – whole body hyperthermia**

To illustrate the differences in Phase II trial size for the Fleming design, A’Hern (2001) uses the study of Van der Zee, van Rhoon, Wike-Hooley et al. (1983) as an example. This describes a Phase II trial of whole-body hyperthermia in 27 patients with various cancers. He supposed a further Phase II trial is planned but only in patients with lung adenocarcinoma in which two of the three patients with this disease had complete remission in this trial. The investigators set the lowest response probability of interest to be \( \pi_0 = 0.15 \) and the treatment would be developed further only if the response was greater than \( \pi_{\text{New}} = 0.50 \). They also require a (one-sided) test size \( \alpha = 0.01 \) and a power \( 1 - \beta = 0.9 \).

Using these values in A’Hern (2001, Table 1) gives \( N_{\text{A’Hern}} = 21 \) with \( r_{\text{A’Hern}} \geq 8 \) responses for acceptance that the higher rate is more plausible. The approximations given by Fleming (1982) give \( N_{\text{Fleming}} = 18 \) and \( r_{\text{Fleming}} \geq 7 \) which suggest three fewer patients and one fewer response required. Such differences are clearly important for the design process.

**TWO-STAGE DESIGN**

In a single-stage design, all the patients are recruited before the response rate is calculated and the decision on level of efficacy made. Should the final response rate turn out to be low, then in a sense, the patients have been exposed to an ineffective regimen. Of course, we do not know this at the commencement of the trial but as the trial progresses interim information on activity does become available. The strategy of a two-stage design is to review this accumulating data (but not too often) so as to keep to a minimum the number of patients treated with the drug should it be ineffective. The implication, at the commencement of Stage 2, is that there is sufficient indication in Stage 1 that there is an acceptable minimum response rate that would enable (ethically) the continued use of this drug. The requirement is that a sufficient number of patients recruited to Stage 2 will be expected to obtain some benefit.

**Gehan Design**

In the approach suggested by Gehan (1961), a minimum requirement of efficacy \( \pi_0 \) is set, as with Fleming’s design, but patients are recruited in two stages, that is, \( n_{G1} \) in Stage 1 and a further \( n_{G2} (\geq 0) \) in Stage 2.

**Study Size**

In these circumstances the probability of \( n_{G1} \) successive patients failing on the drug in Stage 1, if its efficacy is exactly that of the minimum efficacy \( \pi_0 \), is

\[
\beta = (1 - \pi_0)^{n_{G1}}.
\]  

(8.3)
If the value of $\beta$ is specified by the design, then equation (8.3) can be rearranged to give

$$n_{G1} = \frac{\log \beta}{\log(1 - \pi_0)}, \quad (8.4)$$

as the Stage 1 sample size.

If no responses ($r_{G1} = 0$) are observed in Stage 1, no patients are recruited to Stage 2. In these circumstances the estimate of $\pi$ is $p_{G1} = r_{G1}/n_{G1}$ or 0%. This estimate of $\pi$ then has a 95% CI ranging from 0% to $[100 \times 1.96^2/(n_{G1} + 1.96^2)]\% \approx [400/(n_{G1} + 4)]\%$ using the methods described by Newcombe and Altman (2000). For example, if $n_{G1} = 14$ and $r_{G1} = 0$, then the upper limit of the 95% CI is from 0% to 22%, implying a great deal of uncertainty with respect to the true value of $\pi$ at this stage.

