7

Early phase clinical trials

7.1 Introduction

As we noted in Section 1.11, this book focuses on teaching you the statistical methodologies and analyses that are employed in the therapeutic confirmatory clinical trials conducted before a sponsor applies for marketing approval for the drug that they have been developing. We also noted that there are other clinical trials that precede therapeutic confirmatory trials. Two other categories of preapproval trials mentioned are Phase I (human pharmacology) trials and Phase II (therapeutic exploratory) trials. Therefore, before focusing on therapeutic confirmatory trials in Chapters 8–11, it is appropriate to provide an overview of human pharmacology and therapeutic exploratory trials.

The usefulness of numerical information from clinical trials in decision-making is an ongoing theme in this book. The first few clinical studies for new drugs are important because they provide information relevant to the critical decisions that must be made with regard to continued investment in the development program. Ideally, studies are designed to answer research questions, the answers to which provide sufficient information to inform the next step of development, that is, either to go forward (a "go" decision) or not to go forward (a "no-go" decision). The answer to these critical early questions must be "go" if we are to reach the later stages of clinical development. For example, we need to have reasonable confidence that the drug is safe enough to progress to therapeutic exploratory trials in which it will be administered for the first time to participants with the disease or condition of interest.

In addition, we need to have reasonable confidence that a particular selected route of administration will prove successful for administering the drug to patients if and when the drug is approved. Although many other questions must be addressed during later-stage clinical development, these critical early phase questions have significant bearing on the ultimate safety and efficacy attributes of the product, as well as commercial implications (for example, route or schedule of administration).

Discussions in this chapter emphasize statistical considerations in early phase clinical trials. These include study designs employed, the types of data collected, and the usefulness and limitations of these data.

7.2 A quick recap of early phase studies

Human pharmacology studies are pharmacologically oriented trials that typically look for the best range of doses to employ. These trials typically involve healthy adults. Comparison with other treatments (such as a placebo or a drug that is already marketed) is not typically an aim of these trials, which are undertaken in an extremely careful manner in very controlled settings, often in residential or inpatient medical centers. Typically, between 20 and 80 healthy adults participate in these relatively short studies, and participants are often recruited from university medical school settings where trials are being conducted. The main objectives are to assess the safety of the investigational drug, understand the drug's pharmacokinetic profile and any potential interactions with other drugs, and estimate pharmacodynamic activity. A range of doses and/or dosing intervals is typically investigated in a sequential manner.

From a statistical viewpoint, the design of human pharmacology studies has certain implications. They include a relatively small number of participants, but a lot of measurements are collected for each participant. This strategy has both advantages and limitations. The extensive array of measurements made allows the drug's effects to be characterized reasonably thoroughly. However, as so few participants participate in these studies, generalizations to the general participant population are relatively more tenuous than for studies with larger sample sizes.

7.3 General comments on study designs in early phase clinical studies

A disappointing result in early clinical studies, as a result of either a real liability of the investigational drug or chance alone, can doom the prospects for the new drug ever entering the market. No-go decisions are a logical consequence of such disappointing results. To provide optimum quality data and the associated optimum quality information upon which to base go and no-go decisions, early clinical studies are very well controlled, thereby limiting extraneous sources of variation as much as possible. Early clinical studies, especially FTIH (first-time-in-human) studies, are typically conducted at a single investigative center. As a relatively small number of participants are studied in such early phase trials, a single center can feasibly accommodate the study by itself. It can recruit enough participants at that single location, and provide all the necessary resources for investigators at that site to conduct all the study procedures documented in the study protocol. Conducting a study at a single center ensures greater consistency with respect to participant management, study conduct, and assessment of adverse events, and provides for frequent and careful monitoring of study participants.

Participants in early clinical studies are usually healthy adults whose health status is carefully documented at the start of the study through physical examinations, clinical laboratory tests, and medical histories. Limiting early studies to healthy participants allows the sponsor to attribute any untoward findings to the drug, or to a particular dose of the drug, as significant background diseases are all but absent.

Early clinical studies frequently involve the use of a concurrent inactive control. This can be important because the study procedures can be somewhat invasive and associated with some adverse effects themselves – for example, frequent blood draws resulting in a lowering of hematocrit. Without a concurrent control arm (even in a study of healthy participants) study sponsors and investigators would not be able to rule out a drug effect when observing such occurrences, which are expected, easily explained, and non-drug related. In early studies that involve inpatient facilities for close monitoring, other controls may be instituted, for example, standardized meals and set times for study procedures.

7.4 Goals of early phase clinical trials

Early clinical trials used in new drug development typically have the following goals:

- characterize the pharmacokinetic profile of the investigational drug
- describe the safety and tolerability of the investigational drug in study participants who do not have significant medical conditions
- describe the extent to which a pharmacodynamic effect is affected by different doses of the new drug
- begin to identify a dose range that would likely provide adequate exposure to yield an important clinical effect.

Although somewhat overly simplistic (especially to readers who are students of pharmacy) we can consider pharmacokinetic effects as "what the body does to the drug" and pharmacodynamic effects as "what the drug does to the body." For those readers who are less familiar with pharmacokinetics and pharmacodynamics, Tozer and Roland (2006) provide an excellent and very readable introduction to these topics.

Patients with diseases or conditions of interest can have a number of attributes that, although

very important in the context of the eventual use of the new drug, make accurate assessments of the safety and pharmacokinetics of the investigational drug difficult. For example, patients with the disease may have compromised kidney or liver function, which would confound the characterization of the metabolism of the new drug. Similarly, patients with the disease may take several other medications for the disease under study or for other related or unrelated diseases. It then becomes difficult to ascertain in early studies whether potential adverse effects or laboratory abnormalities are attributable to the investigational drug, to concomitant drugs, or to any potential interactions between the investigational drug and other drugs. (Drug interactions are not discussed in this book and readers are referred to Hansten [2004].)

The employment of healthy participants in early clinical studies provides essential information about the pharmacokinetics, pharmacodynamics, and safety of the new drug. This chapter focuses on the research questions relevant to early human studies, the designs used to address them, the data and analysis approaches commonly encountered, and the development decisions that are made as a result of these studies.

We should note here that there are some special cases for which the use of healthy participants is not justified in early studies. For particularly invasive therapies (for example, implantation of a medical device) or therapies with known toxicity (for example, oncologics) it is not ethical to study healthy participants. The use of healthy participants in early studies may also provide a misleading result for future studies of participants with disease. For example, the maximum tolerated dose of new antidepressants or anxiolytics may differ quite markedly between healthy participants and those with the disease.

7.5 Research questions in early phase clinical studies

In the early clinical development of a new drug, the following questions arise:

- How does the magnitude of systemic exposure to the new drug differ as a function of increasing concentrations of the drug?
- How does the magnitude of systemic exposure to the new drug differ as a function of different dosing schedules (for example, once, twice, or three times a day)?
- How do varying degrees of drug exposure modify measurable pharmacodynamic effects?
- How does the total amount of drug exposure from the route of administration being studied (for example, oral) compare with the total amount of drug exposure when administered parenterally (that is, intravenously or intra-arterially)? In other words, how bioavailable is the drug?
- How safe is the new drug? Evaluations include clinical laboratory tests, physical assessments, vital signs, adverse events, and cardiac effects through electrophysiological monitoring via an ECG.

To address the first four research questions listed above, pharmacokinetic data are typically collected at various time points in early clinical studies: These data are discussed in the next section. To evaluate the difference between background variation (influences that are not directly of interest) and changes brought about by the administration of drug (influences that are of interest), measurements are collected on several occasions before the start of the drug, at several times during drug administration, and at least once after the administration of the drug when its effect is likely to be minimal (for example, 24 hours later). Evaluation of the fifth question is discussed in Section 7.10.

7.6 Pharmacokinetic characteristics of interest

Investigations at this stage of a clinical development program focus primarily on a very careful evaluation of how well the drug reaches the bloodstream, and how its concentrations in the bloodstream change over time, that is, on pharmacokinetics. The extent and duration of a drug's presence in the bloodstream determine

how good a chance it has ultimately to exert its intended clinical effect by reaching and interacting with its target receptors, the domain of pharmacodynamic investigation. Therefore, we need to study pharmacokinetic factors before studying the actual clinical effects of the drug in therapeutic exploratory trials, trials in which the relationship between drug concentrations and clinical response are typically addressed for the first time.

As mentioned in Section 2.5, the term "pharmacokinetics" generally refers to the absorption, distribution, metabolism, and excretion (ADME) of a drug. When developing a new drug a great deal of time and effort is devoted to formulating the drug so that it has the most desirable characteristics from the standpoint of safety, efficacy, and commercial concerns (for example, patient convenience and patient adherence to the prescribed regimen). There are several commonly used summary measures that are useful for quantifying absorption and excretion. In contrast, metabolism and distribution are not as easy to define in terms of quantifiable measures, although it is possible to characterize how a drug is metabolized through the identification of certain markers.

7.6.1 Total systemic exposure

Total systemic exposure to an administered drug is usually measured by the area under the drug concentration curve. For each participant the drug concentration (in nanograms/milliliter) can be plotted as a function of time, as displayed in Figure 7.1. The maximal drug concentration (C_{max}) and the time at which it is observed (t_{max}) are also shown. These two parameters are discussed in Section 7.6.2.

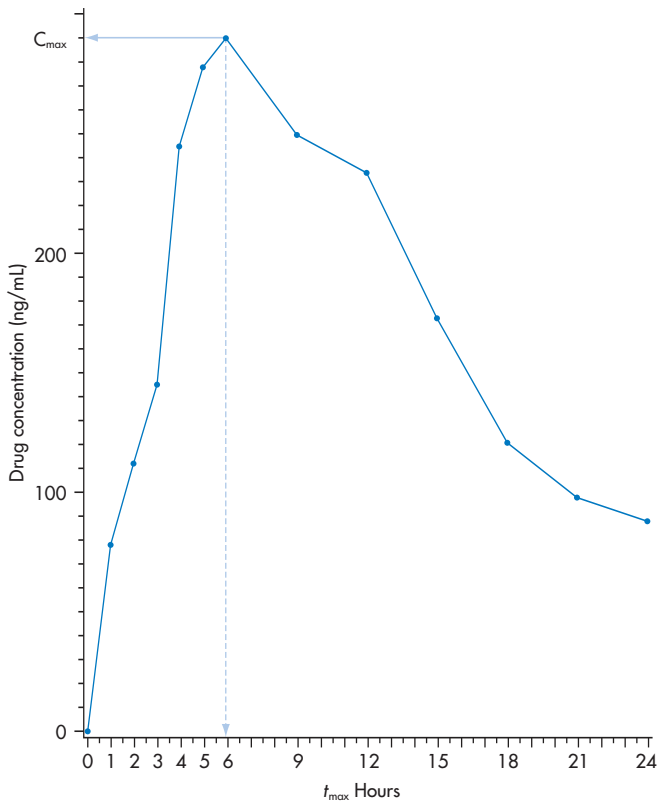

Figure 7.1 Sample drug–concentration time curve for a single participant (C_{max} of 290 ng/mL and t_{max} of 6 hours)

The estimated area under the curve from time point zero to infinity ($AUC_{(0-\infty)}$) is calculated using the trapezoidal rule. There are two steps in this process:

1. Calculate the trapezoidal area between all adjacent time points
2. Sum all areas calculated in the first step.

This calculation is an estimate of the real $AUC_{(0-\infty)}$, but a meaningful and useful estimate. By progressively increasing the sampling frequency we could obtain more and more precise measurements, but pragmatism dictates frequency, and a reasonable frequency produces a useful estimate of $AUC_{(0-\infty)}$. $AUC_{(0-t)}$ denotes the area under the curve from 0 to any time point t.

