CHAPTER 1
Basic ideas in clinical trial design

1.1 Historical perspective

As many of us who are involved in clinical trials will know, the randomised controlled trial is a relatively new invention. As pointed out by Pocock (1983) and others, very few clinical trials of the kind we now regularly see were conducted prior to 1950. It took a number of high-profile successes plus the failure of alternative methodologies to convince researchers of their value.

Example 1.1 The Salk Polio Vaccine trial

One of the largest trials ever conducted took place in the USA in 1954 and concerned the evaluation of the Salk polio vaccine. The trial has been reported extensively by Meier (1978) and is used by Pocock (1983) in his discussion of the historical development of clinical trials.

Within the project, there were essentially two trials, and these clearly illustrated the effectiveness of the randomised controlled design.

Trial 1: Original design: Observed control
1.08 million children from selected schools were included in this first trial. The second graders in those schools were offered the vaccine, while the first and third graders would serve as the control group. Parents of the second graders were approached for their consent, and it was noted that the consenting parents tended to have higher incomes. Also, this design was not blinded so that both parents and investigators knew which children had received the vaccine and which had not.

Trial 2: Alternative design: Randomised control
A further 0.75 million children in other selected schools in grades one to three were to be included in this second trial. All parents were approached for their consent, and those children where consent was given were randomised to receive either the vaccine or a placebo injection. The trial was double blind with parents, children and investigators unaware of who had received the vaccine and who had not.

The results from the randomised controlled trial were conclusive. The incidence of paralytic polio, for example, was 0.057 per cent in the placebo group compared to 0.016 per cent in the active group, and there were four deaths in the placebo group compared to none in the active group. The results from the observed control trial, however, were less convincing with a smaller observed difference (0.046 per cent vs. 0.017 per cent). In addition, in the cases where

Statistical Thinking for Non-Statisticians in Drug Regulation, Second Edition. Richard Kay.

consent could not be obtained, the incidence of paralytic polio was 0.036 per cent in the randomised trial and 0.037 per cent in the observed control trial, event rates considerably lower than those among placebo patients and in the untreated controls, respectively. This has no impact on the conclusions from the randomised trial, which is robust against this absence of consent; the randomised part is still comparing like with like. In the observed control part however, the fact that the *no consent* (grade 2) children have a lower incidence than those children (grades 1 and 3) who were never offered the vaccine potentially causes some confusion in a non-randomised comparison; does it mean that grade 2 children naturally have lower incidence than those in grades 1 and 3? Whatever the explanation, the presence of this uncertainty reduced confidence in the results from the observed control trial.

The randomised part of the Salk Polio Vaccine trial has all the hallmarks of modern-day trials – randomisation, control group and blinding – and it was experiences of these kinds that helped convince researchers that only under such conditions can clear, scientifically valid conclusions be drawn.

1.2 Control groups

We invariably evaluate our treatments by making comparisons – active compared to control. It is very difficult to make absolute statements about specific treatments, and conclusions regarding the efficacy and safety of a new treatment are made relative to an existing treatment or placebo.

ICH E10 (2001): 'Note for Guidance on Choice of Control Group in Clinical Trials'

'Control groups have one major purpose: to allow discrimination of patient outcomes (for example, changes in symptoms, signs, or other morbidity) caused by the test treatment from outcomes caused by other factors, such as the natural progression of the disease, observer or patient expectations, or other treatment'.

Control groups can take a variety of different forms; here are just a few examples of trials with alternative types of control group:

- Active versus placebo
- Active A versus active B (vs. active C)
- Placebo versus dose level 1 versus dose level 2 versus dose level 3 (dose finding)
- Active A + active B versus active A + placebo (add-on)

The choice will depend on the objectives of the trial.

Open trials with no control group can nonetheless be useful in an exploratory, maybe early phase setting, but it is unlikely that such trials will be able to provide confirmatory, robust evidence regarding the performance of the new treatment.

Similarly, external concurrent or historical controls (groups of subjects external to the study either in a different setting or previously treated) cannot

provide definitive evidence in most settings. We will discuss such trials in Chapter 17. The focus in this book however is the randomised controlled trial.

1.3 Placebos and blinding

It is important to have blinding of both the subject and the investigator wherever possible to avoid unconscious bias creeping in, either in terms of the way a subject reacts psychologically to treatment or in relation to the way the investigator interacts with the subject or records subject outcome.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Blinding or masking is intended to limit the occurrence of conscious or unconscious bias in the conduct and interpretation of a clinical trial arising from the influence which the knowledge of treatment may have on the recruitment and allocation of subjects, their subsequent care, the attitudes of subjects to the treatments, the assessment of the endpoints, the handling of withdrawals, the exclusion of data from analysis, and so on'.

Ideally, the trial should be *double-blind* with both the subject and the investigator being blind to the specific treatment allocation. If this is not possible for the investigator, for example, then the next best thing is to have an independent evaluation of outcome, both for efficacy and for safety. A *single-blind* trial arises when either the subject or investigator, but not both, is blind to treatment.

An absence of blinding can seriously undermine the validity of an endpoint in the eyes of regulators and the scientific community more generally, especially when the evaluation of that endpoint has an element of subjectivity. In situations where blinding is not possible, it is important to use hard, unambiguous endpoints and to use independent recording of that endpoint.

The use of placebos and blinding goes hand in hand. The existence of placebos enables trials to be blinded and accounts for the placebo effect – the change in a patient's condition that is due to the act of being treated, but is not caused by the active component of that treatment.

Note that having a placebo group does not necessarily imply that one group is left untreated. In many situations, oncology is a good example, the experimental therapy/placebo is added to an established active drug regimen; this is the add-on study.

