CHAPTER 2

Clinical Trial Evaluation and Biostatistics Primer

Gary Sloskey, PharmD, BCPS

KEY TERMS

blinding
case-control study
case report
categorical data
clinical endpoint
cohort study
composite endpoint
confidence interval
confounder
continuous data
covariate
cross-sectional study
crossover trial
dependent variable
independent variable
intent-to-treat (ITT) population
last observation carried forward (LOCF) method
nominal data
noninferiority trial
nonparametric data
ordinal data
p-value
parallel design
parametric data
per-protocol population
power
pre-post trial
randomized controlled trial (RCT)
standard deviation (SD)
stratification
study objective
surrogate endpoint
type I error
type II error

LEARNING OBJECTIVES

After reading this chapter, the reader should be able to:

1. Identify appropriate descriptive and summary statistics for a given data type.
2. Categorize data according to type (i.e., continuous, parametric vs. nonparametric, categorical, and ordinal).
3. Identify outcomes from a clinical trial and differentiate between surrogate and clinical outcomes.
4. Describe the values generated from the statistical evaluation of clinical trial results and explain what they mean.
5. Explain what *p*-values are, how they are interpreted, and what they do not mean.
6. Recognize the appropriate statistical tests to interpret the data resulting from clinical trials.
7. Differentiate between clinical and statistical differences with regard to the results of comparative study endpoints.
8. Interpret 95% confidence intervals for absolute data and ratio data.
9. Differentiate between types of clinical trial designs (e.g., parallel, crossover, and pre-post), and describe how their statistical assessments differ.
10. Differentiate between types of observational trial designs (e.g., cohort and case-control), and identify the strengths and limitations of each.
11. Differentiate between and recognize study methods influencing the internal and external validity of a study.
12. Define *type I* and *type II error*, and describe the methods used to estimate the likelihood of error (e.g., power, alpha, beta).
13. Analyze a dataset by selecting the appropriate statistical tests (e.g., chi-square, *t*-test, ANOVA, ANCOVA).
14. Critically evaluate the primary literature with regard to the qualitative assessment of an experiment's study design, methods, results, and validity and the applicability of a given conclusion.

INTRODUCTION

As a young pharmacy student, I served my internship in a moderately sized and relatively progressive hospital in the suburbs of Philadelphia. It was there that I had the pleasure of working alongside one of the most pleasant, experienced, and elderly pharmacists that I had known at that stage of my life. One day, while our pharmacists were discussing therapeutic options for a patient who was showing resistance to diuretic therapy, a low and raspy voice from the background kept repeating, "watermelon seeds." Our noted octogenarian pharmacist went on to recommend a "time-honored" practice of administering an infusion of watermelon seeds as a diuretic.

I bring up this story to illustrate our need to approach the therapy of our patients with an eye toward evidence-based medicine (EBM). This case is a little unique; however, recent history has shown that many therapies that the medical community has relied upon for years have now been shown in controlled trials and meta-analyses to be less than effective and sometimes less than safe. The heart of this chapter deals with the importance of EBM and the skills needed to identify and use it. To summarize with a quote, "EBM is the integration of the best available evidence with our clinical expertise and our patient's unique values and circumstances."[1] Pharmacists must have the skills to appropriately evaluate and understand information obtained from the primary literature. This charge requires a working understanding of study designs, research endpoints, and biostatistical evaluation. This chapter will therefore provide an overview of the components of published clinical trials, focusing on research methods, experimental designs, and study results. Of course, no chapter dealing with clinical trial evaluation would be worthwhile without an attempt to clearly explain the significance of study results at a biostatistical level in order to aid students in their understanding of these concepts.

STUDY DESIGN

The more common experimental designs and design components are presented in **Table 2-1**. The table is divided into three sections that correspond to the major types of study designs:

1. Experimental studies: The investigator actively provides an intervention that is then evaluated for its associated outcomes.
2. Observational studies: The investigator does not provide an active intervention but simply prospectively or retrospectively observes relationships between exposures (potential causes) and outcomes (potential effects).
3. Descriptive studies: Clinical cases, survey data, or epidemiologic data are described.

Randomized controlled trials (RCTs) are experimental studies and are considered to be at the top of the experimental research design hierarchy in the healthcare research and medical literature, especially for the purpose of proving hypotheses. In a traditional, parallel design RCT, each patient is randomly assigned to one of two or more independent study groups (study arms) and receives a unique active intervention or a placebo (inactive) intervention. Do not confuse *random* with *arbitrary*. Several acceptable procedures for true randomization are available, including the use of random number tables. An RCT is, by definition, controlled or, more specifically, placebo-controlled. The outcomes of active experimental interventions are compared to those of a placebo, or sham, intervention for a very important reason.

