3

Research questions and research hypotheses

3.1 Introduction

One of the most efficient ways to acquire new knowledge about any topic is to devise questions that guide our investigations. These questions can focus our thinking and keep us on track as we gather information that contributes to our knowledge base. Devising more specific questions as we progress is typically a good strategy. While relatively loosely formed questions can be very helpful early on in the knowledge acquisition process, refining our knowledge is facilitated by asking more specific questions, and, in turn, acquiring more specific information and knowledge.

The scientific method is one particular method of acquiring new knowledge. In scientific research, including drug development, our questions need to be asked in a particular manner. These questions are called research questions, and they lead to the development of research hypotheses. The scientific method requires that these research hypotheses be structured in a certain way, and then tested in a scientific manner. As noted earlier, a fundamental characteristic of these research hypotheses is that they can be disproved. The word "disproved" is not a typo: These hypotheses need to be able to be disproved, not "proved."

3.2 The concept of scientific research questions

In our operational definition, Statistics is regarded as a multifaceted scientific discipline that comprises several activities. The first of these is identifying a research question that needs to be answered. We noted in Chapter 2 that the reason for the development of a new drug is usually the identification of an unmet medical need. The development of a drug that will meet this need requires a series of nonclinical studies that comprise the nonclinical development program, followed by a series of clinical studies that comprise the clinical program. In each case, the study will be designed to answer a specific research question or questions. (From a statistical perspective, it is a good idea to have a small number of research questions that will be answered by an individual study, despite the real temptation to try to collect data that will answer many research questions.)

3.3 Useful research questions

In Section 2.1 we cited Norgrady and Weaver's (2005) definition of a useful drug in the context of drug development. That definition is a good illustration of how the term "useful" is used in scientific research. A precise, operational definition is needed, rather than a vague statement such as "it looks pretty useful to me."

Turner (2007) provided an operational definition of a useful research question:

- It needs to be specific (precise).
- It needs to be testable.

If a candidate drug successfully makes it through the nonclinical development program, the safety and efficacy of the investigational drug will be tested in humans in a series of clinical trials that comprise the clinical development program. Each study that is conducted will test a particular aspect or facet of the drug, and the overall development program will employ these trials in a systematic fashion. We noted in Section 2.6 that three kinds of trials are typically conducted before a new drug is approved for marketing: human pharmacology trials, therapeutic exploratory trials, and therapeutic confirmatory trials. The information that is collected during human pharmacology trials forms the basis for the testing that is done in therapeutic exploratory trials, and the information from both categories of trials forms the basis for the testing that is done in therapeutic confirmatory trials – that is, a body of information and knowledge about the investigational drug is gained in an incremental manner, a hallmark of scientific investigation.

Information is gathered by conducting studies each of which asks a specific, testable research question. A general and vague question such as "Is this investigational drug good for people's blood pressure?" is simply not useful in this context, because it does not facilitate the acquisition of useful information. The same is true in nonclinical studies: the question "Do you think that the drug is pretty safe when given to a bunch of animals?" will not facilitate the acquisition of useful nonclinical information.

3.4 Useful information

You may have noticed that we used the term "useful information" twice in the previous paragraph. Accordingly, it is helpful to provide an operational definition of useful information in the context of drug development. Useful information has the following characteristics:

- It needs to be specific (this is the same as the first characteristic of a useful research question – see Section 3.3).
- It provides a solid basis for further studies that will acquire more useful information.
- It provides the rational basis for decision-making during the drug development process.
- It will be acceptable to regulatory agencies, which will eventually review the reports that provide the information to them.

3.5 Moving from the research question to the research hypotheses

Before discussing the connection between the research question and the associated research hypotheses, it should be noted that the word "hypotheses" here is not a typo: We meant to write the plural form "hypotheses" and not the singular form "hypothesis." As discussed shortly, each research question has two associated research hypotheses.

We noted in Section 3.3 that a research question needs to be useful. We also noted that the question "Is this investigational drug good for people's blood pressure?" is not useful, because it does not provide a precise definition of "good" – that is, the research question is not specific, or precise. A better research question might be "Does the investigational drug alter blood pressure?" At the outset of any scientific inquiry about the potential effects of an investigational hypertensive drug, we must entertain the notion that the drug may, contrary to our expectations, actually increase blood pressure. For reasons that we explain in Chapter 6, this potential is expressed in the statement of the statistical hypotheses. For now, this improved research question addresses the intention of our experiment and drug development program. However, we can do better.

3.6 The placebo effect

An interesting phenomenon in pharmacotherapy is called the "placebo effect." The dictionary currently sitting in our office (*The American Heritage College Dictionary*, 3rd edn) provides several definitions of the word placebo, including:

> A substance containing no medication and given to reinforce a patient's expectation to get well
>
> An inactive substance used as a control in an experiment to determine the effectiveness of a medicinal drug

Both of these definitions are helpful in the present context. The first is a more general one, and relates to the observation that, if a person is given a substance containing no medication, but that person believes that the substance will have a beneficial therapeutic effect, it is not unusual that improvement may be seen in the person's condition. The second definition is a particularly relevant definition in the context of this book, and the term "placebo" will be used extensively in later chapters.

3.7 The drug treatment group and the placebo treatment group

The terms "drug treatment group" and "placebo treatment group" will become very familiar to you as you work your way through this book. Chapter 4 provides more detail, but it is helpful at this point to comment briefly on the following aspect of certain preapproval clinical trials.

It is very common in therapeutic confirmatory trials (and in some therapeutic exploratory trials) to compare the effects of the investigational drug with the effects of a placebo (see also the discussions in Section 3.16.1) – that is, these trials are comparative in nature. One common study design involves giving the investigational drug to one group of individuals, the drug treatment group, and the placebo to a second group of individuals, the placebo treatment group. In addition, and extremely importantly, these individuals do not know whether they are receiving the investigational drug or the placebo. All individuals are treated identically throughout the trial, with the one exception of the treatment that they receive.

A key point to note is that individuals receiving the placebo treatment often show a small improvement in the condition that is the focus of the trial. In some instances, such as studies of antidepressants and analgesics, improvement is self-reported by the individuals themselves and can often be marked. In addition, improvement among individuals can be accounted for by a phenomenon called "regression to the mean." Imagine a clinical therapeutic trial involving an investigational antihypertensive drug where individuals are eligible for enrollment only if they have a documented systolic blood pressure (SBP) of a specified level (say at least 140 mmHg). It is a simple fact that, as a result of expected and naturally occurring random variation, some individuals may initially show this level of SBP even though it does not accurately reflect their true SBP. On subsequent observations the level of the characteristic will return closer to its expected level, resulting in an "improvement" caused simply by the fact that the individual was enrolled at a time when his or her SBP was higher than normal. In trials involving investigational antihypertensive drugs, it is not unusual for individuals in the placebo treatment group to show small decreases in BP during the trial. Therefore, it becomes important to determine whether the investigational drug has a larger effect on BP than the placebo – that is, the trial is comparative in nature and the placebo is the control against which the investigational drug is compared.

3.8 Characteristics of a useful research question

In Section 3.4 we formulated some initial versions of a research question. We decided that the question "Is this investigational drug good for people's blood pressure?" is not useful. We noted that an improved research question might be "Does the investigational drug alter blood pressure?" This is certainly moving in the right direction. However, as noted at the end of that section, we can do better.