On the other hand, if $r_{G1} \geq 1$ responses are observed, then the size of the recruitment to Stage 2 depends on their actual number. Assuming that once Stage 1 is complete, $r_{G1} (\geq 1)$ responses are observed, then the estimated response rate is $p_{G1} = r_{G1}/n_{G1}$ and a further $n_{G2}$ patients are recruited to Stage 2. This gives a total of $N_{Gehan} = n_{G1} + n_{G2}$ patients in all. The value of $n_{G2}$ is chosen to give a required value of the standard error, $SE(p)$ for the final estimate, $p$, of the true activity $\pi$, that is,

$$SE(p) = \sqrt{\frac{p(1-p)}{n_{G1} + n_{G2}}} = \varepsilon, \quad (8.5)$$

where $\varepsilon$ is set by the investigating team. Rearranging equation (8.5), the required number of patients for the second stage is

$$n_{G2} = \frac{p(1-p)}{\varepsilon^2} - n_{G1}. \quad (8.6)$$

However, at the end of Stage 1, we do not know the final estimate $p$, only $p_{G1} = r_{G1}/n_{G1}$, the proportion of successes in the first stage. Thus, to estimate $n_{G2}$ from equation (8.6), we must use $p_{G1}$ rather than $p$. However, $n_{G1}$ is usually so small that the resulting $p_{G1}$ will be very imprecise. As a consequence, rather than using $p_{G1}$ to replace $p$ in equation (8.6), Gehan used $\pi_{UL}$ the one-sided upper (arbitrarily chosen) 75% confidence limit for $\pi$ obtained at the end of Stage 1. This gives the estimate for the Stage 2 sample size.

$$n_{G2} = \frac{\pi_{UL}(1 - \pi_{UL})}{\varepsilon^2} - n_{G1}. \quad (8.7)$$

This depends rather critically on the number of successes $r_{G1}$ observed in Stage 1 of the trial.

If there are $r_{G2}$ responses in Stage 2, then $p = (r_{G1} + r_{G2})/(n_{G1} + n_{G2})$ is the final estimate of the activity of the drug based on $N_{Gehan}$ patients.

### Example – sample size using Gehan Phase II design – non-responsive breast cancer

Lehnert, Mross, Schueller et al. (1998) used the Gehan design for a Phase II trial of the combination dexverapamil and epirubicin in patients with breast cancer. For Stage 1 they set $\pi_0 = 0.2$ and $\beta = 0.05$, obtaining $n_{G1} = 14$. Of these
14 patients, $r_{G1} = 3$ responses were observed, then their requirement of $\varepsilon = 0.1$ implies a further $n_{G2} = 9$ patients were to be recruited. Finally a total of four (17.4%) responses was observed from the $N_{\text{Gehan}} = n_{G1} + n_{G2} = 23$ patients, the result giving a 95% CI for $\pi$ from 7 to 37%.

**Simon – Optimal and Mini-max Designs**

As is clear, from the chosen upper limit of a 75% CI used by Gehan (1961) to determine the number of patients to enter Stage 2 of his design, rather arbitrary assumptions are made when developing statistical designs for Phase II trials. Thus Simon (1989) describes two, two-stage designs with somewhat different properties from those of Gehan. He describes a Phase II design that is ‘optimal’ for Stage 1, in that the sample size is minimised for that stage if the regimen has low activity. The second, or ‘mini-max’, design aims to minimise the maximum total (Stage 1 plus Stage 2) sample size, $N$.

As with the Fleming design, these designs specify the parameters $\pi_0$ and $\pi_{\text{New}}$, where again $\alpha$ is the probability of rejecting the hypothesis $\pi \leq \pi_0$ when it is in fact true and $\beta$ the probability of rejecting the hypothesis $\pi \geq \pi_{\text{New}}$ when that is true. Simon establishes his designs by checking for every total sample size $N$, each possible division into two stages, those that satisfy these conditions. From these he then chooses for the ‘optimal’ design that which has the smallest Stage 1 sample size, and for the ‘mini-max’ that with the minimal total sample size. For each of these designs, the corresponding sample size for each stage is given together with the minimum number of responses required to trigger the start of Stage 2 and the total number of responses required to suggest activity.