1.4 Randomisation

Randomisation is clearly a key element in the design of our clinical trials. There are two reasons why we randomise subjects to the treatment groups:

- To avoid any bias in the allocation of the patients to the treatment groups
- To ensure the validity of the statistical test comparisons

Randomisation lists are produced in a variety of ways, and we will discuss several methods later. Once the list is produced, the next patient entering the trial receives the next allocation within the randomisation scheme. In practice, this process is managed by *packaging* the treatments according to the predefined randomisation list.

There are a number of different possibilities when producing these lists:

- Unrestricted randomisation
- Block randomisation
- Unequal randomisation
- Stratified randomisation
- Central randomisation
- Dynamic allocation and minimisation
- Cluster randomisation

1.4.1 Unrestricted randomisation

Unrestricted (or simple) randomisation is simply a random list of, for example, As and Bs. In a moderately large trial, with, say, $n = 200$ subjects, such a process will likely produce approximately equal group sizes. There is no guarantee however that this will automatically happen and in small trials, in particular, this can cause problems.

1.4.2 Block randomisation

To ensure balance in terms of numbers of subjects, we usually undertake *block randomisation* where a randomisation list is constructed by randomly choosing from the list of potential blocks. For example, there are six ways of allocating two As and two Bs in a *block* of size four:

AABB, ABAB, ABBA, BAAB, BABA, BBAA

and we choose at random from this set of six blocks to construct our randomisation list, for example,

ABBA BAAB ABAB ABBA, …

Clearly, if we recruit a multiple of four patients into the trial, we will have perfect balance and approximate balance (which is usually good enough) for any sample size.

In large trials, it could be argued that block randomisation is unnecessary. In one sense, this is true; overall balance will be achieved by chance with an unrestricted randomisation list. However, it is usually the case that large trials will be multi-centre trials, and not only is it important to have balance overall, but it is also important to have balance within each centre. In practice, therefore, we would allocate several blocks to each centre, for example, five blocks of size four if we are planning to recruit 20 patients from each centre. This will ensure balance within each centre and also overall.

How do we choose block size? There is no magic formula, but more often than not, the block size is equal to two times the number of treatments.

What are the issues with block size?

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Care must be taken to choose block lengths which are sufficiently short to limit possible imbalance, but which are long enough to avoid predictability towards the end of the sequence in a block. Investigators and other relevant staff should generally be blind to the block length…'.

Shorter block lengths are better at producing balance. With two treatments, a block length of four is better at producing balance than a block length of 12. The block length of four gives perfect balance if there is a multiple of four patients entering, whereas with a block length of 12, perfect balance is only going to be achieved if there are a multiple of 12 patients in the study. The problem, however, with the shorter block lengths is that this is an easy code to crack and inadvertent unblinding can occur. For example, suppose a block length of four was being used in a placebo-controlled trial and also assume that experience of the active drug suggests that many patients receiving that drug will suffer nausea. Suppose the trial begins and the first two patients suffer nausea. The investigator is likely to conclude that both these patients have been randomised to active and that therefore the next two allocations are to placebo. This knowledge could influence his/her willingness to enter certain patients into the next two positions in the randomisation list, causing bias in the mix of patients randomised into the two treatment groups. Note the comment in the ICH guideline regarding keeping the investigator (and others) blind to the block length. While in principle this comment is sound, the drug is often delivered to a site according to the chosen block length, making it difficult to conceal information on block size. If the issue of inadvertent unblinding is going to cause problems, then more sophisticated methodologies can be used, such as having the block length itself varying, perhaps randomly chosen from two, four or six.

1.4.3 Unequal randomisation

All other things being equal, having equal numbers of subjects in the two treatment groups provides the maximum amount of information (the greatest power) with regard to the relative efficacy of the treatments. There may, however, be issues that override statistical efficiency:

- It may be necessary to place more patients on active compared to placebo in order to obtain the required safety information.
- In a three-group trial with active A, active B and placebo (P), it may make sense to have a 2:2:1 randomisation to give more power for the A versus B comparison as that difference is likely to be smaller than the A versus P and B versus P differences.

Unequal randomisation is sometimes needed as a result of these considerations. To achieve this, the randomisation list will be designed for the second example with double the number of A and B allocations compared to placebo.

For unequal randomisation, we would choose the block size accordingly. For a 2:1 randomisation to A or P, we could randomly choose from the blocks:

AAP, APA, PAA

1.4.4 Stratified randomisation

Block randomisation therefore forces the required balance in terms of the numbers of patients in the treatment groups, but things can still go wrong. For example, let's suppose in an oncology study with time to death as the primary endpoint that we can measure baseline risk (say, in terms of the size of the primary tumour) and classify patients as either high risk (H) or low risk (L) and further suppose that the groups turn out as follows:

A : HHLHLHHHHLLHHHLHHLHHH $(\mathrm{H}=15, \mathrm{L}=6)$
B : LLHHLHHLLHLHLHLHHLLHLL $(\mathrm{H}=10, \mathrm{L}=12)$

Note that there are 15 (71 per cent) high-risk patients and six (29 per cent) low-risk patients in treatment group A compared to a split of 10 (45 per cent) high-risk and 12 (55 per cent) low-risk patients in treatment group B.

Now suppose that the mean survival times are observed to be 21.5 months in group A and 27.8 months in group B. What conclusions can we draw? It is very difficult; the difference we have seen could be due to real treatment differences or could be caused by the imbalance in terms of differential risk across the groups, or a mixture of the two. Statisticians talk in terms of *confounding* (just a fancy way of saying *mixed up*) between the treatment effect and the effect of baseline risk. This situation is very difficult to unravel, and we avoid it by *stratified randomisation* to ensure that the *case mix* in the treatment groups is comparable.