7.6.2 Maximum concentration

Another important measure of absorption is the peak or maximum concentration or maximum systemic exposure (C_{max}). It may be of interest to know the C_{max} associated with a beneficial effect. However, it is more common to use the value of C_{max} to provide assurance that, despite observing a specific C_{max} value, there was no unwanted toxicity. If the C_{max} is too high for a given dose of drug as measured by a clinical effect, such a finding could guide development of other formulations and treatment schedules. The C_{max} is calculated as the maximum value of the drug concentration during the period of monitoring. The time from administration to achieve the C_{max} is called t_{max}. Depending on the intended clinical use for the new drug, it may be more desirable to have shorter or longer values of t_{max}. For example, when in need of headache pain relief, we might be interested in a t_{max} that is as short as possible. As noted in the previous section, both C_{max} and t_{max} are shown in Figure 7.1, where C_{max} has a value of 290 ng/mL and t_{max} has a value of 6 hours.

7.6.3 Elimination

Elimination of a drug is measured using a quantity called a half-life ($t_{1/2}$). A half-life is the time required to reduce the plasma concentration to half its initial value. Longer half-lives can be associated with desirable characteristics (for example, longer activity requiring less frequent administration of the drug) or undesirable ones (for example, adverse effects).

7.6.4 Excretion

Excretion concerns the removal of a drug compound from the body. Both the original (parent) drug compound and its metabolites can be excreted. The primary mode of investigation here is excretion balance studies. A radiolabeled drug compound is administered and radioactivity is then measured from excretion sites (for example, urine, feces, expired air). These studies provide information on which organs are involved in excretion and the time course of excretion.

7.7 Analysis of pharmacokinetic and pharmacodynamic data

Statistical analyses of pharmacokinetic and pharmacodynamic effects are primarily descriptive in nature. As described in Chapter 6, inferential statistical methods such as hypothesis testing are used to make decisions in the presence of uncertainty, while limiting the likelihood of making decisions with unwanted consequences (for example, marketing an ineffective drug or not bringing to market an effective one). The decisions to be made in pharmacokinetic studies do not have such dire consequences nor are they directly applicable to the real world use of the new drug. Rather, the data acquired in pharmacokinetic studies are used as a starting point to identify doses, dosage forms, and dosage regimens for the new drug which, when studied in individuals with the disease, will allow a reasonable chance at evaluating the potential benefits and risks associated with its use.

Pharmacokinetic measures such as $AUC_{(0-24)}$, C_{max}, and t_{max} are analyzed as continuous measures. As seen in the example in Table 7.1, measures of central tendency and dispersion can be helpful to highlight differences among groups.

Table 7.1 Pharmacokinetic measures: Mean (SD) for three dosage regimens of a new investigational drug

	Dosage regimen		
	20 mg once a day (n = 10)	20 mg twice daily (n = 10)	20 mg three times daily (n = 10)
$AUC_{(0-24)}$ (ng h/mL)	812 (132)	1632 (264)	2237 (412)
C_{max} (ng/mL)	174 (61)	181 (74)	308 (94)
t_{max} (h)	2.4 (0.9)	2.6 (0.6)	7.4 (1.0)

Before discussing the interpretation of the data in Table 7.1, a couple of points about tabular displays like this one are worth pointing out. First, every study or analysis has a primary comparison of interest: In this case it is to compare AUC across groups. The ability of a regulatory reviewer to interpret data with respect to the primary comparison is aided by displaying data across columns for the comparison. In this case, a single summary measure of interest (AUC) represents a row and the groups may be compared by reading left to right. Secondary comparisons should then be placed as rows on a table.

A common example of a secondary comparison in pharmacokinetic studies is the concentration of drug at various time points during the study. It is important to know how the within-group average concentration changes over time, but it is more important to know how the mean concentration differs among groups at one time point. The fundamental nature of clinical trials is comparative, above all else. The second point about tabular displays such as this one is that the table itself is well labeled with titles and column headers. In the regulated world of drug development, presentation is extremely important. "Substance" is our first concern, but "style" is certainly important to convey the substance to a regulatory reviewer.

Descriptive analyses were discussed in Chapter 5, particularly measures of central tendency and dispersion. Those discussions now enable us to examine the pharmacokinetic data presented in Table 7.1. As can be seen, a total of 10 participants were studied in each group. The mean (SD) AUC values were 812 (132), 1632 (264), and 2237 (412), respectively. From these results we can conclude that the three times daily dosage regimen resulted in overall greater systemic exposure to the drug.

7.7.1 Decisions and inferences from FTIH studies

Having completed one or more pharmacokinetic and pharmacodynamic studies in early development, a multidisciplinary team, consisting of clinical scientists, regulatory specialists, pharmacologists, and statisticians, will examine these early clinical data to plan for studies of early efficacy and safety in individuals with the disease or condition of interest. They will interpret the data to decide which combinations of dosage forms, concentration, and regimen resulted in the optimal exposure to the drug with minimal apparent toxicities. In many instances the data may be too ambiguous to make clear decisions, especially as a degree of subjectivity is present in such decisions. For example, two regimens may have similar total exposure (as measured by AUC) but one may be associated with a greater C_{max}, which may lead to adverse effects in subsequent studies. These judgments and decisions are fairly imperfect anyway because the relationship between pharmacokinetics and clinical effects is more relevant.

For this reason alone, early clinical studies are not considered definitive and most sponsors are wise to interpret the data carefully. Ideally, certain combinations of dosage forms, drug concentrations, and regimens can be eliminated from future consideration as a result of early

studies (for example, inadequate exposure, too much exposure). Pharmaceutical companies can then conduct future research on the drug in forms that may realistically provide a benefit. AUC and C_{max} are very useful measures in initial clinical development activities and, accordingly, we need statistical methods and analyses to assess them in a scientific and therefore informative manner. However, these are not the parameters that are of ultimate interest: It is the clinical benefits and risks of the new drug that are ultimately the characteristics of importance. The statistical evaluation clinical benefits (therapeutic efficacy) and risk (adverse events, etc.) are covered in Chapters 8–11.

7.8 Dose-finding trials

A drug's dosing regimen comprises the dose of the drug given and the schedule on which it is administered – that is, both concentration and timing are important characteristics. A variety of dosing regimens may be explored in these trials, and the specific regimens chosen in a specific trial depend on the objectives of the trial and the type of drug being studied.

Dose-finding studies are conducted to provide information that facilitates selection of a safe and efficient drug administration regimen. Chevret (2006a, p 5) defined dose-finding trials as "early phase clinical experiments in which different doses of a new drug are evaluated to determine the optimal dose that elicits a certain response to be recommended for the treatment of patients with a given medical condition." Chevret (2006a) also provided some related definitions that are helpful:

- Dose: The amount of active substance that is given in a single administration or repeated over a given period, as dictated by an administration schedule of equal or unequal single doses at equal or unequal intervals.
- Response: The outcome of interest in study participants. This can be defined in pharmacodynamic terms as the therapeutic points of interest, or in terms of pharmacotoxicity/tolerability of the drug.
- Maximum tolerated dose (MTD): The highest dose that produces an "acceptable" risk for toxicity or, expressed differently, the dose that, if exceeded, would put individuals at "unacceptable" risk for toxicity.
- Minimally effective dose (MED): The dose that elicits a specified lowest therapeutic response.

Before continuing with our current discussions, the word "acceptable" in the third bullet point may initially seem somewhat incongruous here. All drugs lead to some side-effects, that is, some adverse events. Therefore, there is some degree of risk associated with taking any drug. To be useful, a drug needs to have an acceptable benefit–risk ratio – that is, the benefit must be larger than the risk, and it must be larger by a certain amount. Stating the precise amount by which a drug must provide more benefit than it may lead to harm is a difficult judgment call that must be made ultimately by physicians. However, we can make some observations. If a drug that is extremely beneficial to very sick patients shows relatively strong side-effects, a clinician may well decide that the benefit–risk ratio is still acceptable. In contrast, a drug taken for a relatively mild condition such as a headache would need to show relatively much less strong side-effects for the benefit–risk ratio to be acceptable.

Human pharmacology studies often involve dose-finding trials that focus on the evaluation of MTDs such as trials in oncology. It is important to note that studies that aim to define an MTD require clear and consistent definitions of toxicities and toxicity grades. In many disease areas outside cancer it is difficult to define the MTD in a clear manner because the drugs themselves may not be as apparently toxic as a new chemotherapeutic. Dose-finding trials that focus on the evaluation of MED are commonly referred to as early Phase II trials. (As noted in Chapter 2, the categorization of clinical trials into Phase I, II, or III, although very common, can result in confusing and less than definitive nomenclature. Here, the nomenclature "early Phase II trials" is used to distinguish these trials from therapeutic exploratory or "late Phase II trials.") One common design for FTIH studies is a dose-escalation cohort study. In this design the

first cohort consists of participants administered the lowest dose of the drug or placebo. An assessment of the safety of the first dose is undertaken and, if the lowest dose is considered safe, a second cohort of participants is studied at the next highest dose. Additional cohorts are studied in this manner until a dose has been found to have unacceptable risks or the maximum dose has been studied in the final cohort. Chevret (2006b) provides a comprehensive discussion of dose-finding experiments.

7.9 Bioavailability trials

Another type of early clinical study may be conducted with the primary objective of establishing the bioavailability of a particular dosage form, concentration, and regimen. Bioavailability can be defined as the proportion of an administered dose that reaches the systemic circulation in an unchanged form. Maximum bioavailability results after an intravenous injection of the drug. In this case, the bioavailability is by definition 100%. When administered orally, however, a drug experiences first-pass metabolism, also called first-pass loss, before it reaches the systemic circulation.