A good clue as to how we can devise a better research question comes from the discussions that we have just had about the comparative nature of these trials, in which the effect of the investigational drug is compared with the effect of the placebo. Based on these discussions an improved research question can be phrased as

"Does the new drug alter SBP more than placebo?" This version of the research question has several useful characteristics:

- It involves both of the treatments received by individuals in the trial: The investigational drug and the placebo.
- It is comparative. The goal is to compare the effects of the two treatments.
- It is precise. There will be a precise answer – yes it does, or no it doesn't. We will have to define the term "more" in more detail shortly, but in the meantime please trust us that this can indeed be done in a precise manner by using the discipline of Statistics.

3.9 The reason why there are two research hypotheses

In this book all research questions are addressed and then answered via the construction of two research hypotheses, commonly called the null hypothesis and the alternate hypothesis. (Although another name for the alternate hypothesis, the research hypothesis, has its own appeal, we employ the commonly used term "alternate hypothesis" in this book.) Both of these hypotheses are key components of the procedure of hypothesis testing. This procedure is a statistical way of doing business. It is described and discussed in detail in Chapter 6, but it is beneficial to introduce the main concept here.

3.9.1 The null hypothesis

The null hypothesis is the crux of hypothesis testing. (It is important to note that the form of the null hypothesis varies in different statistical approaches. As the main type of clinical trial discussed in this book is the therapeutic confirmatory trial, we talk about this first. We then talk briefly about the forms of the null hypothesis that are used in other types of trials in Section 3.10.) As noted earlier, therapeutic confirmatory trials are comparative in nature. We want to evaluate the efficacy of the investigational drug, and the way that we do this is to compare its efficacy with the efficacy of a control treatment, typically a placebo. The key question, expressed in our research question, is "Does the new drug alter SBP more than placebo?" As noted earlier, we need to provide a precise definition of "more," and we will do this in due course. In this type of trial, called a superiority trial, the null hypothesis takes the following form:

> The average effect of the investigational drug on SBP is equivalent to the average effect of the placebo on SBP.

3.9.2 The alternate hypothesis

The alternate hypothesis reflects the alternate possible outcome of the trial, and therefore the alternate possible answer to the research question: The trial was conducted with the specific goal of providing an answer to the research question. In a superiority trial the alternate hypothesis takes the following form:

> The average effect of the investigational drug on SBP is *not* equivalent to the average effect of the placebo on SBP.

Note that the research question "Does the new drug alter SBP more than placebo?" allows for the fact that the investigational drug could actually *increase* SBP more than placebo, as does the alternate hypothesis, which includes the possibility that the drug could increase SBP. The reason for this is that the statistical information used to decide which hypothesis is more plausible must include the possibility, however remote we believe it to be, that the drug has the opposite effect of what we hope for. Another way of stating this is that exclusion of one side of the alternate hypothesis – that is, that the drug is worse than the placebo – is presumptive and contrary to the scientific process of collecting data to search for the true state of nature. It is certainly true that, if we find ourselves in a position to claim that the alternate hypothesis should be accepted because the drug did more harm than good (that is, increased SBP), we will have answered the research question. For now, we must accept that, despite its seeming incongruence with our preferred research question, the use of a two-sided alternate hypothesis is the norm in the regulated world of drug development.

3.9.3 Facts about the null and alternate hypotheses

It is important to note that, whatever the outcome of the trial, both of these hypotheses cannot be correct. It is also true that one of them will always be correct. Again, we operationally define the term "correct" in this context in due course, but the point to note here is that:

- It is never the case that neither hypothesis is correct.
- It is never the case that both hypotheses are correct.
- It is always the case that one of them is correct and the other is not correct.

The procedure of hypothesis testing allows us to determine which hypothesis is correct.

A helpful way of remembering which hypothesis is which – that is, which form the null hypothesis takes and which the alternate hypothesis takes – is to conceptualize that the alternate hypothesis states what you are "hoping" to find and the null hypothesis states what you are not hoping to find. It must be emphasized here, however, that, although helpful, this conceptualization skates on very thin scientific ice: As Turner (2007, p 101) noted:

> In strict scientific terms, hope has no place in experimental research. The goal is to discover the truth, whatever it may be, and one should not start out hoping to find one particular outcome. In the real world, this ideologically pure stance is not common for many reasons (financial reasons being not the least of them).

When a pharmaceutical or biotechnology company has spent many years and huge sums of money developing an investigational drug, and the drug has made it to the point where a therapeutic confirmatory trial is being conducted, the company hopes that the drug will indeed be more effective than placebo, and in due course be approved for marketing by a regulatory agency. One reason for this hope is that patients will (relatively) soon have the opportunity to receive a new drug that is therapeutically beneficial for them. Another reason, as noted in the previous quote, is that the drug will be approved and make money for the company. Drug companies are for-profit businesses, and this is not a negative judgmental comment. The only way that they can develop future drugs is to sell present drugs for a profit: The costs in pharmaceutical research and development (R&D) are enormous. In this pragmatic sense, the alternate hypothesis (at least one side of it) can meaningfully be conceptualized as stating the outcome that you are hoping for.

3.10 Other forms of the null and alternate hypotheses

The forms of the null and alternate hypotheses are dictated by the goal of the trial. In the therapeutic confirmatory trial discussed so far the goal of the trial is to demonstrate that the investigational drug shows greater efficacy than the control treatment – that is, we are hoping to demonstrate that the investigational drug shows superior efficacy, hence the name superiority trial. Following our earlier memory tip, you can conceptualize the alternate hypothesis as stating what you are hoping to find and the null hypothesis as stating what you are not hoping to find. As we have seen, this leads to the following forms of these hypotheses:

- Null hypothesis: The average effect of the investigational drug on SBP is equivalent to the average effect of the placebo on SBP.
- Alternate hypothesis: The average effect of the investigational drug on SBP is *not* equivalent to the average effect of the placebo on SBP.

When we are hoping to demonstrate something different, however, these hypotheses take different forms. Consider the example of a trial called an equivalence trial. Equivalence trials are conducted to demonstrate that an investigational drug has therapeutic equivalence compared with a control treatment. Equivalence trials are also comparative in nature, but in this case the control treatment is not a placebo but a marketed drug. Here, the control treatment is referred to as an active comparator drug. This active comparator drug is typically the drug that is currently the best, or perhaps the only, treatment available for the disease or condition of interest, and is referred to as the gold standard treatment. The intent in an equivalence trial is to provide compelling evidence that the efficacy of the investigational drug is "equivalent" to that of the active comparator drug. (The term "equivalent" requires a precise statistical definition, and this is provided in Chapter 12.)

Why are equivalence trials important? That is, why would we be interested in a new drug that is only as effective as an existing drug? This is a good question, but one that has an equally good answer. One reason would be that we believe (hope) that the investigational drug is equally effective and also has the considerable advantage that its safety profile is better. This would lead to the same efficacy with less likelihood of side-effects. If the side-effects of the current gold standard drug are particularly unpleasant, this would be a considerable advantage. Other advantages that may justify the use of equivalence trials include convenience of the dosing regimen, and the inability to use an inactive control for ethical reasons.

In an equivalence trial the research question is: Does the new drug demonstrate equivalent efficacy compared with the reference drug? The resultant accompanying research and alternate hypotheses take the following forms:

- Null hypothesis: The investigational drug does not show equivalent efficacy to the comparator drug (the reference drug).
- Alternate hypothesis: The investigational drug does show equivalent efficacy to the comparator drug (the reference drug).