This simply means that we produce separate randomisation lists for the high-risk and the low-risk patients, the strata in this case. For example, the following lists (which are block size four in each case):

H : ABBAAABBABABABABBBAAABBAABABBBAA
L : BAABBABAAABBBAABABABBBAABBAABAAB

will ensure firstly that we end up with balance in terms of treatment group sizes but also secondly that both the high- and low-risk patients will be equally split across those groups, that is, balance in terms of the mix of patients.

Having separate randomisation lists for the different centres in a multi-centre trial to ensure *equal* numbers of patients in the treatment groups within each centre is using *centre* as a stratification factor; this will ensure that we do not end up with treatment being confounded with centre.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'It is advisable to have a separate random scheme for each centre, i.e. to stratify by centre or to allocate several whole blocks to each centre. Stratification by important

prognostic factors measured at baseline (e.g. severity of disease, age, sex, etc.) may sometimes be valuable in order to promote balanced allocation within strata'.

Where the requirement is to have balance in terms of several factors, a stratified randomisation scheme using all combinations of these factors to define the strata would ensure balance. For example, if balance is required for sex and age, then a scheme with four strata – males, <50 years; females, <50 years; males, ≥50 years; and females, ≥50 years – will achieve the required balance.

1.4.5 Central randomisation

In *central randomisation*, the randomisation process is controlled and managed from a centralised point of contact. Each investigator makes a telephone call through an *Interactive Voice Response System* (IVRS) or an *Interactive Web Response System* (IWRS) to this centralised point when they have identified a patient to be entered into the study and is given the next allocation, taken from the appropriate randomisation list. Blind can be preserved by simply specifying the number of the (pre-numbered) pack to be used to treat the particular patient; the computerised system keeps a record of which packs have been used already and which packs contain which treatment. Central randomisation has a number of practical advantages:

- It can provide a check that the patient about to be entered satisfies certain inclusion/exclusion criteria, thus reducing the number of protocol violations.
- It provides up-to-date information on all aspects of recruitment.
- It allows more efficient distribution and stock control of medication.
- It provides some protection against biased allocation of patients to treatment groups in trials where the investigator is not blind; the investigator knowing the next allocation could (perhaps subconsciously) select patients to include or not include based on that knowledge; with central randomisation, the patient is identified and information given to the system before the next allocation is revealed to them.
- It gives an effective way of managing multi-centre trials.
- It allows the implementation of more complex allocation schemes such as minimisation and dynamic allocation (but see comments later on these techniques).

Earlier, we discussed the use of stratified randomisation in multi-centre trials, and where the centres are large, this is appropriate. With small centres however, for example, in GP trials, this does not make sense and a stratified randomisation with *region* defining the strata may be more appropriate. Central randomisation would be essential to manage such a scheme.

Stratified randomisation with more than a small number of strata would be difficult to manage at the site level, and the use of central randomisation is then almost mandatory.

1.4.6 Dynamic allocation and minimisation

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Dynamic allocation is an alternative procedure in which the allocation of treatment to a subject is influenced by the current balance of allocated treatments and, in a stratified trial, by the stratum to which the subject belongs and the balance within that stratum. Deterministic dynamic allocation procedures should be avoided and an appropriate element of randomisation should be incorporated for each treatment allocation'.

Dynamic allocation moves away from having a pre-specified randomisation list, and the allocation of patients evolves as the trial proceeds. The method looks at the current balance, in terms of the mix of patients and a number of pre-specified factors, and allocates the next patient in an optimum way to help redress any imbalances that exist at that time.

For example, suppose we require balance in terms of sex and age (≥65 vs. <65) and part way through the trial we see a mix of patients as in Table 1.1.

Treatment group A contains proportionately more males (12 out of 25 vs. 10 out of 25) than treatment group B but fewer patients over 65 years (7 out of 25 vs. 8 out of 25). Further suppose that the next patient to enter is male and aged 68 years. In terms of sex, we would prefer that this patient be placed in treatment group B, while for age, we would prefer this patient to enter in group A. The greater imbalance however is in relation to sex, so our overall preference would be for treatment group B to help *correct* for the current imbalance. The method of *minimisation* would simply put this patient in group B. ICH E9 however recommends that we have a *random element* to that allocation, and so, for example, we would allocate this patient to treatment group A with, say, probability of 0.7. Minimisation is a special case of dynamic allocation where the random assignment probability (0.7 in the example) is equal to one. Of course with a small number of baseline factors, for example, centre and two others, stratified randomisation will give good enough balance, and there is no need to consider the more complex dynamic allocation. This technique, however, has been proposed when there are more factors involved.

Since the publication of ICH E9 however, there has been considerable debate about the validity of dynamic allocation, even with the random element. There

Table 1.1 Current mix of patients

	A	B
Total	25	25
Male	12/25	10/25
Age ≥65	7/25	8/25

is a school of thought that has some sympathy within regulatory circles that supports the view that the properties of standard statistical methodologies, notably *p*-values and confidence intervals, are not strictly valid when such allocation schemes are used. As a result, regulators are very cautious.

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'…techniques of dynamic allocation such as minimisation are sometimes used to achieve balance across several factors simultaneously. Even if deterministic schemes are avoided, such methods remain highly controversial. Thus applicants are strongly advised to avoid such methods'.

So if you are planning a trial, then stick with stratification and avoid dynamic allocation. If you do have an ongoing trial that is using dynamic allocation, then be prepared at the statistical analysis stage to supplement the standard methods of calculating *p*-values with more complex methods that take account of the dynamic allocation scheme. These methods go under the name of *randomisation tests*.