Metabolism is a complex and tremendously beneficial process in most cases, but one that poses interesting challenges in pharmacological therapy. We are constantly exposed to xenobiotics, substances that are foreign to our bodies. For example, our modern environment is a constant source of xenobiotics that are toxicants. These can enter our bodies via our lungs as we breathe and our stomachs as we eat, and some can enter the body through our skin. In addition, animal and plant food contains many chemicals that have no nutritional value but do have potential toxicity. Fortunately, our bodies are very good at getting rid of bodily toxicants. The processes of metabolism and excretion are involved in this. As noted by Mulder (2006), metabolism can be divided into three phases:

1. Phase 1: The chemical structure of the compound is modified by oxidation, reduction, or hydrolysis. This process forms an acceptor group.
2. Phase 2: A chemical group is attached to the acceptor group. This typically generates metabolites that are more water soluble and therefore more readily excreted.
3. Phase 3: Transporters transport the drug or metabolites out of the cell in which Phase 1 and Phase 2 metabolism has occurred.

Along with all animals, humans have a wide variety of xenobiotic-metabolizing enzymes that convert a wide range of chemical structures to water-soluble metabolites, which can be excreted in urine. Humans have a high concentration of these enzymes in the gut mucosa and the liver. This arrangement ensures that systemic exposure to potentially toxic chemicals is limited. A high percentage of these may be caught in first-pass metabolism. Xenobiotics that are absorbed from the intestine travel via the hepatic portal vein to the liver, the major organ of metabolism, before being circulated systemically, and metabolism in the liver means that damage to the rest of the body is ameliorated. Under normal circumstances this is extremely advantageous.

From the point of view of pharmacological therapy, however, this protective system represents a considerable challenge. Orally administered drugs also travel via the hepatic portal vein to the liver before being circulated systemically. Therefore, before the drug gets a chance to exert any therapeutic activity in the body, it has to withstand this first attempt to degrade it. This first-pass metabolism is more or less effective depending on factors including the drug's chemical and physical properties, but almost certainly there will be some degree of degradation. This means that most orally administered drugs display less than 100% bioavailability.

The most rigorous quantitative way to assess the extent of bioavailability for an orally administered drug is to compare the areas under the respective plasma–concentration curves after oral and intravenous administration of the same dose of drug. The AUC is then calculated for both, and a ratio calculated by dividing the AUC for the oral administration by that for the intravenous administration. If the area ratio for the

drug administered orally and intravenously is 0.5 (which can be expressed as 50%), only 50% of the oral dose was absorbed systemically.

Consider the development of a new drug that is going to be given orally. Assessing its bioavailability is important. An intravenous infusion of the new drug will result in a certain systemic exposure as measured by the AUC. This amount of systemic exposure will by definition be called 100% bioavailability. It is important to identify the dosage form and schedule that provide relatively high bioavailability. In this case, participants may be randomly assigned to receive one of the following drug administration regimens:

- intravenous infusion of the drug for 4 hours
- 10 mg tablet once a day
- 10 mg tablet twice a day
- 10 mg three times a day
- 20 mg tablet once a day
- 20 mg tablet twice a day
- 20 mg three times a day.

At the end of the study, the pharmacokinetic characteristics of the drug would be evaluated, and the systemic exposure for each dosage regimen compared with the intravenous route of administration.

7.10 Other data acquired in early phase clinical studies

As we saw in Section 7.5, one research question of interest in early phase trials is:

- How safe is the new drug? Evaluations include clinical laboratory tests, physical examinations, vital signs, adverse events, and cardiac effects through ECG monitoring.

More extensive discussion of these safety assessments is provided in the following chapters, but it is useful to introduce these topics at this point.

7.10.1 Clinical laboratory tests

There is a very wide range of clinical chemistry tests that can be conducted, including liver (hepatic) and kidney (renal) tests. These are discussed in Section 9.2.

7.10.2 Physical examinations

Although perhaps not as sensitive as other safety assessments, physical examinations are still very helpful, because a general exam may identify more pronounced effects to the drug such as allergic reactions or edema (fluid retention). Data collected from physical exams include a subjective assessment by the investigator as to whether the participant has "normal" or "abnormal" function for each body system (for example, respiratory, dermatologic) examined. If the body system is considered abnormal, additional descriptions of the particular abnormality are also recorded. Data recorded as normal or abnormal are measured on the nominal scale. These data are typically summarized by tabulating the number and percentage of individuals with each result.

7.10.3 Vital signs

Monitoring of vital signs, including heart rate, respiration rate, and blood pressure, is carried out on a regular basis, typically several times a day. Each of these is measured on the continuous scale. Analyses of these outcomes primarily focus on measures of central tendency and dispersion.

7.10.4 Adverse events

The collection of adverse events can be based on observation by either the investigator or participant self-report. Participant self-reports of adverse events can vary according to how the information is elicited from them. It is advisable to standardize the manner in which participants are asked about how they feel during the trial. Data collected from adverse events usually include text descriptions of several characteristics of the adverse event:

- the adverse event, for example, "rash on left forearm"

- the severity or intensity of the adverse event, for example, mild, moderate, severe
- the date and time of onset
- the outcome of the adverse event (resolved without sequelae, resolved with sequelae, or ongoing)
- any treatments administered for the adverse event
- any action taken with the study drug (for example, temporarily discontinued, stopped, none)
- whether or not the adverse event is considered serious.

To standardize the reporting of adverse events, the adverse event descriptions are coded using medical dictionaries such as MedDRA (*Medical Dictionary for Drug Regulatory Affairs* coding dictionary: See, for example, Chow and Liu, 2004, p. 563) or COSTART. The original description of the adverse event provides qualitative information about the finding that may not be captured in the coded event. Both aspects – that is, coded and uncoded – are retained in the scientific database for reporting and analysis.

7.11 Limitations of early phase trials

In this chapter we have discussed the importance, and the strengths, of early phase clinical trials. Before moving on to later phase clinical trials, it is also appropriate to consider their limitations. The word "limitations" should not be seen as a negative assessment in this context. As we will discuss in Chapter 12, later-phase preapproval clinical trials also have their limitations. Acknowledgment of the strength and the limitations of any method of inquiry is legitimate and helpful: As Katz (2001, p xi) noted, "to work skillfully with evidence is to acknowledge its limits."

7.11.1 Studying pharmacokinetics in healthy participants

Studying the pharmacokinetics of a new investigational drug in FTIH studies – that is, in individuals with healthy renal and hepatic systems – results in a pharmacokinetic assessment that is somewhat artificial. In later stages of development, it may be necessary to study the drug in individuals with impaired kidney or liver function, especially if these conditions are expected in the types of patients who will be prescribed the drug if and when it is approved for marketing. However, this initial FTIH assessment can serve as a useful starting point and provide guidance for such later studies.

7.11.2 Extremely tight experimental control

It may seem paradoxical to see tight experimental control listed in a section discussing the limitations of a clinical trial. After all, in Chapter 4 we extolled the merits of such control. The issue here is related to the issue addressed in Section 7.2. Since the investigational drug is administered in such a carefully controlled manner, the generalizability of the results from these studies becomes questionable. If and when the drug is approved for marketing, patients who are prescribed the medication will be unlikely to take the medication in such a precisely controlled manner. As in many places in drug development, there are advantages and disadvantages to this strategy. We have noted the disadvantages and now focus on the advantages.

The advantage of very tight control in early Phase II (therapeutic exploratory) trials is that the "pure" efficacy of the drug can be assessed as well as possible. The drug has every chance to demonstrate its efficacy in these circumstances.

In other words, we can assess how well the drug *can* work. It is not so easy to assess how the drug *will* work if and when approved and prescribed to a very large population of heterogeneous patients who take the drug in various states of adherence with the prescribed regimen, but that is another question for another stage of the clinical development program.

7.12 Review

1. What are some reasons that inferential statistics (that is, hypothesis testing) are not used very often in early phase studies?
2. What information from early phase trials may be used to inform the study designs of therapeutic exploratory and therapeutic confirmatory trials?
3. Name three advantages or strengths of early phase trials as they pertain to the overall development of a new drug.
4. Name three disadvantages of early phase trials as they pertain to the overall development of a new drug.

7.13 References

Chevret S (2006a). Basic concepts in dose-finding. In: Chevret S (ed.), *Statistical Methods for Dose-finding Experiments*. Chichester: John Wiley & Sons, 5–18.

Chevret S (ed.) (2006b). *Statistical Methods for Dose-finding Experiments*. Chichester: John Wiley & Sons.

Chow S-C, Liu J-P (2004). *Design and Analysis of Clinical Trials: Concepts and methodologies*. Chichester: John Wiley & Sons.

Hansten PD (2004). Important drug interactions and their mechanisms. In: Katzung BG (ed.), *Basic and Clinical Pharmacology*, 9th edn. New York: McGraw-Hill, 1110–1124.

Katz DL (2001). *Clinical Epidemiology and Evidence-based Medicine: Fundamental principles of clinical reasoning & research*. Thousand Oaks, CA: Sage Publications.

Mulder GJ (2006). Drug metabolism: inactivation and activation of xenobiotics. In: Mulder GJ, Dencker L (eds), *Pharmaceutical Pharmacology*. London: Pharmaceutical Press, 41–66.

Tozer TN, Roland M (2006). *Introduction to Pharmacokinetics and Pharmacodynamics: The quantitative basis of drug therapy*. Baltimore, MD: Lippincott, Williams & Wilkins.

8

Confirmatory clinical trials: Safety data I

8.1 Introduction

The regulatory standard for the approval of new drugs for marketing can be framed in the following manner: The benefits associated with the new treatment outweigh the risks associated with the new treatment. All pharmaceutical products carry the potential for side-effects, some of which are more serious than others. Therefore, for a given investigational drug to be approved for marketing the regulatory agency needs to be presented with compelling evidence that the likely benefits to the target population with the disease or condition of interest outweigh the likely risks. This requires conducting clinical trials that employ samples selected from the target population, and use of Statistics to design these trials appropriately, collect optimum quality data, analyze and interpret the data correctly, and make inferences about the population from which those samples were drawn.

Judgments about the benefit–risk profile of an investigational drug require, by definition, consideration of both benefit and risk. This means that the therapeutic benefit of the investigational drug needs to be assessed quantitatively, and considered together with quantitative assessments of risk. In this chapter we discuss the assessment of risk in terms of evaluating the drug's safety profile. Even though we typically use the nomenclature benefit–risk profile and not risk–benefit profile, we discuss safety evaluations first because the safety of patients must be our first concern.

Safety analyses in pharmaceutical clinical trials tend to be largely descriptive because there are so many adverse events (AEs) and other safety parameters evaluated, and analysis of them leads to issues of multiplicity (see Section 8.9). As described in Chapter 6, the appropriate use of inferential statistics requires a prespecified hypothesis of interest. As knowledge is gained about an experimental therapy during its development (for example, in therapeutic exploratory trials) a specific hypothesis about the drug's safety may emerge and can then be tested appropriately. In such instances there are inferential statistical analyses that can be used for safety data, and we present some of those applicable to AEs in this chapter. (See also Chow and Liu [2004b, Chapter 13] for additional discussions of safety assessment.)