**Example – Simon’s optimal design – advanced non-small-cell lung cancer**

Baldini, Tibaldi, Ardizzoni et al. (1998) used a Simon’s optimal two-stage design so as to minimise the expected number of patients to be accrued in the case of low activity, in which case, only Stage 1 would be implemented. The use of ‘expected (total) number’ refers to the statistical properties of the design as one does not know while planning the study if Stage 2 will, or will not, be implemented. They state, with some notational changes, in the statistical methods section of their paper: ‘Sample size was calculated on the following assumptions: $\alpha = 0.05$, $\beta = 0.1$; $\pi_0$ (clinically uninteresting true response rate) and $\pi_{\text{New}}$ (sufficiently promising true response rate), defined according to Simon, were set at 10% and 30% respectively’.

This design implied recruiting 18 patients to Stage 1 and if two or fewer responses were observed, the accrual had to be stopped. Otherwise, 17 more patients were to be accrued in Stage 2. The drug combination was considered of interest if seven or more responses were observed out of 35 evaluable patients.
Example – Simon’s mini-max design – metastatic nasopharyngeal cancer

Foo, Tan, Leong et al. (2002) used the mini-max design of Simon (1989) for two Phase II trials that were to be conducted in parallel. In one study chemonaive patients with metastatic nasopharyngeal cancer were recruited and, in the other, those who had received previous chemotherapy for their disease. The investigators determined the design parameters, \( \pi_0 \) and \( \pi_{\text{New}} \) for each trial separately as summarised in Table 8.4, and for both studies set \( \alpha = 0.05 \) and \( \beta = 0.2 \).

Table 8.4 Simon mini-max designs utilised, and results obtained, by Foo, Tan, Leong et al. (2002) in patients with metastatic nasopharyngeal cancer

<table>
<thead>
<tr>
<th>Previous chemotherapy</th>
<th>( \pi_0 )</th>
<th>( \pi_{\text{New}} )</th>
<th>( n_{S1} )</th>
<th>( r_{S1} )</th>
<th>( r )</th>
<th>( n_{S2} )</th>
<th>( N_{\text{Simon}} )</th>
<th>( r_S )</th>
<th>( R (%) )</th>
<th>95% CI for ( \pi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>0.1</td>
<td>0.3</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>25</td>
<td>6</td>
<td>7 (28.0)</td>
<td>15.9 to 44.4</td>
</tr>
<tr>
<td>Yes</td>
<td>0.05</td>
<td>0.2</td>
<td>13</td>
<td>1</td>
<td>7</td>
<td>14</td>
<td>27</td>
<td>4</td>
<td>13 (48.1)</td>
<td>33.2 to 63.4</td>
</tr>
</tbody>
</table>

The Stage 1 results of Table 8.4 allowed both studies to proceed to Stage 2 since \( r > r_{S1} \) in both cases. However, at the close of Stage 2 for the chemonaive patients, efficacy was just claimed with a total of \( R = 7 \) responses observed against the requirement of \( r_S = 6 \). In contrast, for the previously treated group, efficacy was clearly established with \( R = 13 \) responses observed against the requirement of \( r_S = 4 \).

When deciding on which of the two Simon (1989) designs to use for a study, the design team need to balance two consequences: (1) the undesirable prospect of giving too many patients what turns out to be an ineffective drug against (2) minimising the total number of patients necessary to complete the Phase II design. Clearly, if no effective drug is available for the disease under consideration then one may not be so concerned with the first of these and would prefer to keep the overall study size to a minimum.

Tan–Machin Single- and Dual-threshold Designs

In the Phase II designs discussed, the final response rate is estimated by \( R/N \), where \( R \) is the total number of responses from the total number of patients recruited \( N \) (whether obtained from a single- or two-stage design). This response rate, together with the corresponding 95% CI, provides the basic information for the investigators to decide if a subsequent Phase III trial is warranted. However, even after the trial is completed, as in the examples of Table 8.4, there often remains considerable uncertainty about the true value of \( \pi \). Thus even for the high response rate of 48.1% observed in Table 8.4 the corresponding 95% CI is consistent with a true response rate as small as 33% and one
as high as 63%, an almost twofold difference. Thus we would not be confident for the chemonaive patients that \( \pi > \pi_{\text{New}} = 0.3 \). Although for the previously treated patient we may be reasonably confident that \( \pi > \pi_{\text{New}} = 0.2 \).