See Roes (2004) for a comprehensive discussion of dynamic allocation.

1.4.7 Cluster randomisation

In some cases, it can be more convenient or appropriate not to randomise individual patients, but to randomise groups of patients. The groups, for example, could correspond to GPs so that each GP enters, say, four patients, and it is the 100 GPs that are randomised, 50 giving treatment A and 50 giving treatment B. Such methods are used but are more suited to phase IV than the earlier phases of clinical development. Many health interventions in third world countries are frequently evaluated using cluster randomisation.

Bland (2004) provides a review and some examples of cluster randomised trials, while Campbell, Donner and Klar (2007) give a comprehensive review of the methodology.

1.5 Bias and precision

When we are evaluating and comparing our treatments, we are looking for two things:

- An unbiased, correct view of how effective (or safe) the treatment is
- An accurate estimate of how effective (or safe) the treatment is

As statisticians, we talk in terms of *bias* and *precision*; we want to eliminate bias and to have high precision. Imagine having 10 attempts at hitting the bull's-eye on a target board as shown in Figure 1.1. Bias is about hitting the bull's-eye on average; precision is about being consistent.

These aspects are clearly set out in ICH E9.

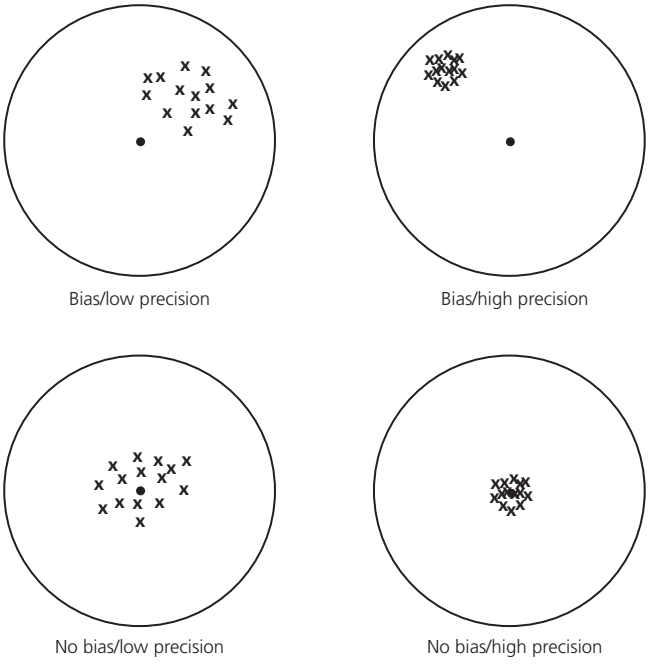

Figure 1.1 Bias and precision

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Many of the principles delineated in this guidance deal with minimising bias and maximising precision. As used in this guidance, the term "bias" describes the systematic tendency of any factors associated with the design, conduct, analysis and interpretation of the results of clinical trials to make the estimate of a treatment effect deviate from its true value'.

What particular features in the design of a trial help to eliminate bias?

- Concurrent control group as the basis for a *treatment comparison*
- Randomisation to avoid bias in allocating subjects to treatments
- Blinding of both the subject and the investigator
- Pre-specification of the methods of statistical analysis

What particular features in the design of a trial help to increase precision?

- Large sample size
- Measuring the endpoints in a precise way
- Standardising aspects of the protocol that impact on patient-to-patient variation
- Collecting data on key prognostic factors and including those baseline factors as covariates in the statistical analysis
- Choosing a homogeneous group of patients

- Choosing the most appropriate design (e.g. using a crossover design rather than a parallel-group design where this is appropriate)

Several of the issues raised here may be unclear at this point; simply be aware that eliminating bias and increasing precision are the key issues that drive our statistical thinking from a design perspective. Also, be aware that if something should be sacrificed, then it is precision rather than bias. High precision in the presence of bias is of no value; you are simply then getting a more precise wrong answer! First and foremost, we require an unbiased view; increasing precision is then a bonus. Similar considerations are also needed when we choose the appropriate statistical methodology at the analysis stage.

One particular point to make clear, and this is a common misunderstanding, is that having a large sample size of itself does not remove bias. If there is a flaw in the trial design, or in the planned methods of statistical analysis, that causes bias, then beating the trial over the head with large patient numbers will not eliminate that bias and you will still be misled, regarding the true treatment difference perhaps even more so because the trial is large. As mentioned earlier, having a large sample size in a flawed clinical trial will just result in a more precise, incorrect answer!

1.6 Between- and within-patient designs

The simplest trial design of course is the *parallel-group design* assigning patients to receive either treatment A or treatment B. For example, suppose we have a randomised parallel-group design in hypertension with 50 patients per group and that the mean fall in diastolic blood pressure in each of the two groups is as follows:

A: $\bar{x}_1 = 4.6$ mmHg

B: $\bar{x}_2 = 7.1$ mmHg

One thing to note that will aid our discussion later is it would be easy (but incorrect) to conclude in light of the data that B is a more effective treatment than A because we have seen a greater fall on average with treatment B than with treatment A, but is that necessarily the case? One thing we have to remember is that the 50 patients in group A are a different group of patients from the 50 patients in group B and patients respond differently, so, the observed difference between the treatments could simply be caused by patient-to-patient variation. As we will see later, unravelling whether the observed difference is reflective of a real treatment difference or simply a chance difference caused by patient-to-patient variation with identical treatments is precisely the role of the *p*-value; but it is not easy.