Following the logic of our memory tip, you will see that the alternate hypothesis in this case, just like in the case of a superiority trial, expresses what we are hoping to find, while the null hypothesis states what we are hoping not to find. The actual natures of the null and alternate hypotheses in an equivalence trial are different from those in a superiority trial, but this sentiment is the same. This is equally true in the case of various other types of trials, including noninferiority trials.

Noninferiority trials are similar to equivalence trials, but require that the investigational drug be in the worst case only trivially worse than the reference to be considered noninferior. A precise statistical definition of "trivially worse" must be agreed upon before the start of the trial. For noninferiority trials the research question is: Does the new drug demonstrate efficacy that is not unacceptably worse (Fleming 2007) than the reference drug? The null and alternate hypotheses corresponding to this research question take the form of:

- Null hypothesis: The efficacy of the investigational drug is unacceptably worse than the efficacy of the comparator drug (the reference drug).
- Alternate hypothesis: The efficacy of the investigational drug is not unacceptably worse than the efficacy of the comparator drug (the reference drug).

As for equivalence trials, additional details about noninferiority designs are provided in Chapter 12.

3.11 Deciding between the null and alternate hypothesis

A research question of interest, then, leads to the null hypothesis and the alternate hypothesis. As noted in Section 3.9.3:

- It is never the case that neither hypothesis is correct.
- It is never the case that both hypotheses are correct.
- It is always the case that one of them is correct and the other is not correct.

(We also noted in that section that we would address the precise operational definition of correct in due course, and we have not forgotten this.) Therefore, statistical analysis of the data acquired in the trial enables us to decide between these two mutually exclusive hypotheses. An ongoing theme in this book is that many decisions have to be made in drug development, and the discipline of Statistics provides numerical representations of information that provide the rational basis for decision-making. At the end of every trial, a decision needs to be made: Which of these two mutually exclusive hypotheses is correct – the null hypothesis or the alternate hypothesis? For the rest of this chapter we talk about superiority trials, but the points made apply equally to equivalence trials and noninferiority trials.

3.12 An operational statistical definition of "more"

Imagine the following results from a superiority trial:

- The average decrease in SBP for the individuals in the drug treatment group was 3 mmHg – that is, on average, the investigational drug lowered SBP by 3 mmHg in this trial.
- The average decrease in SBP for the individuals in the placebo treatment group was 2 mmHg – that is, on average, the placebo lowered SBP by 2 mmHg in this trial.

In this superiority trial we are interested to find out whether the investigational drug is more effective than the comparator treatment, the placebo. As the number "3" is numerically greater than the number "2," the simple mathematical answer is clear: Yes, the investigational drug lowered BP more than the placebo. However, you may well feel that this simple mathematical answer, although true, does not capture the spirit of the findings from the trial. The investigational drug managed to lower blood pressure only 1 mmHg more than the placebo, a substance that has no pharmacotherapeutic capability.

The term "treatment effect" is an important one in drug development, and is defined as the difference between the average response to the investigational drug and the average response to the comparator being used in the trial. In this case of the development of a new antihypertensive drug it is defined as the difference between the average decrease in BP shown by the individuals in the drug treatment group and the average decrease in BP shown by the individuals in the placebo treatment group. The logic here is that, as the placebo resulted in a small decrease, even though it has no pharmacotherapeutic capability, the pharmacotherapeutic capability of the investigational drug should be regarded as the decrease in BP over and above that caused by the placebo. Another interpretation of the treatment effect is that it is the amount of change in SBP attributed to the drug over and above that which would have been observed had the drug not been given. Therefore, the treatment effect here is 1 mmHg. This is > 0, and so the investigational drug is mathematically more effective than the comparator treatment, but, as noted earlier, you may be left feeling that the term "more" is not quite appropriate.

3.12.1 A very important aside: The concept of clinical significance

We are about to introduce you to the concept of a statistically significant difference between the means of two sets of numbers. These numbers can be any kind of data, not just the clinical data that are the focus of this book. No matter how large or how small the difference found between the means of two sets of numbers, it is possible for that difference to be declared statistically significantly different after appropriate analysis of those data. In some circumstances and for some data, a difference of just one unit between the means of the two groups of data may indeed be found to be statistically significantly different, and in such a case the use of the term "more" would be statistically appropriate.

However, in the context of clinical data, another extremely important concept is clinical significance. The clinical significance of a treatment effect is a completely separate assessment from the treatment effect's statistical significance. This is a clinical judgment, not the result of a single numerical calculation. It is perfectly possible for a treatment effect to be found to be statistically significant after an appropriate statistical analysis and yet judged by clinicians to be not clinically significant. As you will see, both statistical significance and clinical significance must be addressed in drug development – an observation that highlights that statisticians and clinicians must work very closely together during this process. Later discussions address the topic of clinical significance in detail.

3.12.2 An operational statistical definition of the term "more"

The discipline of Statistics provides us with methodology that tells us whether or not the use of the word "more" is appropriate from a statistical point of view. The formulation of a scientifically meaningful research question and its two associated hypotheses, the null hypothesis and the alternate hypothesis, allows us to reach an answer in an objective manner by following a prescribed methodology. Moreover, the regulatory and clinical communities acknowledge this methodology. Therefore, for a given set of data that have been collected in a trial, statistical testing provides a precise answer that is couched in statistical terms and that has effectively been agreed upon as objective by all interested parties (see Turner, 2007). This leads us to the concept of a statistically significant difference.

3.13 The concept of statistically significant differences

The concept of statistical significance and its practical implementation are discussed in more detail in Chapter 6, but it is appropriate here to set the scene for those discussions. The words "significant," "significance," and "significantly" are used differently in Statistics than they are in everyday language. In the language of Statistics they have precise quantitative meanings. We focus here on the meaning of the term "significantly" in the discipline of Statistics. The discipline of Statistics facilitates a single, quantitative answer to questions concerning assessments of "more." For a given set of data collected in a superiority trial, the employment of the appropriate statistical analysis will reveal whether the treatment effect attained statistical significance – that is, it will reveal whether or not the investigational drug was statistically significantly more effective than the placebo. In this manner it provides a precise definition of the term "more." If there is a statistically significant difference between the average decrease in blood pressure in the drug treatment group and the placebo treatment group – that is, if there is a statistically significant treatment effect – the use of the term "more" is warranted.

3.14 Putting these thoughts into more precise language

The following are the research question and the two associated research hypotheses that we have formulated so far:

- Research question: Does the new drug alter SBP more than the placebo?
- Null hypothesis: The average effect of the investigational drug on SBP is equivalent to the average effect of the placebo on SBP.
- Alternate hypothesis: The average effect of the investigational drug on SBP is *not* equivalent to the average effect of the placebo on SBP.

The concept of statistical significance allows all three of these to be reframed as follows:

- Research question: Does the new drug alter SBP statistically significantly more than the placebo?
- Null hypothesis: The average effect of the investigational drug on SBP is *not* statistically significantly different from the average effect of the placebo on SBP.
- Alternate hypothesis: The average effect of the investigational drug on SBP is statistically significantly different from the average effect of the placebo on SBP.

Each of these is now expressed in a more precise manner. In addition, the null hypothesis now allows for the treatment groups to differ to a certain extent. In any trial involving any treatments, the group averages will almost certainly differ to some extent: The probability of a treatment effect of zero is extremely small.