One consequence of this type of uncertainty led Tan and Machin (2002) to argue that what is of key relevance to the decision as to whether to proceed to a Phase III trial, is the knowledge of the probability that the response rate, \( \pi \), is greater than, for example, \( \pi_{\text{New}} \). Thus in their two-stage single-threshold design (STD) the investigator first sets the target response rate \( \pi_{\text{New}} \), (which they denote as \( R_U \)) and \( \pi_{\text{Prior}} \), the anticipated value of the drug being tested. However, in place of \( \alpha \) and \( \beta \), \( \lambda_1 \) the required threshold probability following Stage 1 that \( \pi > \pi_{\text{New}} \) and \( \lambda_2 \) the required threshold probability after completion of Stage 2 that \( \pi > \pi_{\text{New}} \) are specified. Further, once the first stage of the trial is completed, the estimated value of \( \lambda_1 \), that is \( l_1 \), can be computed and, should the trial continue to Stage 2, then, on trial completion, \( l_2 \) can be computed.

**Example – Tan–Machin Phase II STD design – gemcitabine in metastatic nasopharyngeal cancer**

Tan, Machin, Tai et al. (2002) re-analysed the Phase II trial of Foo, Tan, Leong et al. (2002) for previously treated patients as if they had been designed by the methodology of Tan and Machin (2002). First they back-calculated from the two-stage Simon mini-max design utilised, that this choice implied for their STD values of \( \lambda_1 = 0.728 \) and \( \lambda_2 = 0.774 \) respectively. Using the actual trial data, they then compute \( l_1 = 0.997 \) (which is clearly greater than \( \lambda_1 = 0.728 \)) and \( l_2 = 0.999 \) (which is clearly greater than \( \lambda_2 = 0.774 \)). So had the STD been used this re-analysis suggests that, at the end of Stage 1 continuation to Stage 2 would have been appropriate. Further information at the end of Stage 2 recommended that gemcitabine was considered to have sufficient activity for Phase III evaluation.

In the two-stage dual-threshold design (DTD) of Tan and Machin (2002) the investigator first sets the target response rates of \( \pi_0 \) and \( \pi_{\text{New}} \) (which they denote as \( R_L \) and \( R_U \)) as with the Fleming design. They then set \( \pi_{\text{Prior}} = (\pi_0 + \pi_{\text{New}})/2 \) as the anticipated value of the drug being tested. Again, in place of \( \alpha \) and \( \beta \), \( \lambda_1 \) is now set as the required threshold probability following Stage 1 that \( \pi < \pi_0 \), while \( \lambda_2 \) remains the required threshold probability after completion of Stage 2 that \( \pi > \pi_{\text{New}} \). Again, once the first stage of the trial is completed, the estimated value of \( \lambda_1 \), that is \( l_1 \), can be computed and should the trial continue to Stage 2 then, on its completion trial, \( l_2 \) can be computed. The latter is then used to help make the decision whether or not a Phase III trial is suggested.
Example – Tan–Machin Phase II DTD design – gemcitabine in metastatic nasopharyngeal cancer

For the chemonaive study of Foo, Tan, Leong et al. (2002), the actual trial data gives $l_1 = 0.01$ and $l_2 = 0.37$. These can be equivalently expressed by the probability that $\pi > \pi_0$ is $1 - l_1 = 0.99$ (1%) but the probability of being greater then $\pi_{\text{New}}$ is 0.37 (37%). These together imply that the response rate is truly in the region of uncertainty, $\pi_0 \leq \pi \leq \pi_{\text{New}}$, has a high probability of 62%.