This design is what we refer to as a *between-patient design*. The basis of the treatment comparison is the comparison between two independent groups of patients.

An alternative design is the *within-patient design*. Such designs are not universally applicable but can be very powerful under certain circumstances. One form of the within-patient design is the *paired design*:

- In ophthalmology – treatment A in the right eye and treatment B in the left eye.
- In a volunteer study in wound care – *create* a wound on each forearm and use dressing of type A on the right forearm and dressing of type B on the left forearm.

Here, the 50 subjects receiving A will be the same 50 subjects who receive B, and the comparison of A and B in terms of, say, mean healing time in the second example is a comparison based on identical *groups* of subjects. At least in principle, drawing conclusions regarding the relative effect of the two treatments and accounting for the patient-to-patient (or subject-to-subject) variation may be easier under these circumstances.

Another example of the within-patient design is the *crossover design*. Again, each subject receives each of the treatments but now sequentially in time with some subjects receiving the treatments in the order A followed by B and some in the order B followed by A.

In both the paired design and the crossover design, there is, of course, randomisation; in the second paired design example earlier, it is according to which forearm receives A and which forearm receives B, and randomisation is to treatment order, A/B or B/A, in the crossover design.

1.7 Crossover trials

The crossover trial was mentioned in the previous section as one example of a within-patient design. In order to discuss some issues associated with these designs, we will consider the simplest form of crossover trial – two treatments A and B and two treatment periods I and II.

The main problem with the use of this design is the possible presence of the so-called carry-over effect. This is the residual effect of one of the treatments in period I influencing the outcome on the other treatment in period II. An extreme example of this would be the situation where one of the treatments, say, A, was very efficacious, so much so that many of the patients receiving treatment A were cured of their disease, while B was ineffective and had no impact on the underlying disease. As a consequence, many of the subjects following the A/B sequence would give a good response at the end of period I (an outcome ascribed to A) but would also give a good response at the end of period II (an outcome ascribed to B) because they were cured by A. These data would give a false impression of the A versus B difference. In this situation, the B data obtained from period II is contaminated and the data coming out of such a trial are virtually useless.

It is important therefore to only use these designs when you can be sure that carry-over effects will not be seen. Introducing a washout period between period I and period II can help to eliminate carry-over so that when the subject enters period II, their disease condition is similar to what it was at the start of period I. Crossover designs should not be used where there is the potential to affect the underlying disease state. ICH E9 is very clear on the use of these designs.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Crossover designs have a number of problems that can invalidate their results. The chief difficulty concerns carryover, that is, the residual influence of treatments in subsequent treatment periods… When the crossover design is used it is therefore important to avoid carryover. This is best done by selective and careful use of the design on the basis of adequate knowledge of both the disease area and the new medication. The disease under study should be chronic and stable. The relevant effects of the medication should develop fully within the treatment period. The washout periods should be sufficiently long for complete reversibility of drug effect. The fact that these conditions are likely to be met should be established in advance of the trial by means of prior information and data'.

The crossover design is used extensively in phase I trials in healthy volunteers to compare different formulations in terms of their bioequivalence (where there is no underlying disease to affect). They can also be considered in diseases, for example, asthma, where the treatments are being used simply to relieve symptoms; once the treatments are removed, the symptoms return to their earlier level.

1.8 Signal, noise and evidence

1.8.1 Signal

Consider the example in Section 1.6 comparing treatments A and B in a parallel-group trial. The purpose of this investigation is to detect differences in the mean reductions in diastolic blood pressure between the two groups. The observed difference between $\bar{x}_1 = 4.6$ mmHg and $\bar{x}_2 = 7.1$ mmHg is 2.5 mmHg. We will refer to this difference as the *signal*, and this captures in part the evidence that the treatments truly are different. Clearly, if the observed difference was larger, we would likely be more inclined to conclude differences. Large differences give strong signals, while small differences give weak signals.

1.8.2 Noise

The signal, however, is not the only aspect of the data that plays a part in our conclusions. If we were to see a large amount of patient-to-patient variation, then we would be less inclined to conclude differences than if all the patients

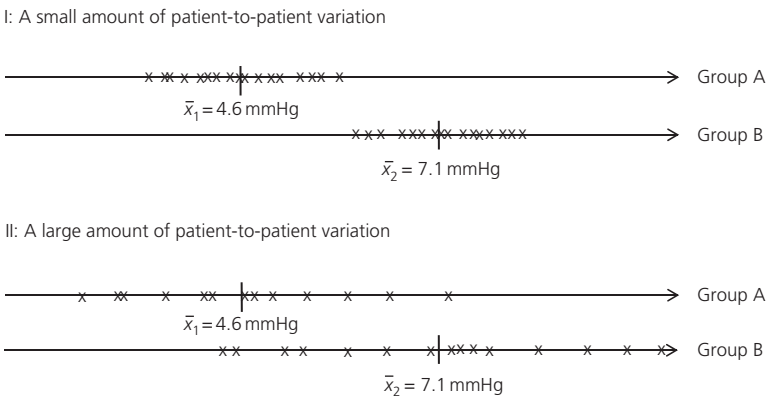

Figure 1.2 Differing degrees of patient-to-patient variation

in treatment group A had reductions tightly clustered around 4.6 mmHg, while those in treatment group B had values tightly clustered around 7.1 mmHg. As can be seen in Figure 1.2, the evidence for a real treatment difference in situation I is much stronger than the evidence seen in situation II although the mean values for both groups are actually the same in each case. We refer to the patient-to-patient variation as the *noise*, and clearly the extent of the noise will influence our willingness to declare differences between the treatments. An observed difference of 2.5 mmHg based on a small amount of noise is much stronger evidence for a true treatment difference than an observed difference of 2.5 mmHg in the presence of a large amount of noise.