8.2 The rationale for safety assessments in clinical trials

When a clinician prescribes a new treatment for a patient for the first time, the clinician and indeed the patient may be interested in the following questions about the safety of the drug:

- How likely is it that my patient will experience an adverse drug reaction? (The term "adverse drug reaction" refers to an unwanted occurrence caused by a drug. Hence, a prescribing clinician [and researchers conducting postmarketing surveillance studies] is concerned with adverse drug reactions. During preapproval clinical trials, we do not know which treatment an individual is receiving, so unwanted occurrences are called AEs. Formal definitions are provided shortly.)
- How likely is it that my patient will experience an adverse drug reaction that is so serious that it may be life threatening?
- How will the risk of an adverse drug reaction vary with different doses of the drug?

- How will the risk of an adverse drug reaction change with the length of treatment?
- Are the typical adverse drug reactions temporary or permanent in nature?
- Are there specific clinical parameters that should be monitored more closely in my patient while he or she is receiving this treatment because of increased risks from the newly marketed drug?

At the time that a new drug receives marketing approval, the best information available upon which the clinician can form an answer to these questions is the information gathered during the preapproval clinical trials. This information is provided to the clinician (and to all patients receiving an approved drug) in the package insert. (This situation changes in due course as additional [and more detailed] safety evaluation takes place during the process of postmarketing surveillance [see Mann and Andrews, 2007]. However, this process may take several years to acquire meaningful data, and so the statement in the text may remain true for quite a while.)

A number of clinical parameters are assessed during preapproval clinical trials. This information provides the basis upon which the clinician will formulate answers to these questions. The precise set of clinical parameters employed in a given trial may vary according to the disease and the type of drug under study. In general, the safety evaluation of new drugs is intended to detect quantifiable effects in as many organs and systems as possible. In other words, when looking for risks associated with a new drug, the strategy is to "cast a wide net."

8.3 A regulatory view on safety assessment

The view of the US Food and Drug Administration (FDA) concerning safety reviews is presented in their guidance document on the safety review of new drug applications (US FDA, 2005, p 5). As this guidance states, most therapeutic exploratory and therapeutic confirmatory trials are carefully designed to establish that a new drug is efficacious, while controlling the probability of committing a type I or II error. Unless safety concerns have arisen in earlier stages of the clinical development program, these trials typically do not involve assessments of safety that are as sensitive as those designed for establishing the efficacy of the investigational drug. Quoting from this guidance:

> In the usual case, however, any apparent finding emerges from an assessment of dozens of potential endpoints (adverse events) of interest, making description of the statistical uncertainty of the finding using conventional significance levels very difficult. The approach taken is therefore best described as one of exploration and estimation of event rates, with particular attention to comparing results of individual studies and pooled data. It should be appreciated that exploratory analyses (for example, subset analyses, to which a great caution is applied in a hypothesis testing setting) are a critical and essential part of a safety evaluation. These analyses can, of course, lead to false conclusions, but need to be carried out nonetheless, with attention to consistency across studies and prior knowledge. The approach typically followed is to screen broadly for adverse events and to expect that this will reveal the common adverse reaction profile of a new drug and will detect some of the less common and more serious adverse reactions associated with drug use.
>
> US FDA (2005, p 5)

Safety evaluations of investigational drugs focus primarily on estimating the risk of unwanted events associated with the drug, and, more specifically, on the risk of those events relative to what would be expected in the patient population as a whole if the drug were to be approved. Although more specialized tests and assays may be evaluated in certain instances, in this chapter and in Chapter 9 we describe statistical approaches used for the most common clinical data used to assess the safety of new drugs: AEs, clinical laboratory data, vital signs, and changes in ECG parameters. This chapter focuses on discussions of AEs. Adverse events are nominal data, and therefore summaries of AEs are based on counts.

8.4 Adverse events

ICH Guidance E6 (R1) (1996, p 2) provides the following definition of the term adverse event:

> Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with the treatment. An adverse event (AE) can therefore be any unfavourable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not associated with the medicinal (investigational) product.

ICH Guidance E2A (1995, p 3) provides a definition of the term "adverse drug reaction" that is applicable during preapproval clinical experiences with a new medicinal product:

> All noxious and unintended responses to a medicinal product related to any dose should be considered adverse drug reactions.

There are various types of AEs, as shown in Table 8.1.

The length of observation for AEs is typically specified in the study protocol. In most instances, on-treatment AEs (also called treatment-emergent AEs) are considered to be those events with an onset from the time that study treatment has been initiated through the protocol-defined follow-up period. For example, a protocol may specify that AEs occurring within 30 days of the last exposure to the study drug be reported. In some therapeutic areas it may be desirable to assess separately those AEs that occur once treatment has been discontinued, for example, to evaluate withdrawal or rebound effects during the follow-up period.

In the hypothetical data presented in Table 8.1, the numbers of participants in the drug and placebo groups are deliberately similar but not identical. This is why provision of both absolute numbers and percentages is so informative when making comparisons between the treatment groups.

8.5 Reporting adverse events

Adverse events are typically reported in one of two ways:

1. By study investigators on the basis of their own observations (for example, from a physical exam)
2. By the study participant as a self-reported event.

In the second case, it is advisable to elicit AEs from participants using a standardized script to ensure that they are collected as accurately as possible. For example, a question such as "Have you noticed anything different or had any health problems since you were last here?" is a

Table 8.1 Participant accountability (Safety Population: Study AB0001)

Adverse events (AEs)	Number (%) of participants	
	Placebo (n = 2603)	Drug (n = 2456)
Pre-treatment AEs	24 (1)	31 (1)
On-treatment AEs[a]	297 (11)	386 (16)
Drug-related AEs[b]	31 (1)	42 (2)
Serious AEs	20 (1)	27 (1)
AEs leading to withdrawal	12 (< 1)	17 (1)

[a]AEs that occur on any treatment, whether active or nonactive.

[b]"Drug-related" is a designation made by an investigator who decides that there is a reasonable chance that the AE was caused by the treatment being taken.

way of asking a participant about potential AEs without leading him or her to answer in a certain way.

Study personnel who interact with participants are trained to capture the essence of any self-reported AEs on a case report form (CRF), one of the most important documents in clinical trials. Examples of reported AEs include "shortness of breath," "rash on left wrist," "dry mouth," and "vomiting." In addition to the description of the nature of the AE, additional information such as the following is typically collected:

- the severity
- the date and time of onset
- the resolution date (if the event resolved), any action that was taken with the study drug (for example, stopped, dosage reduced)
- the presumed relationship to the study treatment
- whether or not the AE was considered "serious" according to a regulatory definition.

8.6 Using all reported AEs for all participants

The first question listed in Section 8.2 was: "How likely is it that my patient will experience an adverse drug reaction?" We turn this question around, and reframe it in terms of assessing how likely it is that a participant in a preapproval clinical trial will experience an AE. The data that are typically used to answer this question are all on-treatment AEs for all participants treated (or exposed) in each treatment group. The probability that a participant in a particular treatment group will report any AE is estimated by the proportion of participants in the group who reported any AE.

When describing proportions, it is important to note what event is being counted in the numerator and what event in the denominator. Many times it is clear what the appropriate numerator for a proportion should be, but not so clear what the appropriate denominator should be. The simplest starting point for determining which participants should be counted in the denominator is to identify all those who are at risk of experiencing the event of interest. For example, the proportion of participants experiencing an AE in the first 90 days should be calculated by counting the number of participants who were treated for at least 90 days in the denominator and the number of participants who were treated for at least 90 days and reported an AE in the first 90 days in the numerator.

As described earlier, proportions are numbers between 0 and 1. We have also noted that it is common for proportions to be multiplied by 100 so that the quantity being assessed is expressed in percentage terms. In the present context, we are interested in the percentages of participants experiencing a certain event.

The probability of an individual reporting an AE in a trial is estimated by the following proportion:

$$\frac{\text{[Number of participants who were administered the treatment and reported any AE]}}{\text{[Number of participants who were administered the treatment]}}$$

Some participants will have reported more than one AE. For this analysis, we count participants only once if they experienced any AE(s).

As noted in Chapter 6, this calculated proportion is considered a point estimate, because it was obtained from a single sample and the estimate does not take into account any variability attributed to sampling. In most clinical study reports (and, ultimately, package inserts for marketed products), the point estimate of the proportion of individuals experiencing AEs is expressed as a percentage of individuals. This quantity can be thought of as a rate (ratio of individuals experiencing an event among those exposed to the treatment) or, in the terminology used in the discipline of epidemiology, the incidence of AEs.

Calculating the proportion (or, equivalently, the percentage) of individuals reporting any AE for all treatment groups in a study enables us to see whether AEs are more or less likely in the test treatment group than in other groups. The use of an inactive control group (for example, a

placebo) in a study allows us to compare the probability attributed to the test treatment group to what can be thought of as the background risk, which is approximated for by the risk in the inactive control group.

8.7 Absolute and relative risks of participants reporting specific AEs

Similar analysis approaches are used to describe the risk, in both the absolute and relative (comparative) sense, of individuals reporting specific AEs. These analyses are much more useful clinically because not all AEs are created equal. One example is to estimate the proportion of individuals in a given group who reported a headache. To do this in a standardized manner it is necessary to "code" the AE descriptions (for example, "tension headache", "achy head"). The use of the MedDRA coding dictionary for this purpose is now widely accepted, and in some instances may be required. Coding is performed before statistical analysis and the "coded" terms are used in statistical summaries that require counting of participants reporting each event.

The proportion from the sample in the study can be estimated as:

$$\frac{\text{[Number of participants who received the treatment and reported a headache during the study]}}{\text{[Number of participants who received the treatment]}}.$$

For example, if 25 participants received treatment A and, among them, 5 reported a headache, the estimated proportion of participants reporting a headache is 5/25 = 0.20, which can also be expressed as 20%. When such an analysis is repeated for all AEs reported, and the quantities are expressed as percentages and displayed in tabular form in a package insert, it is relatively easy for prescribing physicians and their potential patients to answer their questions.

Suppose that an investigational antihypertensive drug is evaluated at multiple doses in a parallel-group placebo-controlled study. Participants in this therapeutic exploratory study were randomly assigned to receive either placebo or one of three possible doses of the test treatment (low, medium, or high). The treatment period was for 6 weeks. The number and percentage of participants experiencing any AE, and particular AEs, are displayed in Table 8.2.