3.15 Hypothesis testing

We have now formulated our research question, null hypothesis, and alternate hypothesis in precise statistical language. This facilitates the strategy of hypothesis testing. Hypothesis testing revolves around two actions after an appropriate statistical analysis: Rejecting the null hypothesis or failing to reject the null hypothesis. The language used to express these two actions is very important. After thinking about these two actions for a few minutes, you might think that these actions could be expressed as "accepting the alternate hypothesis" and "accepting the null hypothesis," respectively. In everyday thinking this might be thought very reasonable. However, in the discipline and the language of Statistics, the terminology of rejecting the null hypothesis or failing to reject the null hypothesis is deliberately employed. While data from two groups (for example, means) may suggest that the alternate is more plausible than the null hypothesis, the sample size of the study also has a direct bearing on our ability to reject the null hypothesis. If there is at least a small difference between groups in a study (which will almost certainly be the case), it is possible that a null hypothesis that is not rejected in a study of a certain size would be rejected if the study had been larger. The relationship between sample size and the probability of rejecting the null hypothesis or failing to reject the null hypothesis is explored in Chapter 12.

The statistical convention of using the expression "failing to reject the null hypothesis" reflects the position that null hypotheses of no difference can always be rejected if enough observations are studied. Statistical methodology necessitates making a choice here. One of these two actions, rejecting or failing to reject the null hypothesis, has to be taken at the end of all hypothesis testing. The action taken is precisely determined by the result obtained from the statistical technique used to analyze the data.

3.16 The relationship between hypothesis testing and ethics in clinical trials

The issue of ethical considerations in clinical trials was introduced in Section 2.8. The procedure of hypothesis testing illustrates one important ethical consideration, clinical equipoise, particularly well. The fact that we have two research hypotheses that express two opposing possible occurrences makes it clear that we do not know which best represents the actual state of affairs. The bottom line is that, at the point in time when the study is planned and started, we do not know whether or not the investigational drug will be more effective than placebo. This uncertainty is a necessary prerequisite of conducting a trial: If it were known that the investigational drug were more effective it would be unethical for people with the disease or condition of interest to be given a placebo.

The philosophy that makes it acceptable that some individuals receive a placebo in a clinical trial is that a comparative clinical trial in which some receive a placebo while others receive the investigational drug is the best way to find out if that drug is indeed effective. If it is, the individuals who received the placebo would not themselves have benefited on this occasion, but their participation in the trial was a crucial component contributing to the later treatment of patients with the approved drug. As noted earlier individuals take part in clinical trials for the greater good, not for their own immediate benefit.

Uncertainty is therefore a fundamental prerequisite to conducting a therapeutic exploratory or therapeutic confirmatory trial. The results of the trial will be used in a decision-making process at the end of the trial: In the light of the prevailing uncertainty, we need to decide whether the results of the trial have provided compelling evidence that the investigational drug is indeed effective. If the statistical results show that the drug is indeed effective – that is, the treatment is statistically significant – the next step is for clinicians to decide if the treatment is also clinically significant (as noted in Section 3.12.1 we discuss clinical significance later in the book). If the treatment effect is deemed to be both statistically and clinically significant, the study team is likely to decide to move forward to the next study in their clinical development program.

3.16.1 Ethics and the use of placebo controls

At this point, it is appropriate to point out diverging views on the use of placebo control treatment in any clinical trial. Some authors have expressed the view that the comparator should always be an active control if possible – that is, in cases where there is already at least one drug on the market that has been demonstrated to be effective for treating the disease or condition of concern, the comparator should be one of these drugs. This issue is reviewed very effectively by Temple and Ellenberg (2000), and we strongly recommend that you read their paper. We agree with the arguments that they present supporting the use of placebo controls in appropriate circumstances. In addition, ICH Guidance E12A (2000) comments on this issue specifically in the context of the evaluation of efficacy in clinical trials for an investigational antihypertensive drug. It states that, for several important reasons, short-term (defined as 4–12 weeks in duration), blinded, placebo-controlled studies are "essential." It also states that long-term studies (defined as 6 months or more) should also be conducted to demonstrate maintenance of efficacy and to assess long-term safety, and that these trials would typically use an active control.

In this book we have chosen to focus on teaching the computational aspects of statistical analyses by using examples involving designs in which the investigational drug is compared with a placebo in short-term trials. In Section 4.7, we describe two potential measurement scenarios that are possible during a 12-week trial: Taking measurements at baseline and at the end of every 2-week period, and simply taking measurements at baseline and at the end of week 12, the end-of-treatment measure. In this book we use the latter design for the sake of simplicity. We have therefore chosen this form of example because it is ethical as discussed in ICH Guidance E12A, and, as noted in the previous section, a comparative trial using both an investigational drug and a placebo treatment group is the best way to find out if the former is indeed effective.

3.17 The relationship between research questions and study design

Having introduced and discussed research questions in this chapter, the following quotes from authoritative sources emphasize their importance:

> The most critical and difficult prerequisite for a good study is to select an important feasible question to answer. Accomplishing this is primarily a consequence of biological knowledge. (Piantadosi, 2005)
>
> The essence of rational drug development is to ask important questions and answer them with appropriate studies. (ICH Guidance E8, 1997)

Between them, these quotes capture the notion that, once an important research question has been formulated, a "good" and "appropriate" study must be conducted to answer the question. As you have seen already, we provide operational definitions of terms that are used in statistical contexts. Accordingly, Chapter 4 provides operational definitions of the terms "good" and "appropriate" in the context of deciding how best to answer an important research question.

Before that, however, it is informative to consider the second sentence in the quote from Piantadosi (2005) – "Accomplishing this [selecting an important feasible question to answer] is primarily a consequence of biological knowledge." This book discusses the employment of the discipline of Statistics in a particular context, the development of a new drug. It is certainly true that the individual statistical analyses that we teach you can be informatively applied in other areas of investigation, but in this book their application is in the development of a biologically active drug that will influence patients' biology for the better (see Turner, 2007).

3.18 Review

1. What are characteristics of a useful research question?
2. Why is a placebo control used in many clinical trials?
3. When would a placebo control not be used?
4. How do the null and alternate hypotheses relate to the objective of a clinical trial?
5. What form do the null and alternate hypotheses take for a superiority trial with a placebo control?
6. What form do the null and alternate hypotheses take for an equivalence trial with an active control?

3.19 References

Fleming TR, DeMets DL (1996). Surrogate endpoints in clinical trials: are we being misled? *Ann Intern Med* **125**:605–613.

ICH Guidance E8 (1997). *General Consideration of Clinical Trials*. Available at: www.ich.org (accessed July 1 2007).

ICH Guidance E12A (2000). *Principles for Clinical Evaluation of New Antihypertensive Drugs*. Available at: www.ich.org (accessed July 1 2007).

Norgrady T, Weaver DF (2005). *Medicinal Chemistry: A molecular and biochemical approach*, 3rd edn. Oxford: Oxford University Press.

Piantadosi S (2005). *Clinical Trials: A methodologic perspective*, 2nd edn. Chichester: John Wiley & Sons.

Temple R, Ellenberg S (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. *Ann Intern Med* **133**:455–463.

Turner JR (2007). *New Drug Development: Design, methodology, and analysis*. Hoboken, NJ: John Wiley & Sons.

4

Study design and experimental methodology

4.1 Introduction

Chapter 3 introduced the central topic of research questions in clinical trials. It also introduced the null and alternate hypotheses. This chapter discusses the relationship between research questions and study design, and shows how optimum experimental methodology is critical to the successful implementation of studies conducted to answer research questions.

Each study in a clinical development program addresses one or more research questions (we noted in the previous chapter that it is a good idea to limit the number of research questions in any given trial). In Chapter 3 we also noted two characteristics that a research question must possess to be considered useful:

- It needs to be specific (precise).
- It needs to be testable.