The sample size plays an additional role in our willingness to conclude treatment differences and in a sense serves to compensate for the extent of the noise. If there is a large amount of patient-to-patient variation (a large amount of noise), then a large sample size is needed before we are able to *see what is happening on average* and conclude that the true means are indeed separated. In contrast, with a small amount of patient-to-patient variation, it is somewhat easier to recognise that the means truly are different, even with a small sample size.

1.8.3 Signal-to-noise ratio

These concepts of signal and noise provide a way of thinking for statistical experiments. In declaring differences, we look for strong signals and small amounts of noise, that is, a large *signal-to-noise ratio*. If either the signal is weak or the noise is large or both, then this ratio will be small and we will have little evidence on which to *declare* differences. The sample size can then be added into this mix. The value of a signal-to-noise ratio based on a small sample size is less reliable than the value of a signal-to-noise ratio based on a large sample

size, and clearly this is also going to influence our willingness to declare treatment differences.

In one sense, the signal is out of our control; it will depend entirely on what the true treatment difference is. Similarly, there is little we can do about the patient-to-patient variability, although we can reduce this by having, for example, precise measures of outcome or a more homogeneous group of patients. The sample size however is very much under our control, and common sense tells us that increasing this will provide a more reliable comparison and make it easier for us to detect treatment differences when they exist.

Later, in Chapter 8, we will discuss power and sample size and see how to choose sample size in order to meet our objectives. We will also see in Section 3.3 how, in many circumstances, the calculation of the p-value is based on the signal-to-noise ratio, which when combined with the sample size allows us to numerically calculate the *evidence* in favour of treatment differences. We will see in that section, for example, that when comparing two treatment means with n subjects per group, the *evidence* for treatment differences is captured by the square root of $n/2$ multiplied by the signal-to-noise ratio.

1.9 Confirmatory and exploratory trials

ICH E9 makes a very clear distinction between *confirmatory* and *exploratory* trials. From a statistical perspective, this is an important distinction as certain aspects of the design and analysis of data depend upon this confirmatory/exploratory distinction.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'A confirmatory trial is an adequately controlled trial in which the hypotheses are stated in advance and evaluated. As a rule, confirmatory trials are needed to provide firm evidence of efficacy or safety'.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The rationale and design of confirmatory trials nearly always rests on earlier clinical work carried out in a series of exploratory studies. Like all clinical trials, these exploratory studies should have clear and precise objectives. However, in contrast to confirmatory trials, their objectives may not always lead to simple tests of pre-defined hypotheses'.

Typically, later phase trials tend to contain the confirmatory elements, while the earlier phase studies – proof of concept, dose finding, etc. – are viewed as exploratory. Indeed, an alternative word for confirmatory is pivotal. It is the

confirmatory elements of our trials that provide the pivotal information from a regulatory perspective.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Any individual trial may have both confirmatory and exploratory aspects'.

Usually, it is the primary and secondary endpoints that provide the basis of the confirmatory claims. Additional endpoints may then provide the basis for exploratory investigations.

1.10 Superiority, equivalence and non-inferiority trials

A clear distinction needs to be made between superiority, equivalence and non-inferiority trials.

In a *superiority* trial, our objective is to demonstrate either that our treatment works by demonstrating superiority over placebo or that we are superior to some reference or standard treatment.

In an *equivalence* trial, we are looking to show that we are similar to some reference treatment; bioequivalence trials are the most common examples of this type of trial.

Finally, in a *non-inferiority* trial, we are trying to demonstrate that we are no more than a certain, pre-specified, usually small amount worse than (*at least as good as*) some active reference treatment.

In therapeutic equivalence trials and in non-inferiority trials, we are often looking to demonstrate efficacy of our test treatment indirectly. It may be that for ethical or practical reasons, it is not feasible to show efficacy by undertaking a superiority trial against placebo. In such a case, we compare our test treatment to a control treatment that is known to be efficacious and demonstrate either strict equivalence or *at least as good as* (non-inferiority). If we are successful, then we can be confident that our test treatment works.

Alternatively, there may be commercial reasons why we want to demonstrate the non-inferiority of our treatment against an active control. Maybe our treatment potentially has fewer side effects than the active control, and we are prepared to pay a small price for this safety advantage in relation to efficacy. If this were the case, then of course we would need to show advantages in terms of a reduction in side effects, but we would also need to demonstrate that we do not lose much with regard to efficacy.

Non-inferiority trials are becoming more and more common as time goes on. This in part is due to the constraints imposed by the revised Helsinki Declaration (2004) and the increasing concern in some circles regarding the ethics of placebo use. These trials however require very careful design and conduct, and we will discuss this whole area in Chapter 12.

1.11 Data and endpoint types

It is useful to classify the types of data and endpoints that we see in our clinical investigations.

The most common kind of data that we see is *continuous* data. Examples include cholesterol level, exercise duration, blood pressure, FEV_1 and so on. Each of these quantities is based on a continuum of potential values. In some cases, of course, our measurement technique may only enable us to record to the nearest whole number (e.g. blood pressure), but that does not alter the basic fact that the underlying scale is continuous.

Probably, the second most common data type is *binary*. Examples of binary data include cured/not cured, responder/non-responder and died/survived. Here, the measure is based on a dichotomy.

Moving up from binary is *categorical* data where there are more than two categories that form the basis of the *measurement*. The following are examples of categorical variables:

- Death from cancer causes/death from cardiovascular causes/death from respiratory causes/death from other causes/survival
- Pain: none/mild/moderate/severe/very severe

The categories are non-overlapping and each patient is placed into one and only one of the outcome categories. Binary data is a special case where the number of categories is just two.