Tan and Machin (2002) suggest planning values for $(\lambda_1, \lambda_2)$ as (0.6, 0.7), (0.6, 0.8) or (0.7, 0.8). These imply, for this trial, that Stage 2 should commence since $(l_1 = 0.99 > 0.6)$, as was indeed the case, but gemcitabine should not be recommended for a Phase III trial on the basis of the final evidence available since $l_2 = 0.37 < 0.7$.

Bryant and Day – Toxicity and Response Design

Bryant and Day (1995) point out that a common situation when considering Phase I and Phase II trials is that although the former primarily focuses on toxicity and the latter on efficacy, each in fact considers both. This provides the rationale for their Phase II design which incorporates toxicity and activity considerations. Essentially they combine the optimal two-stage Simon design for activity with a similar design for toxicity where one is looking for acceptable toxicity but high activity.

Their design implies the same, two, one-sided hypotheses, are to be tested as for the Fleming and Simon designs which are that the true rates $\pi$ are either $\leq \pi_0$ or $\geq \pi_{\text{New}}$. But now, these values have to be set for both response and toxicity. It is then necessary to specify $\alpha_R$ the probability of rejecting the hypothesis $\pi_R \leq \pi_{R0}$ and similarly $\alpha_T$ for the hypothesis $\pi_T \leq \pi_{T0}$ when they are, in fact, true and $\beta$ is set as the probability of failing to recommend a treatment that is acceptable with respect to activity and toxicity. Since both toxicity and response are assessed in the same patient, the distributions of response and toxicity are not independent, and these two are linked by means of

$$
\phi = \frac{\eta_{00} \eta_{11}}{\eta_{01} \eta_{10}}.
$$

(8.8)

Here $\eta_{00}$ is the true proportion of patients who both fail to respond and also experience unacceptable toxicity, $\eta_{01}$ is the proportion of patients who fail but have acceptable toxicity, $\eta_{10}$ is the proportion of patients who respond but who have unacceptable toxicity, and finally $\eta_{11}$ is the proportion of patients who respond and also have acceptable toxicity. Fortunately the designs suggested by Bryant and Day (1995) turn out to be little affected by the magnitude of $\phi$, and so in Table T10 levels of toxicity and activity are assumed independent, in which case $\phi = 1$. For pragmatic reasons, when selecting the designs, Bryant and Day (1995) restrict their choice to those for which the size of Stage 2 is $n_2 \leq 1.25 n_1$.  


Example – Bryant–Day toxicity and response Phase II design – ifosfamide and vinorelbine in ovarian cancer

González-Martín, Crespo, García-López et al. (2002) used the Bryant and Day two-stage design with a cutoff point for the response rate of 10% and for severe toxicity, 25%. Severe toxicity was defined as grade 3–4 non-haematological toxicity, neutropenic fever or grade 4 thrombocytopenia. They do not provide full details of how the sample size was determined but their choice of design specified a Stage 1 of 14 patients and Stage 2 a further 20 patients. In the event, in these advanced platinum-resistant ovarian cancer patients, the combination of ifosfamide and vinorelbine was evidently very toxic. Hence the trial was closed after 12 patients with an observed toxicity level above the 25% contemplated.

Randomised ‘Selection’ Designs

In situations where there is more than one agent available for Phase II testing and all (or at least several) of them prove to be potentially worthwhile there is a difficulty in proceeding to the Phase III stage. This is because with many options it may not be possible to test all of them against the current standard treatment for the disease in a definitive Phase III trial as the sample size then required for a multi-arm trial would be unacceptably large. Thus an alternative strategy is to first screen the new therapies in a Phase II trial, but in a design setting the aim of which is to select only one to test in Phase III. In this screen of two or more agents, patients are assigned at random to the alternatives in the Phase II trial. Such ‘selection’ designs have been proposed by Simon, Wittes and Ellenberg (1985), but their use is really confined to agents or combinations of agents that indicate real promise from earlier studies. This approach chooses the observed best treatment for the Phase III trial, however small the advantage over the others. The trial size is determined in such a way that if a treatment exists for which the underlying efficacy is superior to the others by a specified amount, then it will be selected with a high probability.