We can now take this thinking one step further. Formulating a good research question and then fine-tuning it is critical to the potential success of a trial. The research question is the driving force behind the way that the trial will be designed and implemented, because certain trial designs, or study designs, are needed to permit the acquisition of data that can be used successfully to answer the research question. The best research question in the world cannot be answered by the acquisition of inappropriate data via the conduct of an inappropriately designed trial, no matter how well the data are collected.

A useful research question suggests how a study needs to be designed to provide the appropriate information to answer the question. Choosing the best study design to answer the research question is therefore critical. The word "best" in the previous sentence is meaningful because there may be more than one study design that is capable of providing data that enable the question to be addressed and answered, but one of these designs may be more appropriate than the other possibilities.

This occurrence illustrates an important point. It is certainly true that the discipline of Statistics contains precise aspects that provide definite answers in situations where that answer is the only possible correct answer. However, it is also true that the successful practice and implementation of the discipline of Statistics require a considerable amount of well-informed judgment. It is therefore vital that professional statisticians are involved in all aspects of clinical trials. This comment may initially come as somewhat of a surprise: This is because there is a widespread tendency to think of statisticians being involved only at the end of a trial when all the data have been collected. This misperception is as unfortunate as it is widespread. The conduct of a successful trial requires that statisticians are involved from Day 1, which can be thought of as the time when a research team first decides that knowledge about a certain characteristic of the investigational drug is needed, to the end of the entire process of acquiring and disseminating the trial's results. This includes submitting the results from the trial to a regulatory agency (multiple agencies if marketing permission is desired in multiple countries) and publishing the results in a clinical communication for other research scientists and clinicians to read.

This chapter also discusses experimental methodology. As well as employing the best study design to facilitate the collection of data to answer a research question most appropriately and successfully the data acquired for this purpose must be of optimum quality. For example, each and every individual's blood pressure must be measured as accurately as possible every single time that a measurement needs to be made. The appropriate choice of the best study design and the implementation of optimum quality methodology work hand in hand to facilitate the acquisition of optimum quality data with which to answer the research question that led to the clinical trial being conducted.

4.2 Basic principles of study design

At the end of Chapter 3 we noted that we would provide operational definitions of the terms "good" and "appropriate study design" in the context of answering an important research question of biological (clinical) importance. Two more quotes from Piantadosi (2005) and ICH Guidance E8 (1997) are illuminating:

> Conceptual simplicity in design and analysis is a very important feature of good trials . . . . Good trials are usually simple to analyze correctly.
>
> Piantadosi (2005, p 130)

> Clinical trials should be designed, conducted, and analyzed according to sound scientific principles to achieve their objectives.
>
> ICH Guidance E8 (1997, p 2)

The first quote provides an excellent operational definition of the term "good" in the context of the design of clinical trials. It also captures a sentiment to which we return time and time again in this book. Study design and statistical analysis "are intimately and inextricably linked: the design of a study determines the analysis that will be used once the data have been collected" (Turner, 2007, p 5). Conceptual simplicity, as Piantadosi (2005) noted, is very important when designing a trial. As design and analysis are intimately linked, conceptual simplicity in design leads to conceptual simplicity in the associated statistical analyses.

Our operational definition of the term "appropriate" in the context of the design of clinical trials has two aspects, one of which comes from ICH Guidance E8 (1997) as cited earlier: Trials need to be designed, conducted, and analyzed according to sound scientific principles. This chapter discusses the scientific experimental methodology that is appropriate for the design and conduct of trials. The second aspect of our operational definition of the term "appropriate" is that the design employed must be capable of providing the data needed to answer the research question of interest. There are many study designs, each of which is appropriate for providing the data necessary for specifically formulated research questions. A design that cannot possibly provide the data to answer the research question of interest is not appropriate, and the decision not to use that design is therefore clear cut. In some circumstances more than one study design is capable of providing the data needed to answer the research question. In this case the decision as to which one is the most appropriate requires a decision based on an informed judgment, and statisticians must be involved in this decision.

Clinical trials embody several fundamental principles of experimental design (Piantadosi, 2005). Three of these are:

1. replication
2. randomization
3. local control.

4.2.1 Replication

Replication refers to the fact that clinical trials employ more than one individual in each treatment group. The reason for this is that there is considerable variation in how individuals respond to the administration of the same drug, so it is not appropriate to choose only one individual to receive the investigational drug and another to receive the placebo: There is no way of knowing how representative the individuals' responses are of the typical responses of people in general.

Replication allows two important features of individuals' responses to the investigational drug to be assessed. One is just how different their responses are from each other. It might be that all individuals show responses that are pretty close to each other, or that there is a considerable difference between individuals' responses. The second is to evaluate the "typical" response of all the individuals. Both of these features are important assessments in the discipline of Statistics. Chapter 5 talks about these assessments and puts these ideas into statistical language. It also provides operational definitions of the term "typical": We use the plural term "definitions" because the typical response can be operationally defined in several ways, each of which is appropriate in certain circumstances.

4.2.2 Randomization

The goal of randomization is to eliminate bias or, in practical terms, to reduce bias to the greatest extent possible. Bias is the difference between the true value of a particular quantity and an estimate of the quantity obtained from scientific investigation. Various influences can introduce error into our assessment of treatment effects, and these are discussed at various points in the following chapters. At this point we discuss an example of systematic error, or bias.

Randomization involves randomly assigning experimental individuals to one of the treatment groups, the drug treatment group or the placebo treatment group. The premise of randomization is simple: Many potential influences on the drug response of individuals participating in the trial (for example, differences in the heights and weights of participants, differences in metabolic pathways involved in the metabolism of the investigational drug) cannot readily be controlled for. It is therefore important that, to the best of our ability, we take steps to ensure that these characteristics are likely to be equally represented in both treatment groups. If all of the individuals in one treatment group share a characteristic that is not present in any of the individuals in the other treatment group, it is not possible to ascribe differences between the groups to the one influence of central interest, that is, the different treatments received by the two groups. Putting all relatively tall individuals into one treatment group and all relatively short individuals into the other treatment group would be an example of systematic bias. In this scenario, height would have a direct impact on the formation of the treatment groups and, therefore, if height were to be a source of influence on the blood pressure change demonstrated by individuals, height could be a cause of systematic bias in the results obtained.

The preapproval clinical trials discussed in this book are experimental studies: The data collected comprise a series of observations made under conditions in which the influence of interest, the type of treatment received, is controlled by the research scientist. (The term "experimental" is used here as defined by Piantadosi [2005]. The converse of an experimental study is a nonexperimental study. Nonexperimental studies are often called observational studies, but this term is inadequate, because it does not definitively distinguish between nonexperimental studies and experimental studies, for example, preapproval clinical trials, in which observations are also made. The methodology employed in preapproval clinical trials is experimental: It comprises a series of observations made under conditions in which the influences of interest are controlled by the research scientist. The methodology employed in other types of study can be nonexperimental; the research scientist collects observations but does not exert control over the influences of interest. The term "nonexperimental" is not a relative quality judgment compared with experimental; the nomenclature simply distinguishes different methodological approaches [Turner, 2007].)