These two examples however are different; in the first example, the categories are unordered, while in the second example, there is a complete ordering across the defined categories. In the latter case, we term the data/endpoint type either *ordered categorical* or *ordinal*.

Ordinal data arises in many situations. In oncology (solid tumours), the RECIST criteria record outcome in one of four response categories (National Cancer Institute, www.cancer.gov):

- Complete response (CR) = disappearance of all target lesions
- Partial response (PR) = 30 per cent decrease in the sum of the longest diameter of target lesions
- Progressive disease (PD) = 20 per cent increase in the sum of the longest diameter of target lesions
- Stable disease (SD) = small changes that do not meet the aforementioned criteria

When analysing data, it is important of course that we clearly specify the appropriate order, and in this case, it is CR, PR, SD and PD.

Other data arise as *scores*. These are frequently as a result of the need to provide a measure of some clinical condition such as depression or anxiety. The Hamilton Depression (HAMD) Scale and the Hamilton Anxiety (HAMA) Scale provide measures in these cases. These scales contain distinct items that are scored individually, and then the total score is obtained as the sum of the

Table 1.2 Categorisation

Group	Cigarettes per day
1	0
2	1–5
3	6–20
4	>20

individual scores. For the HAMD Scale, there are usually 17 items – depressed mood, self-depreciation, guilt feelings, etc. – each scored on a three-point to five-point scale. The five-point scales are typically scored 0 = absent, 1 = doubtful to mild, 2 = mild to moderate, 3 = moderate to severe and 4 = very severe, while the three-point scales are typically 0 = absent, 1 = probable or mild and 2 = definite.

Finally, data can arise as *counts* of items or events; number of epileptic seizures in a 12-month period, number of asthma attacks in a 3-month period and number of lesions are just a few examples.

As we shall see later, the endpoint type to a large extent determines the class of statistical tests that we undertake. Commonly for continuous data, we use the t-tests and their extensions – analysis of variance and analysis of covariance. For binary, categorical and ordinal data, we use the class of chi-square tests (Pearson chi-square for categorical data and the Mantel–Haenszel chi-square for ordinal data) and their extension, logistic regression.

Note also that we can move between data types depending on the circumstances. In hypertension, we might be interested in:

- The fall in diastolic blood pressure (continuous)
- Success/failure with success defined as a reduction of at least 10 mmHg in diastolic blood pressure and diastolic below 90 mmHg (binary)
- Complete success/partial success/failure with complete success = reduction of at least 10 mmHg and diastolic below 90 mmHg, partial success =reduction of at least 10 mmHg but diastolic 90 mmHg or above and failure = everything else (ordinal)

There are further links across the data types. For example, from time to time, we group continuous, score or count data into ordered categories and analyse using techniques for ordinal data. For example, in a smoking cessation study, we may reduce the basic data on cigarette consumption to just four groups (Table 1.2), accepting that there is little reliable information beyond that.

We will continue this discussion in the next section on endpoints.

1.12 Choice of endpoint

1.12.1 Primary variables

Choosing a single primary endpoint is part of a strategy to reduce multiplicity in statistical testing. We will leave discussion of the problems arising with

multiplicity until Chapter 10 and focus here on the nature of endpoints both from a statistical and a clinical point of view.

Generally, the primary endpoint should be that endpoint that is the clinically most relevant endpoint from the patients' perspective.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The primary variable ("target" variable, primary endpoint) should be that variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial'.

This choice should allow, among other things, a clear quantitative measure of benefit at the individual patient level. As we will see, identifying new treatments is not just about statistical significance, but it is also about clinical importance, and the importance of the clinical finding can only ever be evaluated if we can quantify the clinical benefit for patients.

Usually, the primary variable will relate to efficacy, but not always. If the primary objective of the trial concerns safety or quality of life, then a primary variable(s) relating to these issues would be needed.

The primary endpoint should not be confused with a summary measure of the benefit. For example, the primary endpoint may be a binary endpoint, survival beyond two years/death within two years, while the primary evaluation is based upon a comparison of two-year survival rates between two treatments. The primary endpoint *is not* the proportion surviving two years, but it is the binary outcome survival beyond two years/death within two years, the variable measured at the patient level.

The primary endpoint must be pre-specified in a confirmatory trial as specification after unblinding could clearly lead to bias. Generally, there would be only one primary endpoint, but in some circumstances, more than one primary endpoint may be needed in order to study the different effects of a new treatment. For example, in acute stroke, it is generally accepted that two primary endpoints are used – one relating to survival free of disability and a second relating to improvement in neurological outcome. See CPMP (2001) 'Note for Guidance on Clinical Investigation of Medicinal Products for the Treatment of Acute Stroke' for further details on this.

1.12.2 Secondary variables

Secondary variables may be defined that support a more detailed evaluation of the primary endpoint(s), or alternatively, such endpoints may relate to secondary objectives. These variables may not be critical to a claim but may help in understanding the nature of the way the treatment works. In addition, data on secondary endpoints may help to embellish a marketing position for the new treatment.

If the primary endpoint gives a negative result, then the secondary endpoints cannot generally recover a claim. If, however, the primary endpoint has

given a positive result, then additional claims can be based on the secondary endpoints provided these have been structured correctly within the confirmatory strategy. In Chapter 10, we will discuss hierarchical testing as a basis for such a strategy.

1.12.3 Surrogate variables

Surrogate endpoints are usually used when it is not possible within the timeframe of the trial to measure true clinical benefit. Many examples exist as seen in Table 1.3.