Consider the following: An investigator studied and compared the efficacy of two different antiemetic regimens for the prevention of postoperative nausea and vomiting (PONV). The study recruited and randomized patients scheduled for surgical procedures under various forms of anesthesia. The results of the study showed no difference in efficacy between the two regimens. Few patients receiving either regimen experienced significant PONV. Many hospital pharmacy directors were quite pleased with the results of this trial. One of the "equally efficacious" regimens was associated with an impressively low cost compared to the standard therapy and represented the potential for hospital savings approaching a million dollars a year. When something seems too good to be true, it probably is. Because this study was not controlled (did not include a placebo group), critics doubted the validity of the study and suggested the very real possibility that the risk for PONV without prophylaxis was very low for all of this study's postoperative subjects. If a placebo control group had been studied and a significantly higher incidence of PONV had been associated with the placebo compared to the active regimens, then this trial may have been valid. As designed, the results were not valid.

Table 2-1 Study Designs in the Primary Literature

Type of Study	Common Features
Experimental Studies	
Randomized controlled trials (RCTs) vs. clinical trials (CTs)	Comparison of different interventions provided to two or more independent subject groups (arms). Subjects are randomly assigned to a study group. One of the study groups receives a sham, or placebo, intervention as a control. Some clinical trials are not randomized and/or may not be placebo-controlled. In fact, some trials are not comparative at all. Although these are not at the top of the trial hierarchy, they are valid and often are important.
Superiority vs. noninferiority	Although the majority of studies traditionally have been designed to determine if one intervention is superior to another active intervention (or placebo), noninferiority studies are becoming more common and are designed to determine if one active intervention is *not* inferior to another active intervention.
Parallel studies	A typical parallel study evaluates the outcomes of different interventions provided to different (independent) subject groups over the same prospective period of time or "in parallel" (e.g., while some subjects are being studied on intervention A, other subjects are studied on intervention B). Crossover and pre-post studies are examples of nonparallel studies.
Crossover trials	All of the enrolled subjects receive, in a sequential manner, all of the interventions being compared in the study. The outcomes of the interventions are compared even though the patients receiving these interventions are in common. The order of the interventions with which subjects are studied is usually randomized, and washout periods between different interventions are usually included.
Pre-post trials	Study subjects receive a therapeutic intervention and their baseline (preintervention) condition is compared to their intervention-treated condition.
Observational Studies	
Cohort studies	Usually, the researcher prospectively observes but does not intervene. A cohort study follows a sample of subjects with a common exposure to determine if this exposure is associated with an outcome to a greater degree than a cohort of subjects without that exposure. *A cohort study starts with a known exposure and looks for an effect or outcome.*
Case-control studies	The researcher retrospectively observes but does not intervene. A case-control study looks at the history of a sample of subjects with a common outcome or condition to determine if this condition is associated with an exposure more frequently than what is seen in a sample of patients without the condition. *A case-control study starts with a known outcome and looks for a possible cause.*
Descriptive Studies	
Cases and case-series reports	Describe or report an interesting clinical case or a collection of similar cases. Although case reports are not valuable for hypothesis validation, they sometimes are very valuable for hypothesis generation.
Cross-sectional studies	Describes a sample, or *cross-section*, of a population at a point in time. Also called *survey studies* or *epidemiologic studies*. Do not confuse with crossover trials.

Clinical trials of medication therapies are frequently designed to show a medication's superiority over another medication or a placebo treatment. **Noninferiority trials**, however, are becoming more and more common in the medical literature.[2] Noninferiority studies have some unique features. What constitutes "inferiority" is prospectively defined by the investigator in each noninferiority study. For example, for antibiotic A to be considered inferior to antibiotic B, an investigator would have to determine

that antibiotic A fails to provide an anti-infective response at least X% more often than antibiotic B (e.g., a –10% threshold for inferiority). The size of the threshold varies between studies and is based on statistical, clinical, and regulatory considerations. A concern that I hear occasionally from students preparing for journal club or seminar is that the noninferiority study that they are evaluating does not include a biostatistical test. When comparing the responses of two interventions, noninferiority trial investigators calculate a 95% confidence interval (CI) of the mean difference between the two therapies. The threshold for inferiority is *not* compared to the mean difference in response rate between the two drugs; it is compared to the lower (or sometimes upper) bound of the mean response difference's 95% CI. For example, a study showed that antibiotic A resulted in a 75% clinical cure rate while antibiotic B resulted in an 80% cure rate. In this instance, the mean difference in cure rate and its 95% CI was reported as: mean = –5%; 95% CI (–8% to +2%) for antibiotic A. Therefore, antibiotic A is noninferior to antibiotic B because the –8% in the lower bound of the 95% CI was not worse than the threshold for inferiority pre-specified above as –10%.

The majority of studies in the primary literature have a **parallel design**. Between the time of active recruitment and completion of the clinical trial, patients are continually randomized into independent groups and studied in a prospective fashion. At any given time during the study, some patients will be receiving one intervention while others are receiving the comparative intervention.

Nonparallel studies include crossover trials and pre-post trials. In a **crossover trial**, all of the patients enrolled in the study will sequentially receive all of the interventions to be compared. Appropriate washout periods are usually included between different interventions. Randomization is used in crossover studies only to determine the sequence of interventions for each patient. Similar to parallel studies, the outcomes of the different interventions are evaluated and compared, but all of the interventions are evaluated in the same subjects. In essence, the potential for the dissimilarity of subjects between interventions affecting the results is no longer a concern. In other words, in crossover trials, the outcomes of the interventions are not influenced by between-group patient variability. Because of this, crossover studies require fewer subjects. Crossover studies require more stringent "paired" statistical methods for the evaluation of their results. If an outcome of a crossover study is evaluated using a standard statistical test, such as the Student's *t*-test, instead of the more stringent paired *t*-test, there is a risk of overestimating the result's statistical significance.