There are various types of randomization strategies. The strategy employed in the trials discussed in this book is called simple randomization, which involves assigning treatments to individuals in a completely random way. Other more complex randomization techniques include block randomization, stratified randomization, and cluster randomization: These are not addressed in this book in any detail (see Turner, 2007, for a brief review). Randomization techniques have an important role in clinical research in general and in drug development in particular because they allow for balanced assignment of treatments within strata of interest (stratified randomization), minimize the possibility of a long run of assignments to the same treatment (block randomization), and facilitate the assignment of large groups of individuals to the same treatment (cluster randomization).

4.2.3 Local control

Another important feature of conducting, or running, clinical trials is local control. This topic takes us into the realm of methodology. Tight control on all aspects of methodology – for example, the manner in which the treatments are administered, the manner in which blood pressure measurements are made, and the apparatus used to make these measurements – must be exercised at all investigative sites. As an example, it is not appropriate that blood pressures for all individuals in one treatment group be measured using one strategy and measuring device whereas blood pressures for all individuals in the other treatment group are measured differently. This naïve strategy could bias the results of the study. As noted in Section 4.2.2, an important objective of control in clinical trials is to remove as much error from the results as possible, that is, to reduce potential bias.

Environmental conditions should also be controlled as much as possible. Taking measurements and evaluating some individuals in relatively cold conditions and others in a relatively warmer environment is not recommended. Taking this example further, and considering factors such as ease of access to the investigative site and the general atmosphere (relaxed, frenetic) of the site and its investigators, it is not appropriate to have all individuals in one treatment group enrolled at one investigative site and all individuals in the other treatment group enrolled at a different site.

4.3 A common design in therapeutic exploratory and confirmatory trials

As noted in Section 4.2, there are many study designs that are employed during a clinical development program. Some of these are typically used in early human pharmacology trials, whereas others are typically used in later therapeutic exploratory and therapeutic confirmatory trials. We discuss one particular study design more than any others, but it must be emphasized that this does not mean that it is more important than other designs. Rather, its employment as our central example allows us to introduce you to statistical methodology and statistical analysis in our chosen way.

The design that is predominantly discussed in this book is the randomized, concurrently controlled, double-blind, parallel group design. The four descriptors in this title – randomized, double-blind, concurrently controlled, and parallel group – identify different aspects of this design. We start with the last two descriptors, parallel group and concurrently controlled, because these capture the fundamental nature of the design. We then discuss the first two descriptors, randomized and double blind.

4.3.1 The concurrently controlled, parallel group design

Individuals participating in a parallel group trial are randomly assigned to one of two or more distinct treatments. Those who are assigned to the same treatment are frequently referred to as a treatment group. While the treatments that these groups receive differ, all groups are treated equally in every other regard, and they complete exactly the same procedures. This parallel activity on the part of the groups of individuals is captured in the term "parallel group design."

The term "concurrently controlled" captures two aspects of this study design. We have already come across the concept that to quantify meaningfully the effect of the investigational drug (that is, the treatment effect), it is necessary to compare the average blood pressure reduction of the group of individuals receiving the investigational drug with the average of those receiving the placebo. The two parallel groups here are the drug treatment group and the placebo treatment group, and the latter functions as the control group. Sometimes this design is called a placebo-controlled parallel group design, because the control employed is a placebo and not another active, marketed drug. This is perfectly valid.

The term "concurrently" refers to the fact that the individuals in the placebo treatment group are participating in the trial at the same time as those in the drug treatment group. This is an important aspect of the study design. If all the individuals in the drug treatment group participated first, followed at some later time by all those in the placebo treatment group, several potential influences could impact the results of the trial. For example, the staff at the investigational sites at which individuals participate in the trial might have changed considerably, and aspects of the overall operation of these sites may have changed. The goal of experimental methodology is to control for all influences other than the type of treatment (drug or placebo) received by individuals, and so having all the individuals in one group participate at one time under one set of conditions and all those in the other group(s) participate at a later time under a potentially different set of conditions is not desirable. (The goal of study protocols is to detail the experimental procedures to be employed during the trial in sufficient detail that they will be executed identically by all research staff at all times. This should therefore minimize the differences just described. However, practical reality sets in here and, in the extreme example employed in the text, the study protocol probably would not be 100% successful.) This point links well with the discussion of local control in Section 4.2.3.

This last point is well acknowledged, and it is unlikely that a trial would not be "concurrently" controlled. Therefore, if the term "placebo controlled" is seen in a published report of a trial, it is almost certainly fair to assume that the trial is concurrently controlled. However, assumption is a dangerous thing, and you should check the details of the trial presented in the report to confirm this.

4.3.2 The crossover design

In contrast to the parallel design, individuals in a crossover design are assigned to receive two or more treatments in a particular sequence. For example, an individual in a crossover study may receive the drug treatment in the first period. Then, after a suitably long washout period during which the individual is off drug, he or she will receive placebo. Other individuals will receive placebo in the first period and then cross over to the drug treatment in the second period. Such a study would be considered a two-period, two-treatment, two-sequence crossover design. Crossover designs may involve a number of treatments, sequences, and periods. In a crossover design, individuals are randomized to treatment sequences, not treatment groups.

The greatest advantage of the crossover design is that individuals receive more than one treatment so that they act as their own controls. This results in more statistical efficiency and therefore smaller sample sizes. Crossover designs can, for this reason, be particularly useful in early pharmacology studies. Another advantage of crossover designs is that they can aid in recruitment of study participants when a serious condition is being treated and they would like to have access to a potentially helpful investigational drug.

Crossover designs also have some disadvantages – one is that their results can be difficult to interpret. As all individuals receive more than one treatment there can be a carryover effect from one or more early periods to subsequent periods, leading to a biased estimate of the treatment effect. When an individual does not contribute data to one of the periods, none of the data can be used in the most straightforward analyses – attributing adverse events or other untoward effects to a single treatment can be difficult. Another disadvantage is that they are not applicable to all therapeutic indications. Crossover trials are ideally suited for indications that are chronic in nature and do not vary in severity over time. For example, studying a new analgesic for migraines may not be feasible because, thankfully, migraines do not occur as frequently or predictably as would be required to evaluate a number of treatments over the course of two or more periods. Crossover designs would be better suited for chronic conditions such as hypercholesterolemia or hypertension. However, having two or more observations from the same individual under different experimental conditions introduces additional complexities (that is, dependence) for statistical analyses. These methods are beyond the scope of this book.

4.3.3 Randomization and the descriptor "randomized"

The topic of randomization was discussed in Section 4.2.2. Any trial that has employed a randomization strategy in its design is called a randomized trial.

4.3.4 Blinding and double-blind trials

We discussed blinding earlier. As a brief recap, making a drug and a placebo look, taste, and smell the same ensures one part of the double blind: It means that individuals do not know which treatment they are receiving. A second component of the blinding process is needed to ensure that the investigators – that is, those administering the treatments – do not know which treatment individuals are receiving. This component necessitates packaging the drug and placebo products at their site of manufacture so that investigators receiving them at the investigational sites cannot tell which is which. A system of codes guarantees that, when the blind is eventually broken once all the data have been acquired, it will be known which treatment each and every individual received.

The importance of double-blind trials can be expressed in both scientific and regulatory terms, with the second being a consequence of the first. These trials are the scientific gold standard, and the results from a trial that is run in a double-blind manner are afforded particular weight by regulatory agencies and clinicians.

4.4 Experimental methodology

Experimental methodology is concerned with all the aspects of implementation and conduct of a study. Experimental methodology and study design work hand in hand to ensure that optimum quality data are collected from which optimum quality answers to the research question can be provided. As we have seen, an appropriate study design must be used to allow the collection of optimum quality data, and we can think of study design as providing the opportunity to collect such data. To take advantage of this opportunity, optimum experimental methodology must be used in the acquisition of the data. Optimum quality methodology is no use if the wrong study design has been employed, and the appropriate study design can lead to optimum quality data only if optimum experimental methodology is employed.