We now have data with which to begin to answer the question: How likely is a patient to experience an AE after use of the new treatment? As this study included three doses of the test treatment, we need to consider the dose in our answer. Examining the top row in Table 8.2, it seems that the overall chance of observing an AE at all doses of the active treatment is similar to that for the placebo group: The percentages for "Any event" range from 10% to 13% across the groups with no apparent relationship to dose. From these data, our best guess as to the probability of an individual treated with the test treatment experiencing any AE is between 10% and 13%. However, the probability of experiencing

Table 8.2 Number and percentage of participants reporting adverse events (AEs) by group

AE	Placebo (n = 98)	Low (n = 101)	Medium (n = 104)	High (n = 97)
Any event	12 (12%)	13 (13%)	10 (10%)	12 (12%)
Headache	6 (6%)	8 (8%)	9 (9%)	8 (8%)
Dizziness	1 (1%)	3 (3%)	4 (4%)	3 (3%)
Upper respiratory infection	4 (4%)	1 (1%)	2 (2%)	2 (2%)
Nausea	1 (1%)	2 (2%)	1 (1%)	3 (3%)

an AE is almost equal to the probability of experiencing an AE after treatment with placebo. The implication of this result is that the risk of experiencing an AE after treatment with the new drug is no different than if the participant had not been treated with the drug.

Looking at the specific AEs in Table 8.2, there is really little difference among the groups with one exception, the AE of dizziness. Only 1% of participants in the placebo group reported dizziness compared with 3–4% of participants treated with the active drug. How might a regulatory reviewer interpret these data? The first conclusion is that dizziness was not reported very often in any participant group, so, if the drug is approved and marketed, most patients treated with the new drug would probably not have a problem. However, the difference in the percentage of participants might generate some concern.

Initially, the absolute difference in dizziness rates (2–3%) may not seem extreme. However, when considering the rates in relative terms, those treated with the investigational drug are three to four times as likely to experience dizziness as someone who did not receive the active drug. This measure of risk is called a relative risk and is calculated as follows:

$$\text{Relative risk} = \frac{\text{The probability of the event in group A}}{\text{The probability of the event in group B}}.$$

In many instances, the communication of a risk (probability of experiencing the event) is most clear with an absolute measure (such as the point estimate for a group) and a relative measure (such as the relative risk). The relative risk is a ratio of two probabilities, and can therefore range from zero to infinity.

8.8 Analyzing serious AEs

ICH Guidance E2A (1994) provides the following definition of a serious event: A serious adverse event (experience) or reaction is any untoward medical occurrence that at any dose:

- results in death
- is life threatening (note that the term "life threatening" in the definition of "serious" refers to an event in which the patient was at risk of death at the time of the event; it does not refer to an event that hypothetically might have caused death if it were more severe)
- requires inpatient hospitalization or prolongation of existing hospitalization
- results in persistent or significant disability/incapacity
- is a congenital anomaly/birth defect.

A similar analysis can be performed to address another question of interest: How likely is it that an individual will experience an adverse effect that is potentially life threatening? The data used to answer this question include all of the AEs that were rated as serious at the time of reporting. We would estimate the probability by calculating the proportion of participants treated in each group who experienced a serious AE. The proportions (or, equivalently, the percentages) of patients could be compared across groups to see if there was an increased risk of a serious AE associated with the new treatment.

8.9 Concerns with potential multiplicity issues

As noted earlier, safety analyses in clinical trials tend to be largely descriptive because so many AEs and other safety parameters are evaluated. If we were to perform hypothesis tests for the large number of parameters evaluated – for example, for all AEs reported in a trial – it is probable that at least one of the tests would be nominally statistically significant at the $\alpha = 0.05$ level. In most instances the statistical analysis is planned so that the probability of making a type I error is ≤ 0.05. In Table 8.2 rates were presented for five AEs (including any event). If we were interested in identifying statistically significant differences for each active dose group versus placebo, we would need to conduct 15 hypothesis tests (three dose groups to be tested against placebo for five AEs). As we saw in Chapter 6, if we test a number of hypotheses without taking into account multiplicity of comparisons we will likely commit a type I error. For the 15 tests that could

be conducted using the data in Table 8.2, it is certainly possible that one of them might have a nominal p value ≤ 0.05 by chance alone. Committing a type I error in this setting would mean concluding that the new treatment was associated with an excess risk of an AE when that really is not the case.

If a single test were considered nominally statistically significant after looking at so many other AEs, the result should be treated with a great deal of skepticism. Before making any regulatory or business decisions on the basis of such a result, medical, clinical, and statistical experts should, at a minimum, evaluate the medical and statistical plausibility of the result. Ideally, additional data would be collected to provide supporting evidence for such a finding. As we have pointed out a number of times already, statistical results such as these aid in decision-making, in concert with insights and evidence from other disciplines. This view, as it relates to analyses of safety data, is perfectly in line with the EMEA's Committee for Proprietary Medicinal Products (CPMP) (2002, p 4) guidance, *Points to Consider on Multiplicity Issues in Clinical Trials*:

> In those cases where a large number of statistical test procedures is used to serve as a flagging device to signal a potential risk caused by the investigational drug it can be generally stated that an adjustment for multiplicity is counterproductive for considerations of safety. It is clear that in this situation there is no control over the type I error for a single hypothesis and the importance and plausibility of such results will depend on prior knowledge of the pharmacology of the drug.

8.10 Accounting for sampling variation

Hypothesis tests and interval estimates of proportions are frequently presented in clinical study reports, especially in earlier studies of development when late phase studies are being planned. Accordingly, discussion now turns to analysis methods that can be used to account for sampling variation and, therefore, determine if the results observed are likely due to chance alone.

In Chapter 6 we described the basic components of hypothesis testing and interval estimation (that is, confidence intervals). One of the basic components of interval estimation is the standard error of the estimator, which quantifies how much the sample estimate would vary from sample to sample if (totally implausibly) we were to conduct the same clinical study over and over again. The larger the sample size in the trial, the smaller the standard error. Another component of an interval estimate is the reliability factor, which acts as a multiplier for the standard error. The more confidence that we require, the larger the reliability factor (multiplier). The reliability factor is determined by the shape of the sampling distribution of the statistic of interest and is the value that defines an area under the curve of $(1 - \alpha)$. In the case of a two-sided interval the reliability factor defines lower and upper tail areas of size $\alpha/2$.

If the shape of the sampling distribution is symmetric (for example, the Z or t distributions), the reliability factor used for the lower and upper limits is exactly the same, but with a change in sign. Some sampling distributions are not symmetric (for example, the F distribution for the ratio of two variances) and, therefore, the reliability factors for the lower and upper limits are not equal.

Let us now look at how we would calculate a confidence interval for a single proportion, such as a within-treatment group proportion of participants experiencing an AE.

8.11 A confidence interval for a sample proportion

The estimator for a sample proportion can be defined as follows:

$$\hat{p} = \frac{\text{number of observations with the event of interest}}{\text{total number of observations at risk of the event}},$$

which is an unbiased estimator of the unknown population proportion, P. The standard error of the estimator,

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}},$$

where $\hat{q} = 1 - \hat{p}$, the sample proportion of observations without the event of interest. The estimator $\hat{p}$ is approximately normally distributed for large samples (that is, when $\hat{p}n > 5$) so the reliability factor for interval estimates will come from the Z distribution. For now, we will consider only two-sided confidence intervals. Hence, the reliability factor $Z_{1-\alpha/2}$ will be the specific value of Z such that an area of $(1 - [\alpha/2])$ lies to the right of the cutoff value. A two-sided $(1 - \alpha)\%$ confidence interval for a sample proportion, $\hat{p}$ is:

$$\hat{p} \pm z_{1-\alpha/2}\mathrm{SE}(\hat{p}).$$

This is also a confidence interval for the parameter p, probability of success, of the binomial distribution. The use of the Z distribution for this interval is made possible because of the Central Limit Theorem. Consider the random variable X taking on values of 0 or 1, such that the sampling distribution of the sample mean (the proportion) is approximately normally distributed. A table of the most commonly encountered values of the standard normal distribution is provided in Table 8.3 for quick reference. Others are provided in Appendix 1.

Table 8.3 Selected values of Z for two-sided confidence intervals

α (two sided)	$Z_{1-\alpha/2}$
0.10	1.645
0.05	1.96
0.01	2.576
0.001	3.3

This methodology can be used to answer a question about the data presented in Table 8.2, where the percentages of participants reporting headache during the 6-week study were 6%, 8%, 9%, and 8% for the placebo, low-dose, medium-dose, and high-dose groups respectively. Headaches may be reported fairly often among people with hypertension as a matter of course, but these data suggest that the proportion (expressed here as a percentage) of individuals reporting headache is a bit higher for individuals treated with the active treatment than for those in the placebo group. We can calculate the 95% confidence interval for the proportion of participants in the combined active dose groups reporting a headache. We can also calculate the corresponding confidence interval for the placebo group and compare the two.

The research question

Is the risk (or probability) of experiencing a headache after treatment with the active drug (all doses combined) higher than the risk after treatment with placebo?

Study design

In this study, an investigational antihypertensive drug was evaluated at multiple doses in a parallel-group, placebo-controlled study. An important feature of the design was randomization to treatment, which provides us with unbiased (accurate) estimates of treatment differences. Another feature of the design of the statistical analysis is that we have chosen to compare the rates of one particular AE among many only after seeing the results (that is, *a posteriori*). As we have already seen, any difference between treatments that we may find at this point may be a type I error resulting from the large number of AEs that could have been selected for this particular analysis.

Data

The data for this analysis are the counts of participants treated in each group (that is, the denominator for within-group proportions) and the counts of participants within each group who reported a headache during the study (that is, the numerator for the within-group proportions). As the research question involved all active dose groups combined (that is, any dose of the drug) it is necessary to pool the data across the active dose groups to calculate the confidence interval of interest. Having done that, we now have the following data for our example: 6 out of 98 participants in the placebo group reported a headache, and 25 out of 302 participants in the combined active groups.

Statistical analysis

The statistical analysis approach is to calculate 95% confidence intervals for the proportion of participants in each group (placebo and combined active) reporting a headache. This analysis approach is reasonable because the sample size is sufficiently large (that is, the values, $\hat{p}n$, in each group are at least five). Satisfying this assumption enables us to use the Z distribution for the reliability factor.

The first step is to calculate the point estimate of the proportion. For the placebo group the proportion is 0.06. The second step is to calculate the standard error. For this estimator the standard error is calculated as follows:

$$\sqrt{\frac{(0.06)(0.94)}{98}} = 0.02.$$

The third component of the interval estimate is the reliability factor. As we are calculating a two-sided 95% confidence interval, we select the value of Z from Table 8.3 corresponding to α of 0.05, that is, 1.96.

With all of the components now available, the last step is to calculate the confidence interval. The lower limit is $0.06 - 1.96(0.02) = 0.02$. The upper limit is $0.06 + 1.96(0.02) = 0.10$. We write the 95% confidence interval as (0.02, 0.10). Repeating these steps for the combined active dose group, we obtain a 95% confidence interval of (0.04, 0.12). (We leave it to you to verify this calculation.)