Trial Size

Table T11 gives the sample size requirements for randomised Phase II selection designs with binary outcomes with $g=2, 3$ and 4 groups. The improvement in response rate in one group (labelled group $g$ for convenience) is anticipated to be at least $\delta=0.15$ or 0.20 over the remainder (termed the baseline).

Except in extreme cases, when $\pi_i$ is small or large, Table T11 indicates the sample size is relatively insensitive to these baseline response rates, that is, the response rates of groups 1 through to $g-1$. Since precise knowledge of these may not be available, a conservative approach to trial design is to always use the largest sample size for each $g$. For example, with $\delta=0.15$ (which may result from many possibilities for the...
components of $\pi_{\text{New}} - \pi_i$ but all leading to the same value of $\delta$) use the row of Table T11 giving the largest number of patients. This is the row with $\pi_i=0.45$, $\pi_{\text{New}}=0.60$ giving 37, 55 and 67 patients per group for $g=2$, 3 and 4, respectively. Similarly with $\delta=0.20$ use the row with $\pi_i=0.40$, $\pi_{\text{New}}=0.60$ giving 21, 31 and 38 patients per group. When randomisation is conducted, the $g$ groups form a natural block size. For example, if four compounds are to be compared the experimental design may be configured in a way similar to Figure 4.3. In this case there would be $b=38$ balanced blocks of size four each containing the $t=4$ different treatments (compounds), or alternatively $b=19$ balanced blocks of size 8, with $r=2$ patients receiving each of the $t=4$ treatments.

Unfortunately, with $g \geq 4$ groups these designs lead to relatively large randomised trials and this may limit their usefulness.

**Example – randomised Phase II design – non-Hodgkin’s lymphoma**

Itoh, Ohtsu, Fukuda *et al.* (2002) describe a randomised two-group Phase II trial comparing dose-escalated (DE) with biweekly (dose-intensified) CHOP (DI) in newly diagnosed patients with advanced-stage aggressive non-Hodgkin’s lymphoma. Their design anticipated at least a 65% complete response rate (CR) in both groups. To achieve a 90% probability of selecting the better arm when the CR rate is 15% higher in one arm than the other, at least 30 patients would be required in each arm. [The more detailed tabulations of Table T11 give 29 as opposed to 30.]

In the event, they recruited 35 patients to each arm and observed response rates with DE and DI of 51% and 60% respectively. The follow-on study, a randomised Phase III trial, compares DI CHOP with the standard CHOP regimen.

**WHICH DESIGN TO USE**

With such a plethora of different options for Phase II designs, it is clearly important that the investigators choose that which is best for their purpose. In some cases the choice will be reasonably clear, for example, if one has several compounds to test at the same time then the randomised selection design will be preferred to (say) a series of parallel single-arm studies. In other circumstances, the patient pool may be very limited and a key consideration will be the maximum numbers of patients that might have to be recruited. Features to guide investigators in their choice are summarised in Figure 8.6.

The essential difference between the Tan–Machin designs and the others is that in the former the statistical design parameters are set through $\lambda_1$ and $\lambda_2$ rather than $\alpha$ and $\beta$. 
Phase II – design and conduct issues

Clearly define patient eligibility

Clearly define the measures of response (and toxicity)

Choose a single- or two-stage design

Consider the importance of not proceeding to Stage 2 if activity low

Consider whether a CI or threshold probability approach is to be used for interpretation

Consider the possibility of a randomised selection design

Ensure that all patients are registered

Ensure all evaluations are made

Ensure the final report details information on all patients
8.5 PRACTICALITIES

Although PK, Phase I and Phase II studies are often of modest or even small size, the temptation to conduct these studies without due attention to detail should be resisted. In fact, these studies (imprecise though they may be) provide key information for the drug development process. It is therefore essential that they are carefully designed, painstakingly conducted and meticulously reported in full.