Consider also the data analysis and interpretation that occur once data have been acquired in a trial. First, the appropriate analysis has to be employed as determined by the study design. However, the employment of this analysis alone is not enough to ensure optimum quality answers to the research question. A computationally perfect execution of the appropriate analysis, and the most meaningful interpretation of the results obtained, will not yield optimal answers if the data being analyzed are of less than optimal quality. Therefore, experimental methodology is also of critical importance.

At this point it is worth considering the length of time it takes to run a therapeutic confirmatory clinical trial. Such trials are often conducted as multicenter trials. Although the total numbers of individuals who participate in trials vary, we noted earlier that a typical number for a therapeutic confirmatory trial is 3000–5000 individuals. Each of these individuals needs to have the disease or condition of interest. It may be that 50 investigational sites are needed to enroll the total number of individuals needed for the trial. Imagine a hypothetical scenario where 5000 individuals are recruited at 50 sites, with the typical number of individuals per site at around 100 (in reality, the number of individuals recruited at each of the sites might differ considerably). Imagine also that the treatment length employed in this trial is 12 weeks. That is, investigational site 01 recruits a total of 100 individuals, and each individual receives either the investigational drug or the placebo for a period of 12 weeks.

The question of interest is: How long does it take to complete the trial? Although the answer "12 weeks" tends to come to mind when first thinking about this, the answer is that it will almost certainly take much longer than that because not all of the 100 individuals will start their participation in the trial on the same day. They will be recruited into the trial, and hence start participation in the trial, in a staggered manner. It is therefore quite possible that the last of the 100 individuals might start his or her 12-week participation months (and possibly years) after the first. This will likely be true at all the investigational sites. The expression "first participant first visit to last participant last visit" is often used to describe the length of the entire trial.

In addition to giving you a feeling for how long it takes to run real-life trials, we mention this point because it emphasizes that it is essential that methodological considerations receive constant vigilance in all studies because some trials can last several years.

4.5 Why are we interested in blood pressure?

The domain of experimental methodology embraces many aspects of conducting a trial, and we do not discuss the vast majority in this book. However, it is important to make you aware of the need for optimum quality methodology, and you can learn more about this from other sources: In particular we recommend Piantadosi (2005). Our discussions focus on one aspect of methodology that is directly relevant to a central theme of this book, namely the measurement of blood pressure. Before discussing this topic, however, it is worth considering why we want to develop drugs that lower blood pressure in the first place, and why optimum quality blood pressure measurements are therefore critical.

4.5.1 Clinically relevant observations

It is possible to make all sorts of observations about people. For example, some are tall, some are short, some have blonde hair, some have dark hair, some love dogs, some love cats, some have relatively high blood pressure, and some have relatively low blood pressure. In drug development we are interested in clinically relevant observations and in making these observations during a trial. (Recall the definition of experimental studies presented earlier: In experimental studies, observations are made when the influence of interest is under the control of the researcher.) In the trials discussed in this book, we are interested in observing (measuring) blood pressure for the duration of individuals' participation in the trial with the ultimate goal of assessing the investigational drug's treatment effect, that is, how much more the investigational drug lowers blood pressure than a placebo. Therefore, the question of interest here is: Why is blood pressure a clinically relevant observation? This takes us into the realm of surrogate endpoints.

4.5.2 Surrogate endpoints

Two clinical endpoints of particular relevance are morbidity and mortality: Morbidity can lessen quality of life and make mortality more likely, and mortality speaks for itself. Not surprisingly, pharmacotherapy (along with other medical interventions) is concerned with reducing both these clinical endpoints. However, the development of morbidity can be prolonged, and the impact of drug therapy on mortality during a clinical trial can be very difficult to evaluate. As it is very unlikely that many individuals will die during (most) clinical trials, the difference in death rates between the drug treatment group and the placebo treatment group is likely to be very small, and quite possibly zero. (Mortality is unfortunately not uncommon in clinical trials in some therapeutic areas and in trials involving very ill or terminally ill patients. On these occasions, it may well be possible to detect the beneficial influence of an investigational drug by focusing on the clinical endpoint of mortality.)

It therefore becomes important in clinical trials to evaluate the influence of the investigational drug on other endpoints of relevance. These can be termed "clinically relevant endpoints" or "surrogate endpoints." Surrogate endpoints are biomarkers or other indicators that substitute for the clinical endpoint by predicting its likely behavior. Justification for the choice of these endpoints is of fundamental importance: The endpoint chosen as the surrogate needs to represent the clinical endpoint in a meaningful manner. How can this be demonstrated? The following are characteristics of meaningful and useful surrogate endpoints (see Oliver and Webb, 2003; Machin and Campbell, 2005):

- Biological plausibility: A detailed knowledge of the pathophysiology of the disease or condition of interest is helpful, as is demonstration that the surrogate endpoint of interest in the clinical trial is on the causal pathway to the clinical endpoint of primary interest.
- A detailed knowledge of the drug's mechanism of action: Coupled with similar knowledge of the pathophysiology of the disease or condition of interest, this can provide a solid basis for believing that the drug will be beneficial. (A drug can certainly be clinically beneficial even if we do not know its mechanism of action, but this does not mitigate the point made here in the context of good surrogate endpoints.)
- The surrogate endpoint predicts the clinical endpoint consistently and independently.
- They are particularly useful in cases where the clinical endpoints occur after long periods.

The choice of endpoints used in studies of new therapies may evolve over time as knowledge is gained about the natural history of the disease or the reliability of surrogate endpoints. Endpoints used to evaluate the benefits of new drugs are provided in Table 4.1 for a number of diseases. Many diseases are associated with numerous medical conditions of consequence to the patient (for example, pain and disability resulting from rheumatoid arthritis), which may be the target for a particular new therapy.

Fleming and DeMets (1996) suggested that the use of surrogate endpoints is most helpful in early therapeutic exploratory studies to study activity and decide if larger, more definitive studies are warranted. Establishing the acceptability of a surrogate endpoint is a difficult undertaking. Fleming and DeMets (1996) cautioned about the use of surrogate endpoints in Phase III confirmatory trials. One frequently cited example (CAST Investigators 1989, 1992) of a misleading surrogate endpoint is from the Cardiac Arrhythmia Suppression Trial (CAST). Ventricular arrhythmia has been established as a risk factor for sudden death. In this study, three drugs that had been approved for the control of arrhythmias (encainide, flecainide, and moricizine) were evaluated for their effect on mortality among individuals with myocardial infarction and ventricular arrhythmia. The results from this study were surprising. All three drugs were associated with higher risks of death than placebo. Hence the benefit of these drugs with respect to arrhythmia did not extend to the underlying clinical endpoint.

We are interested in high blood pressure (hypertension) because of our interest in cardiovascular disease, a leading cause of morbidity and mortality. High blood pressure is a meaningful and useful cardiovascular surrogate endpoint because it is well established that chronic high blood pressure causes cardiovascular and cerebrovascular events.