Interpretation and decision-making

Using these two confidence intervals we can now make some conclusions about the unknown population proportion of participants who experience headache after exposure in each group. In the case of the placebo group, we are 95% confident that the population proportion of participants experiencing a headache is enclosed in the interval (0.02, 0.10). For the combined active dose group, we are 95% confident that the population proportion of participants experiencing a headache is enclosed in the interval (0.04, 0.12). Although it may initially have seemed that there may be an increased risk of headache associated with the active treatment, the overlapping within-group confidence intervals suggest that there is insufficient evidence to conclude that the observed difference is real (that is, not due to chance).

8.12 Confidence intervals for the difference between two proportions

There is also another way to answer this research question. If the proportions of individuals reporting headache are the same among participants in the active dose groups and the placebo group, the difference between the two proportions would be 0. Further, because of the influence of sampling error, with which we are now very familiar, we would not necessarily expect the difference to be exactly 0 (just like we do not expect precisely equal numbers of heads and tails in a series of coin tosses). In this approach, therefore, we calculate a confidence interval about the difference in proportions for two independent groups. This interval estimate allows us to exclude implausible values of the difference. This method and others throughout this book require independence of groups (for example, two groups of participants). Examples of groups that are not considered independent are measurements on the same study participant (for example, in ophthalmology left eyes are not considered independent of right eyes in the same individual).

For this method we have sample proportions for independent groups 1 and 2 defined as above:

$$\hat{p}_1 = \frac{\text{number of observations in group 1 with the event of interest}}{\text{total number of observations in group 1 at risk of the event}} \text{ and}$$

$$\hat{p}_2 = \frac{\text{number of observations in group 2 with the event of interest}}{\text{total number of observations in group 2 at risk of the event}}.$$

The estimator for the difference in the two sample proportions is $\hat{p}_1 - \hat{p}_2$ and the standard error of $\hat{p}_1 - \hat{p}_2$ is:

$$\text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},$$

where $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$.

For large samples (that is, when $\hat{p}_1 n_1 > 5$ and $\hat{p}_2 n_2 > 5$) the estimator $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed with mean,

$$p_1 - p_2$$

and variance,

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

So the reliability factor for interval estimates will come from the Z distribution. Then a two-sided $(1-\alpha)$% confidence interval for the difference in sample proportions, $\hat{p}_1 - \hat{p}_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\mathrm{SE}(\hat{p}_1 - \hat{p}_2).$$

While this form of the confidence interval is widely used we suggest the use of a correction factor,

$$\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

attributed to Yates (see Fleiss et al., 2003). This continuity correction factor accounts for the fact that the normal distribution is being used as an approximation to the binomial. With the correction factor, a two-sided $(1-\alpha)$% confidence interval for the difference in sample proportions, $\hat{p}_1 - \hat{p}_2$, is:

$$(\hat{p}_1 - \hat{p}_2) \pm \left(z_{1-\alpha/2}\mathrm{SE}(\hat{p}_1 - \hat{p}_2) + \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

As an example, we look at the headache AE data again and calculate a two-sided confidence interval for the difference in sample proportions.

Data

As above, the data are in the form of counts: 6 out of 98 participants in the placebo group reported a headache and 25 out of 302 participants in the combined active groups reported a headache.

Statistical analysis

As with the previous method, the first step is to calculate the point estimate, but this time the point estimate of the difference in sample proportions. For the placebo group the proportion is 0.06. For the active group the proportion is 0.08. So the point estimate for the difference is $0.06 - 0.08 = -0.02$.

The next step is to calculate the standard error, which is:

$$\sqrt{\frac{(0.06)(0.94)}{98} + \frac{(0.08)(0.92)}{302}} = 0.03.$$

The third component of the interval estimate is the reliability factor. The Z value will be the same as for the previous example (that is, 1.96). For this interval estimate, we also use the continuity correction factor. The continuity correction is calculated as $0.5(1/98 + 1/302) = 0.007$.

We now have all the components of the interval calculation. The lower limit is given as follows:

$$-0.02 - 1.96(0.03) - 0.007 = -0.09.$$

The upper limit is given as follows:

$$-0.02 + 1.96(0.03) - 0.007 = 0.04.$$

Note that the calculated limits do not appear to be equidistant from the point estimate, as we might have expected. This is the result of rounding to two significant digits in the calculations. The calculated 95% confidence interval about the difference in proportion of participants reporting headache as an AE is written as follows:

$$95\%\ \mathrm{CI} = (-0.09, 0.04).$$

Interpretation and decision-making

Given its importance, it is worth restating the interpretation of this confidence interval. We are 95% confident that the true difference in proportions of individuals reporting headache as an AE is within the interval $(-0.09, 0.04)$. As the interval includes 0, there is not enough evidence to suggest that the two groups are statistically significantly different with respect to the risk of headache as an AE. Following this conclusion, we could reasonably continue with further studies in our clinical development program of the active drug, with some assurance from these limited data that the active treatment did not increase the risk of headache.

Suppose, however, that a skeptical colleague insisted that the risk of headache had to be

higher for participants treated with the active drug than with placebo. Using these data, how confident could he or she be that this was really the case? What if all of the headaches in the active treated group were reported in the first week of treatment, whereas in the placebo group the events were spread evenly over the entire 6-week treatment period? Would your view of the relationship between the active treatment and the risk of headache change? A methodology called time-to-event analysis is useful here.

8.13 Time-to-event analysis

An illustration of this scenario is given in Figure 8.1, which shows data from a hypothetical study, study 1 (we discuss another hypothetical scenario, study 2, in due course). There are two treatment groups represented: Active and placebo. Suppose for this example that there are 10 participants in each group, and the length of treatment is 20 days. On the x axis of each panel is time, that is, the number of days since the start of study treatment. Different study participants are represented on the y axis of each panel. Participants numbered 1–10 are in the placebo group and participants numbered 11–20 are in the active group. The occurrence of an AE ("A") is represented with an "X." Completion of the study on day 20 without the AE is denoted by an open circle. The time to either the first report of the AE or the completion of the study is represented by the length of the line from day 1 to the event. Note that it is possible for participants to report more than one instance of the same AE, but only the first occurrence is represented in Figure 8.1.

Here is a descriptive summary of the data displayed in Figure 8.1. For both groups (placebo and active), 5 out of 10 (50%) of the participants reported the particular AE. So, if we were to report these rates and a 95% confidence interval about the difference in proportions, there would not appear to be any difference between these two groups. However, when we look at the times relative to the start of study treatment, this is not so clear any more. In the placebo group, the AE was reported on days 4, 9, 11, 14, and 18. In contrast, the AE was reported much earlier among participants in the active group, on days 1, 2, 4, 5, and 6. The remainder of participants in both groups completed the study on day 20 without experiencing the AE. It appears as if the probability of experiencing the AE (as estimated by the proportion of participants reporting it) is the same between the groups, but that there is a temporal relationship between the start of the study treatment and the time at which the AE is reported. How might we report such a result?

One possibility that might come to mind, although it is not recommended for reasons we discuss shortly, would be to calculate the average number of days to the reported AE. This is problematic, however, because we can calculate such a quantity only for those participants who actually reported the AE. The mean number of days is 11.2 and 3.6 among participants reporting AE "A" in the placebo and active groups, respectively. This analysis completely ignores those who did not report the AE. It hardly seems accurate to exclude these individuals from our analysis. In fact, although half of the participants in both groups did not report "A," they might have eventually reported it if we had followed them longer. Such an estimate of the expected time at which an AE is reported is biased, because not all participants were part of the estimate.

This example suffers from an oversimplification that we have to deal with in the real world, namely that study participants do not always complete the study for the full length of the follow-up period. Participants may drop out of studies for a number of reasons, some of which reflect their experience with the drug (for example, it may be poorly tolerated). Therefore, the "time at risk" differs from individual to individual within the same trial, and it can differ to a considerable degree from trial to trial throughout a clinical development program.

The most important points to remember here are as follows. Simply comparing the relative frequency (that is, the proportion of participants reporting the AE) of the AE between two groups does not tell the whole story: Such an analysis does not address the potential temporal relationship between exposure to the study treatment and the AE of interest. As we saw in this

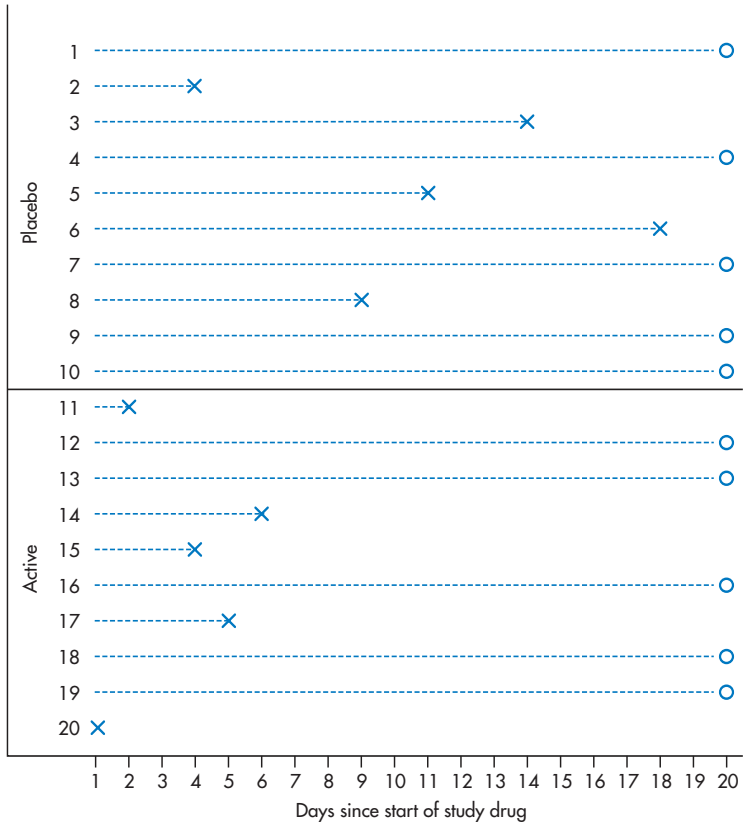

Figure 8.1 Days since the start of study drug at which adverse event "A" was first reported: Study 1 with no dropouts

example, exactly the same proportion of participants in both groups reported the AE. However, the AE occurred in the first 6 days among participants in the active group, whereas the AE reported by participants in the placebo group occurred at evenly spaced intervals over the course of the time at risk. Such a difference in times of the events would suggest that there is a cause-and-effect relationship between the active treatment and the AE.