Although we have discussed design, care must also be taken to prepare for the unexpected to occur. Perhaps a level or type of toxicity not anticipated may occur and one should think of ways in which the basic design may have to be modified in such an eventuality.

It is also important that all patients are registered for the trial (and hence are in the trial database) and that the final report includes information on all these patients. This is particularly important if a review process of, for example, each objective response in a Phase II trial reveals that certain patients admitted to the trial either were not truly eligible, or had not received the full treatment as specified by the protocol or could not be evaluated for the endpoint. Perhaps it is unclear whether or not they had sufficient tumour shrinkage for a satisfactory response. It must be clear in the study protocol itself, and in the subsequent report of the study results, whether these ‘ineligible’, ‘non-compliant’ and ‘non-evaluable’ patients are or are not included in the reported response rates. This equally applies for any assessment of toxicity, whether or not toxicity is a formal endpoint for the design as it is in Phase I studies and the Bryant–Day design of Phase II.

Recruitment

One difficulty with some Phase I designs is that the results from each patient must be known before the dose for the next patient can be determined. This almost certainly implies inbuilt delays in the recruitment process and hence studies of lengthy duration. For the same reason, there may be delay between Stage 1 and Stage 2 of a two-stage Phase II design. However, continuous monitoring of the patient responses may trigger Stage 2 before the formal recruitment to Stage 1 is complete, if there are already sufficient responses. However, this may be difficult in a multicentre setting and so a formal review, once Stage 1 is complete, may be justified before embarking on Stage 2.

For both Phase I and II trials of any design, in most circumstances patient numbers, and often response rates, are quite low; so investigators should always use exact confidence interval methods when reporting their results.

8.6 TECHNICAL DETAILS

THE BINOMIAL DISTRIBUTION

In Phase II trials, the underlying reponse rate is assumed to be constant and to have probability $p$. If $N$ patients are recruited to a study then $r = 0, 1, \ldots, N$ responses may be observed. The probability of $r$ responses from $N$ patients is given by the binomial distribution as
\[ P(r) = \frac{N!}{r!(N-r)!} \pi^r (1-\pi)^{N-r}, \quad (T8.1) \]

where \( r! = 1 \times 2 \times 3 \times \ldots \times (r-1) \times r \) and if \( r=0 \), then \( r!=0!=1 \). This distribution has mean \( \mu = N\pi \) and standard deviation \( \sigma = \sqrt{[N\pi(1-\pi)]} \).

In circumstances when \( N \) is reasonably large, and \( \pi \) is not close to 0 or to 1, this can be approximated to by a Normal distribution with the same mean, \( \mu \), and standard deviation, \( \sigma \). This was the approximation used by Fleming (1982). However, in Phase II trials \( N \) may not be large and small values of \( \pi \) may be anticipated so A’Hern (2001) used the exact binomial probabilities themselves when tabulating the designs.

CONFIDENCE INTERVALS

If a Phase II study is conducted in \( N \) subjects and \( r \) patients respond, then the estimate of the true proportion of responses \( \pi \), is given by \( p = r/N \) and this has a standard error (SE) estimated by

\[ SE(p) = \sqrt{\frac{p(1-p)}{N}}. \quad (T8.2) \]

The use of the Normal distribution leads to the following approximate 100(1-\( \alpha \))% CI for the true probability \( \pi \) as

\[ p - z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{N}} \quad \text{to} \quad p + z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{N}}. \quad (T8.3) \]

However, as we have pointed out, when \( N \) is small, as will often be the case, and particularly if \( \pi \) is small (as it may be in Phase II trials), Newcombe and Altman (2000) provide exact CIs and these should replace equation (T8.3) in all circumstances.