Table 4.1 Examples of endpoints used in clinical trials of experimental drugs

Disease	Example endpoints
Cancer[a]	Survival
	Objective response (reduction in tumor size for a minimum amount of time)
	Time to progression of cancer symptoms
Rheumatoid arthritis[b]	Improvement in signs and symptoms
	Radiological progression of disease
Uncomplicated urinary tract infection[c]	Eradication of bacterial pathogen
Hypertension[d]	Change from baseline SBP
Postmenopausal osteoporosis[e]	Bone mineral density
	Bone fractures

[a] Food and Drug Administration (US Department of Health and Human Services or DHHS, FDA, 2007).
[b] Food and Drug Administration (DHHS, FDA, 1999).
[c] Food and Drug Administration (DHHS, FDA, 1998).
[d] ICH Guidance E12A (2000).
[e] Food and Drug Administration (DHHS, FDA, 1994).

4.6 Uniformity of blood pressure measurement

One method of measuring blood pressure is to use a stethoscope and a sphygmomanometer; you may have experienced this in your doctor's clinic/office. Other methods include the use of various automated devices. Although we do not go into these in detail here, the important point is that considerable attention must be paid to methodological considerations. It is important that the same measurement technique be used at all the investigational sites in a trial, and that time is taken before the trial starts to train every site in the correct use of whichever measuring device is chosen. This might happen at an investigators' meeting or a central meeting of all principal investigators held before the start of the trial to address procedural consistency across sites. It is also important that every measurement at each site be made correctly, and that any routine calibration of the measurement device is conducted as mandated.

4.7 Measuring change in blood pressure over time

As antihypertensive drugs are intended to lower blood pressure, their evaluation in clinical trials requires at least two measurements. One of these is an initial measurement, typically called a baseline measurement, and the other is a measurement some time later, such as at the end of the treatment phase (the end-of-treatment measurement). These two measurements allow us to calculate a change score that represents the change in blood pressure from the start to the end of the treatment phase. Change scores can be calculated in several ways. One of these, and the method that is used in all of the examples in this book, is simply to calculate the arithmetic difference between each individual's baseline measurement and his or her end-of-treatment measurement.

It is also possible that blood pressure may be measured more than twice in a trial. If the treatment period is 12 weeks long, measurements might be taken, for example, at baseline, week 2, week 4, week 6, week 8, week 10, and week 12 (end of treatment). By taking several measurements, the change across the treatment phase can be examined in more detail. Suppose that an individual's SBP decreases by 20 mmHg from baseline to end of treatment. There are many possible patterns of change across time here. For example, most of the individual's decrease in blood pressure could happen in the first few weeks, it could decrease steadily across the 12 weeks, or most of the decrease could occur during the last few weeks. Although this level of analysis is of interest in some trials, we focus on change scores calculated by using two measurements, the baseline measurement and the end-of-treatment measurement.

4.8 The clinical study protocol

When the clinical research team has decided on their research question, and the appropriate study design and methodology to acquire optimum quality data with which to answer this question, all this information needs to be documented. The clinical study protocol is the document that is written for this purpose. Chow and Chang (2007, p 1) noted that the study protocol is "the most important document in clinical trials, since it ensures the quality and integrity of the clinical investigation in terms of its planning, execution, conduct, and the analysis of the data."

The study protocol is a comprehensive plan of action that contains information concerning the goals of the study, details of individual recruitment, details of safety monitoring, and all aspects of design, methodology, and analysis. Input is therefore required, for example, from clinical scientists, medical safety officers, study managers, data managers, and statisticians. Consequently, although one clinical scientist or medical writer may take primary responsibility for its preparation, many members of the study team make critical contributions to it.

The following are some of the fundamental components in a study protocol for a therapeutic confirmatory trial for an investigational antihypertensive drug:

- How the disease or condition of interest will be diagnosed, that is, participating individuals need to be diagnosed as hypertensive. The protocol will state the precise criteria that constitute high blood pressure in this particular study, and how and by whom determining measurements will be taken.
- Inclusion and exclusion criteria: These provide detailed criteria for individual eligibility for participation in the trial. These eligibility criteria can often represent a compromise among several perspectives, such as regulatory, medical, and logistical. For example, the most valuable information about the benefits of the new treatment will be obtained from a group of study individuals who are most representative of the patients to whom the drug will be prescribed. On the other hand, "real world" patients may be taking a number of medications or have concurrent illnesses that may confound the ability to evaluate the investigational treatment. It may be logistically impossible to study individuals with poor reading abilities because they will not comply with study procedures. Eligibility criteria define the study population, a term that is discussed in greater detail in Chapter 5.
- The primary objective and any secondary objectives (it is a very good idea to limit the number of objectives): These must be stated precisely.
- Measures of safety: The criteria to be used to evaluate safety are provided. These will typically include adverse events, clinical laboratory assays, electrocardiograms (ECGs), vital signs, and physical examinations.
- Measures of efficacy: The criteria to be used to determine efficacy are provided. Decrease in blood pressure will be the primary measurement of interest. Also, it may be the case that average decreases of a certain magnitude are required for the investigational drug to be deemed effective.
- Drug treatment schedule: Route of administration, dosage, and dosing regimen are detailed. This information is also provided for the control treatment.
- The statistical analyses that will be used once the data have been acquired. The precise

analytical strategy needs to be detailed, here and/or in an associated statistical analysis plan.

A study protocol is often supplemented with another very important document called the statistical analysis plan (sometimes referred to by similar names such as a data analysis plan or reporting analysis plan). The statistical analysis plan often supplements a study protocol by providing a very detailed account of the analyses that will be conducted at the completion of data acquisition. The statistical analysis plan should be written in conjunction with (and at the same time as) the protocol, but in reality this does not always happen. At the very least it should be finalized before the statistical analysis and breaking of the blind. In many instances (for example, confirmatory trials) it may be helpful to submit the final statistical analysis plan to the appropriate regulatory authorities for their input.

4.9 Review

1. What is the importance of replication, randomization, and local control in experimental design?
2. Define the following aspects of clinical trial study design:
   (a) double blind
   (b) concurrent control
   (c) parallel group.
3. What is the difference between a clinical endpoint and a surrogate endpoint?
4. What information is included in a study protocol?

4.10 References

Cardiac Arrhythmia Suppression Trial Investigators (1989). Preliminary report: Effect of encainide and fleicanide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* **321**:406–412.

Cardiac Arrhythmia Suppression Trial Investigators (1992). Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* **327**:227–233.

Chow S-C, Chang M (2007). *Adaptive Design Methods in Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC.

Fleming TR, DeMets DL (1996). Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* **125**:605–613.

ICH Guidance E8 (1997). *General Consideration of Clinical Trials*. Available at: www.ich.org (accessed July 1 2007).

ICH Guidance E12A (2000). *Principles for Clinical Evaluation of New Antihypertensive Drugs*. Available at: www.ich.org (accessed July 1 2007).

Machin D, Campbell MJ (2005). *Design of Studies for Medical Research*. Chichester: John Wiley & Sons.

Oliver JJ, Webb DJ (2003). Surrogate endpoints. In: Wilkins MR (ed.), *Experimental Therapeutics*. Boca Raton, FL: Taylor & Francis Group, 145–165.

Piantadosi S (2005). *Clinical Trials: A methodologic perspective*, 2nd edn. Chichester: John Wiley & Sons.

Turner JR (2007). *New Drug Development: Design, methodology, and analysis*. Hoboken, NJ: John Wiley & Sons.

US Department of Health and Human Services, Food and Drug Administration (1998). *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*. Available from www.fda.gov (accessed July 1 2007).

US Department of Health and Human Services, Food and Drug Administration (2007). *Challenge and Opportunity on the Critical Path to New Medical Products*. Available from www.fda.gov (accessed July 1 2007).