A more informative approach would be to take into account the time of the event relative to the start of treatment. Ideally, we should use the data from all participants in this approach and should account for varying lengths of time at risk for experiencing the event. O'Neill (1987) advocated such an approach especially for serious AEs caused by the shortcomings of simply describing the incidence (or "crude rate" as he defines it) of AEs:

> For drugs used for chronic exposure, one number or rate such as the crude rate is not likely to be informative without reference to time. To be useful as a summary measure of combined safety data from several studies and which would estimate an overall rate that describes experiences of all participants exposed for varying time periods, there is a need to stratify for time as well as other factors. (O'Neill 1987, p 20)

The next section in this chapter addresses just such a method.

8.14 Kaplan–Meier estimation of the survival function

The analysis method attributed to Kaplan and Meier (1958) enables us to analyze the time to the first reported AE while accounting for different lengths of time at risk. To illustrate this method fully, we have modified the data from the previous example slightly, as shown in Figure 8.2. We refer to this new example as study 2.

The proportion of participants with the event is still equal between the groups (this time 0.6 in both). As seen in Figure 8.2, some participants dropped out of the study before reporting the AE, which are denoted by the open circles at days before day 20. When analyzing data in this way, observations for which the event of interest was not recorded during the time at risk are called censored observations. As noted earlier it is conceivable that, if we had followed these participants for a longer period of time, or if they had not dropped out of the study, they may have experienced the AE of interest.

When analyzing the time to the AE, we need an analytic way to deal with these censored observations. Although we do not know what would have happened for these participants, we do know that they were at risk for some period of time and "survived" their time in the study without experiencing the AE. Accordingly, the main objective of this analysis is to describe how long participants survive without experiencing the event.

The name survival analysis reflects one situation in which this type of analysis is used. When the participants in a clinical trial are very ill, the measurement of efficacy can be the length of time that they live, that is, death is the "event."

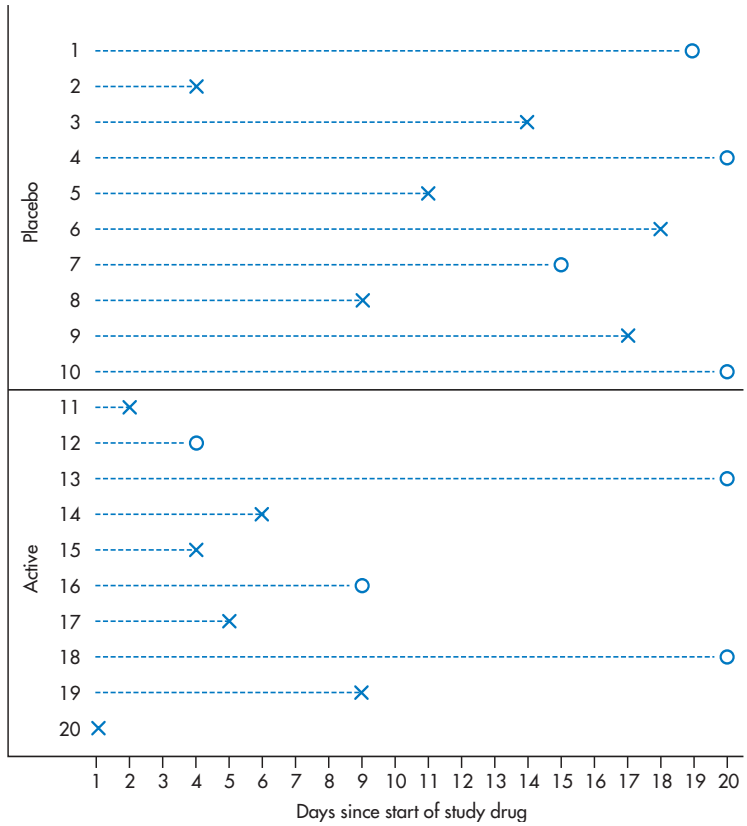

Figure 8.2 Days since the start of study drug at which adverse event "A" was first reported: Study 2 with dropouts

An example may be an oncology trial in which one group receives the investigational drug and the other receives an active control, usually the current gold standard of therapy for the specific type of cancer. Of interest is whether those receiving the investigational treatment survive longer than those receiving the active control. However, survival analysis, as will be seen in our examples, can also be used to measure the time to any defined event.

In this analytic methodology the data for each participant are expressed in a different manner. We present the event times for every participant, defined in one of two ways:

1. The day at which the participant reported the AE
2. The last day the participant was "at risk" for reporting the AE without having done so. This type of participant is labeled parenthetically as "censored."

The data are therefore as follows:

- placebo: 4, 9, 11, 14, 15 (censored), 17, 18, 19 (censored), 20 (censored), 20 (censored)
- active: 1, 2, 4, 4 (censored), 5, 6, 9, 9 (censored), 20 (censored), 20 (censored).

Before discussing the formal definition of this method, it is instructive to think through how we might interpret these data. Let us start with the active group. At the start of the study, all 10 participants are at risk of reporting the AE. Therefore, at day 0 (the day before the start of study treatment), the probability of surviving day 0 without having experienced the AE is 1.00 (we accept this as a given when we define this analysis formally). On day 1, 1 participant out of 10 at risk reported the AE. The probability of an AE on day 1 is 1/10 or 0.10 (that is, 10%). This participant is no longer at risk of reporting the AE later. On day 2 there are nine participants at risk and on this day one more participant reported the event. The probability of an AE on day 2 is 1/9 or 0.11.

This also leaves eight participants at risk on day 3. On day 3 no participant reported the event. Of the eight participants who were at risk on day 4, one reported the AE and one dropped out (that is, was "censored" from the analysis). As before, the probability of an AE occurring is calculated relative to the number at risk, that is, 1/8 or 0.13. On day 5 there are only six participants still at risk. These data are provided in the first five columns of Table 8.4 for the active group, and the same interpretation follows through the end of the 20-day study.

The primary interest in this analysis is not what happens at a single time point, but rather what happens at time t and all points preceding time t. This leads us to the final column of Table 8.4. The numbers in this last column are the estimated probabilities of participants surviving the interval time t without having reported the AE. Given these data, it becomes possible to compare among treatments the probability of a participant not having the event of interest at any given time t.

This method has two desirable characteristics that a simple comparison of proportions does not have. First, it takes into account the variable timing of AEs, which can occur if there is a cause-and-effect relationship of drug to AE. Second, it takes into account the possibility that not all participants will remain at risk for the same amount of time.

Having thought about this methodology in conceptual terms, we now address the necessary calculations for arriving at the data presented in the final column in Table 8.4. This methodology is called the survival function.

8.14.1 The survival function

We introduced you to Bayes' theorem in Chapter 6. According to this theorem, the conditional probability of A given B can be written as:

$$P(A \mid B) = \frac{P(B \mid A)}{P(B)} \times P(A).$$

Or, equivalently, as:

$$P(A) = \frac{P(A \mid B)}{P(B \mid A)} \times P(B).$$

In this methodology we define A as surviving through time t, and B as surviving through time $t - 1$. Then, $P(A|B)$ is the probability of a participant surviving through time t given that he or she has survived through all preceding times

Table 8.4 Event times for the active group in study 2

Time (day), t	Individuals at risk for the AE before time t	Individuals reporting AE at time t	Probability of AE at time t among those at risk	Individuals dropping out at time t	Probability of surviving through time t without AE
0	10	0	0	0	1.00
1	10	1	0.10	0	0.90
2	9	1	0.11	0	0.80
3	8	0	0	0	0.80
4	8	1	0.13	1	0.70
5	6	1	0.17	0	0.58
6	5	1	0.20	0	0.46
7	4	0	0	0	0.46
8	4	0	0	0	0.46
9	4	1	0.25	1	0.35
10	2	0	0	0	0.35
11	2	0	0	0	0.35
12	2	0	0	0	0.35
13	2	0	0	0	0.35
14	2	0	0	0	0.35
15	2	0	0	0	0.35
16	2	0	0	0	0.35
17	2	0	0	0	0.35
18	2	0	0	0	0.35
19	2	0	0	0	0.35
20	2	0	0	2	0.35

$(t-1)$, $(t-2)$, . . . , (1). In addition, $P(B|A)$ is the probability of surviving through $t-1$ given that the participant survived through time t. By definition, that probability is 1.00. Therefore, to calculate the conditional probability of surviving through time t, we need two pieces of information:

1. The probability of surviving through time t given that the participant survived the previous time
2. The probability of surviving the previous interval.

At day 0 (before any participants are at risk), the probability of surviving through time t is 1.00 by definition. On day 1 the probability of surviving through day 1 is the probability of surviving through day 1 given survival through day 0 (that is, 1 minus the probability of the event on day 1 among those at risk), which is equal to $1 - 0.10 = 0.90$ times the probability of surviving through day 0 (1.00). That is, the probability is $0.90 \times 1.00 = 0.90$. Therefore, to calculate the probability in the last column we use the cumulative survival probability (last column) for the previous time and the probability of the event in the interval among those at risk.

Sometimes, these data are presented in a shorter table that displays only those time points at which an individual had an event or was censored, and thus the only values of time for which the probability of survival changes. It is more common, however, to see analyses of this type displayed graphically. The Kaplan–Meier estimate of the survival distribution is displayed for both groups in Figure 8.3. The survival curves displayed in the figure are termed "step functions" because of their appearance. We return to the interpretation of Figure 8.3 after we have fully specified the survival distribution function.

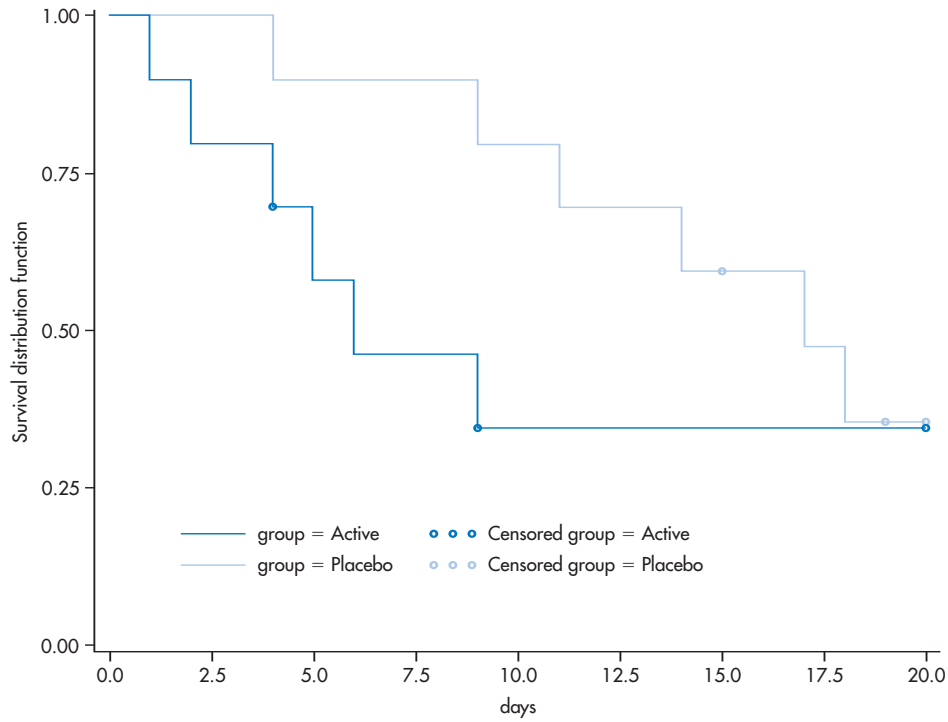

Figure 8.3 Kaplan–Meier estimate of the survival distribution for adverse event A

8.14.2 Kaplan–Meier estimation of a survival distribution

The survival function is the probability that a participant survives (that is, does not experience the event) longer than time t:

$$S(t) = P \text{ (participant survives longer than } t).$$

By definition, a participant cannot experience the event until he or she is at risk of the event, so will survive longer than time 0 or, equivalently, $S(0) = 1$. Also, we accept as a given that, if we waited an infinite amount of time, an individual would eventually experience the event no matter how rare. Therefore, the survival distribution at infinity is defined to be 0, or $S(\infty) = 0$. We also define the following:

- t_i is the unique event time, where $i = 1, 2, \ldots, i$.
- n_i is the number of participants who are at risk just before t_i
- m_i is the number of participants with events at time t_i
- c_i is the number of participants censored in the interval (t_i, t_{i+1}).