Unfortunately, many treatments that have shown promise in terms of surrogate endpoints have been shown not to provide subsequent improvement in terms of the clinical outcome. Fleming and DeMets (1996) provide a number of examples where we have been disappointed by surrogate endpoints and provide in each of these cases possible explanations for this failure of the surrogate. One common issue in particular is that a treatment may have an effect on a surrogate through a particular pathway that is unrelated to the underlying disease process or the clinical outcome.

Treatment effects on surrogate endpoints therefore do not necessarily translate into treatment effects on clinical endpoints, and the validity of the surrogate depends not only on the variable itself but also on the disease area and the mode of action of the treatment. Establishing new valid surrogates is very difficult. Fleming and DeMets conclude that surrogates are extremely valuable in phase II *proof-of-concept* studies, but they question their general use in phase III confirmatory trials.

Table 1.3 Surrogate variable and clinical endpoints

Disease	Surrogate variable	Clinical endpoint
Congestive heart failure	Exercise tolerance	Mortality
Osteoporosis	Bone mineral density	Fractures
HIV	CD4 cell count	Mortality
Hypercholesterolemia	Cholesterol level	Coronary heart disease

Example 1.2 Bone mineral density and fracture risk in osteoporosis

Li, Chines and Meredith (2004) quote three clinical trials evaluating the effectiveness of alendronate, risedronate and raloxifene in increasing bone mineral density (BMD) and reducing fracture risk in osteoporosis. These treatments are seen to reduce fracture risk by similar amounts (47 per cent, 49 per cent and 46 per cent, respectively), yet their effects on increasing BMD are somewhat different (6.2 per cent, 5.8 per cent and 2.7 per cent, respectively). Drawing conclusions on the relative effectiveness of these treatments based solely in terms of the surrogate BMD would clearly be misleading.

1.12.4 Global assessment variables

Global assessment variables involve an investigator's overall impression of improvement or benefit. Usually, this is done in terms of an ordinal scale of categories. While the guidelines allow such variables, experience shows that they must at the very least be accompanied by objective measures of benefit. Indeed, both the FDA and the European regulators tend to prefer the use of the objective measures only, certainly at the primary endpoint level.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'If objective variables are considered by the investigator when making a global assessment, then those objective variables should be considered as additional primary, or at least important secondary, variables'.

1.12.5 Composite variables

In some circumstances, it may be necessary to combine several events/endpoints to produce a combined or composite endpoint. The main purpose for doing so is to avoid multiple testing, and more will be said about this in Chapter 10. In addition, combining endpoints/events will increase the absolute numbers of events observed, and this can increase sensitivity (power) for the detection of treatment effects.

1.12.6 Categorisation

In general, a variable measured on a continuous scale contains more information and is a better reflection of the effect of treatment than a categorisation of such a scale. For example, in hypertension, the clinical goal may be to reduce diastolic blood to below 90 mmHg; that is not to say that a reduction down to 91 mmHg is totally unacceptable, while a reduction down to 89 mmHg is a perfect outcome. Having a binary outcome that relates to achieving 90 mmHg is clearly only a somewhat crude measure of treatment benefit. The CHMP recognise that the original variable contains more information, and although they support the presentation of the proportion of responders in order to gauge clinical benefit, they suggest that statistical testing be undertaken on the original scale.

CPMP (2002) 'Points to Consider on Multiplicity Issues in Clinical Trials'

'When used in this manner, the test of the null hypothesis of no treatment effect is better carried out on the original primary variable than on the proportion of responders'.

Nonetheless, categorisation can be of benefit under some circumstances. In an earlier section, we discussed the categorisation of number of cigarettes to a four-point ordinal scale, accepting that measures on the original scale may be subject to substantial error and misreporting; the additional information contained in the number of cigarettes smoked is in a sense spurious precision.

There may also be circumstances where a categorisation combines responses measured on different measurement domains, for example, to give a single dichotomous responder/non-responder outcome. There are connections here with global assessment variables. This approach is taken in Alzheimer's disease where the effect of treatment is in part expressed in terms of the 'proportion of patients who achieve a meaningful benefit (response)'; see the CHMP (2007) 'Draft Guideline on Medicinal Products in the Treatment of Alzheimer's Disease and Other Dementias'. In oncology, the RECIST criteria may be used simply to give the proportion of patients who achieve a CR or PR. This reduces the sensitivity of the complete scale but may make it easier to quantify the clinical benefit in what is often termed a *responder analysis*. For an interesting exchange on the value of dichotomisation, see Senn (2003) and Lewis (2004). Royston, Altman and Sauerbrei (2006) are against categorisation for data analysis as this tends to waste information and consequently is less able to detect treatment differences should they exist. Both these sets of authors however recognise that such analyses can be beneficial in terms of data presentation and communication.

Finally, a few words about the use of the visual analogue scale (VAS). A value on this 10 mm line gives a continuous measure (the distance between the left-hand end and the marked value), and these are used successfully in a number of therapeutic settings. Their advantage over an ordinal four- or five-point scale, however, is questionable as again there is an argument that the additional *precision* provided by VAS is of no value. A study by Jensen *et al.* (1989) in the measurement of post-operative pain showed that information relating to pain was best captured using an 11-point scoring scale (0, 1, 2, …, 10) – sometimes referred to as a *Likert scale* – or a verbal rating scale with five points (mild, discomforting, distressing, horrible, excruciating). In addition, around 10 per cent of the patients were unable to understand the requirement for completion of the VAS for pain. These ordered categorical scales may well be as precise or more precise than the VAS and at the same time prove to be more effective because patients understand them better.