The Kaplan–Meier estimate of the survival function at time t is:

$$\hat{S}(t) = \prod_{t_i < t} (1 - \frac{m_t}{n_t}).$$

We can write this series of products out in full as follows:

$$\hat{S}(t) = 1 \times (1 - \frac{m_1}{n_1}) \times (1 - \frac{m_1}{n_1})(1 - \frac{m_2}{n_2}) \times \ldots \times (1 - \frac{m_{t-1}}{n_{t-1}})(1 - \frac{m_t}{n_t}).$$

This expression means that the probability of surviving past time t is the product of the probability of surviving time t conditional upon surviving all preceding time points and the probability of surviving all other preceding time points.

The variance of the survival distribution function at time t is:

$$\text{var}[\hat{S}(t)] = [\hat{S}(t)] \sum_{t(i)<t} \frac{m_t}{n_t(n_t - m_t)}.$$

Consequently, we can take the square root of the variance to obtain the standard error and calculate a $(1 - \alpha)$% confidence interval:

$$\hat{S}(t) \pm Z_{1-\alpha/2} \sqrt{\text{var}[\hat{S}(t)]}.$$

A common measure of central tendency from the Kaplan–Meier estimate is the median survival time (note that this can be estimated only if more than half the participants experience the event). The median survival time is the earliest value of t such that the probability of survival is < 0.5. Note that when observations are censored any estimate of the mean is biased because, technically, the event would eventually occur if we followed participants indefinitely.

We now return to our example to work with some of these expressions. Looking at the last column in Table 8.4 (the estimated survival distribution), we can see that the probability of surviving day 5 is 0.58. Similarly the probability of surviving day 6 is 0.46. Therefore, the estimated median time to an AE in the active group is 6 days, the earliest time at which the probability of survival is < 0.5. For a comparison, the median time to an AE is 16 days in the placebo group. The graphical representation of the survival distribution in Figure 8.3 can also be used to estimate the median time to event.

In Figure 8.3 the survival distribution is plotted against time. As can be seen from the tabular presentation of these estimates in Table 8.4, the survival estimate changes only when there is an event. In the active group on day 1 the estimate is 0.9 and then it drops down to 0.8 on day 2. An important property of the step function defined using discrete event times is that it is a discontinuous function (that is, not defined) between event times. For example, the survival distribution function is 0.46 on days 6, 7 and 8, and then at day 9 the estimate is 0.35. Looking at the Kaplan–Meier curve for the active group you could read day 9 as having an estimate of 0.35 or 0.46, but it is appropriate to remember that the outside edge of the step (right at day 9) is discontinuous, and thus the estimated probability of survival for day 9 or later is 0.35.

Using this guideline we can read off the median survival times by drawing a reference line across Figure 8.3 at $S(t) = 0.50$ and finding the earliest value of time on the curve below the reference line. We leave it to you to verify the median times of 6 and 16 days for the active and placebo groups, respectively, using this method.

The point estimate of the probability of surviving past day 6 is 0.46 for the active group. Using the notation above, we write $\hat{S}(6) = 0.46$. We can now calculate a 95% confidence interval about this estimate. The first step is to calculate the variance about the estimate. Using the expression above and point estimate and the number of events and participants at risk at each time point before day 6, we obtain the following:

$$\text{var}[\hat{S}(6)] = (0.46)^2 \left[\frac{1}{10(9)} + \frac{1}{9(8)} + \frac{1}{8(7)} + \frac{1}{6(5)}\right]$$

$$= (0.46)^2 \left[\frac{1}{90} + \frac{1}{72} + \frac{1}{56} + \frac{1}{30}\right]$$

$$= 0.016.$$

As we have chosen a confidence level of 95%, the corresponding value of Z (the reliability factor) is 1.96. Finally, the 95% confidence interval is calculated as follows:

$$0.46 \pm 1.96 \ (0.016), \text{ i.e., } (0.43, 0.49).$$

That is, we are 95% confident that the true probability of not experiencing the event (surviving) past day 6 is in the interval (0.43, 0.49).

The Kaplan–Meier estimate is a non-parametric method that requires no distributional assumptions. The only assumption required is that the observations are independent. In the case of this example, the observations are event times (or censoring times) for each individual. Observations on unique study participants can be considered independent. The confidence interval approach described here is consistent with the stated preference for estimation and description of risks associated with new treatments. A method for testing the equality of survival distributions is discussed in Chapter 11.

8.14.3 Cox's proportional hazards model

Although we do not cover them in detail, there are parametric methods to analyze time to event data of this type, the most notable of which is Cox's proportional hazards model.

A hazard can be thought of as the risk of the event in a small interval of time, given survival up to the start of the short interval. Parametric approaches to time-to-event data such as Cox's model have a number of advantages, including the ability to adjust for other explanatory effects in a model and to extend them to recurring events for a single individual. In this case, event times would not be independent because within-participant event times would be correlated. Such an approach is appealing statistically because it makes use of more data. However, the main disadvantage of Cox's model is that the single parameter of the model, the ratio of the hazards of two groups, is assumed to be constant over time. The risk of an AE for participants treated with an active drug could vary in a nonconstant manner over time relative to the risk for placebo-treated participants, making such an assumption tenuous.

8.14.4 Considerations for the use of Kaplan–Meier estimation for AEs

We suggest the use of the Kaplan–Meier estimate for a better understanding of the risk of AEs in clinical trials for two reasons. First, the proportional hazards model has important assumptions which must be made. Secondly, the Kaplan–Meier method is easier to implement and interpret. The analysis of AEs using the Kaplan–Meier method allows us to account for the different lengths of time at risk without making any significant assumptions about the shape of the underlying distributions of the survival or hazard functions. Reviewing the rather exaggerated data from Figure 8.3 it may seem obvious that ignoring the time at risk could be problematic. Employing an appropriate method of analysis (for example, properly accounting for all individuals and calculating an interval estimate for the proportion) does not necessarily mean that the analysis is the most appropriate one. Consideration should be given to the varying lengths of follow-up or "time at risk" when reporting AEs. It is wise to consider the denominator carefully when making any statement about probabilities.

A final word of caution here is that, although the Kaplan–Meier method (and other methods for time-to-event data) appropriately accounts for the time at risk of an event within a group, if the pattern of censoring is dependent on the treatment (for example, suppose the dropout rate is dose dependent as might be seen with chemotherapy), any treatment group comparisons of the estimate of the risk of AEs would be potentially biased. Thus, a more complete analysis would include first an assessment of censoring times (visually at a minimum) and the reason for drop out, and then the appropriate analysis to account for the time at risk. Failing to quantify the probability of an AE accurately during drug development can have significant implications for sponsors, regulatory authorities, prescribing clinicians, and patients.

8.15 Review

1. What measures are taken to ensure that AE data are of a high quality?
2. Refer to Table 8.2. Calculate a two-sided 99% confidence interval for the proportion of participants reporting any event in the:

   (a) placebo group
   (b) active dose groups combined.
3. In a therapeutic exploratory trial, 22 participants out of 140 reported an AE:

   (a) What is the 95% confidence interval for the sample proportion of participants reporting an AE?
   (b) What is the 99% confidence interval for the sample proportion of participants reporting an AE?
   (c) How confident would you be that the true population proportion of participants reporting an AE does not exceed 0.18?

4. A total of 290 participants were studied in the first therapeutic exploratory trial of an investigational antihypertensive drug. Of the 150 individuals treated with the test treatment, 32 reported fatigue. Of the 140 treated with placebo, 19 reported fatigue:
   (a) Calculate a 90% confidence interval for the difference in proportions of participants reporting fatigue.
   (b) Calculate a 95% confidence interval for the difference in proportions of participants reporting fatigue.
   (c) Calculate a 99% confidence interval for the difference in proportions of participants reporting fatigue.
   (d) What is the statistical interpretation of these results?
   (e) How might these results influence the course of future development of the drug?
5. Why is it important to account for the time that individuals are at risk of an AE?
6. Describe in your own words what a survival function is.

8.16 References

Chow S-C, Liu J-P (2004). *Design and Analysis of Clinical Trials: Concepts and methodologies*. Chichester: John Wiley & Sons.

EMEA Committee for Proprietary Medicinal Products (CPMP) (2002). *Points to Consider on Multiplicity Issues in Clinical Trials*. London: EMEA.

Fleiss JL, Paik MC, Levin B (2003). *Statistical Methods for Rates and Proportions*, 3rd edn. Chichester: John Wiley & Sons.

ICH Guidance E2A (1995). *Clinical Safety Data Management: Definitions and Standards for Expedited Reporting*. Available at: www.ich.org (accessed July 1 2007).

ICH Guidance E6 (R1) (1996). *Good Clinical Practice*. Available at: www.ich.org (accessed July 1 2007).

Kaplan EL, Meier P (1958). Nonparametric estimation from incomplete observations. *J Am Statist Assn* **53:**457–481.

Mann R, Andrews E, eds (2007). *Pharmacovigilance*, 2nd edn. Chichester: John Wiley & Sons.

O'Neill RT (1987). Statistical analyses of adverse event data from clinical trials: special emphasis on serious events. *Drug Information J* **21**:9–20.

US Food and Drug Administration (2005). *Conducting a Clinical Safety Review of a New Product Application and Preparing a Report on the Review*. Available from www.fda.gov (accessed July 1 2007).