Pre-post trials do not compare the difference between interventions but simply study the change in subjects' baseline (prestudy) parameters or condition due to the intervention being investigated. Because these patients serve as their own control, similar to the crossover trial, "paired" statistical methods are used for the evaluation of these results. Some words of caution: You may occasionally come across a trial that appears to be a comparative parallel study, but it only compares baseline parameters (pre) to on-therapy endpoints (post) for each intervention.[3] Don't be fooled. The clue is that these studies will report a separate *p*-value for each intervention result.

In observational studies, including cohort studies and case-control studies, the investigator does not actively intervene or prescribe therapies. **Cohort studies** are usually prospective in design. The investigator identifies subjects having a common exposure and follows them to uncover a potentially associated outcome. An example of a cohort study is one that identified a large sample of female subjects (nurses) who reported the routine use of aspirin. The investigators then followed these patients over an extended period of time and compared their cardiovascular outcomes to a cohort, or sample, of similar subjects who were not aspirin users.[4]

Case-control studies, in contrast to cohort studies, identify patients with a common outcome or condition and then retrospectively evaluate past exposures as a potential cause of the condition. An example is a case-control study that identified a group of patients with a diagnosis of pulmonary hypertension. The investigators then reviewed the patients' medical histories and found that the use of stimulant anorexic weight-control agents was more common than in control cases of patients without pulmonary hypertension.[5]

In the two observational study examples discussed, note that other unknown factors may have significantly influenced the outcomes.

CLINICAL PEARL

Because observational studies generally lack a control, they are not reliable for answering questions definitively or for proving existing hypotheses or proving causal relationships. However, they can be quite valuable for generating new hypotheses and areas for further study.

Descriptive studies include case and case series reports as well as cross-sectional studies. **Case reports** and case series reports simply describe and discuss individual clinical cases or a series of cases. A case report or even a series of cases reporting that a particular drug treatment was successful does not offer strong or reliable evidence of the efficacy of that treatment. Although this type of publication is ranked rather low on the hierarchy list of primary literature, the importance of case reports is still significant. In fact, the medical community's awareness of HIV and AIDS was first brought to light by case and case series reports describing patients with unusual tumors and immunologic deficiencies.[6]

Cross-sectional studies are, essentially, snapshots of data collected from a cross-section, or sample in time, of a population. They include survey studies and epidemiological research. Although this type of research seems simplistic on the surface, epidemiologists apply scientific principles and a study design in order to obtain meaningful and valid results. A commonly cited example of a poorly constructed survey recounts the telephone-based survey of potential voters prior to the 1948 U.S. presidential election between Thomas Dewey and Harry Truman.[7] This survey, and other similar ones, predicted that Dewey, the Republican, would be the likely winner of the election. However, Truman won the election. The fact that three times as many Republicans owned telephones compared to their Democratic counterparts may have been overlooked and may have introduced a serious sampling bias, leading to the faulty study.

Note that there are many look-a-like, sound-a-like terms in the literature. Do not confuse the terms *cross-sectional* and *cross-over*. These terms are great fodder for trick exam questions.

COMPONENTS OF A PRIMARY LITERATURE PUBLICATION

Table 2-2 lists the eight common components of published clinical trials. Experienced practitioners are generally more effective and more efficient at reading and comprehending the primary literature

Table 2-2 Components of Published Clinical Trials

Component	Comments and Considerations
Title	Avoid making clinical conclusions and decisions based on a title of a trial even when the title reads something like, "Drug X is highly effective for the treatment of Y disease."
Authors	The list of authors can provide a wealth of information regarding the investigators' expertise and authority and the potential quality of the study. Likewise, it can suggest the presence or absence of potential study bias or conflicts of interest.
Abstract	Avoid making clinical conclusions from a study abstract. Abstracts do not adequately describe potential experimental weaknesses. They are provided to give the reader an idea of content relevance and a reason to critically read the study.
Introduction	A valuable section of the publication that provides background information, the study's general objectives, and, often, the rational for the study's specific uniqueness or importance.
Methods and results	This chapter will discuss and put into perspective a number of types of study methods and endpoints and how they influence the applicability of the study's findings and the type of statistical assessment most appropriate for the study results. See Tables 2-3 and 2-4 for detailed information on the methods and results sections of a publication.
Discussion/conclusion	Avoid making clinical decisions based solely on a study's discussion or conclusion section. The discussion/conclusion is a valuable source of insight from the investigators and commonly points out the strengths and weaknesses of the investigation and the importance of the results. However, this section is occasionally a place where investigator bias emerges.
References	In addition to the obvious purposes of a reference list following any publication, references in a clinical trial can provide the reader with a rich source of information detailing background information about disease and therapy. Additionally, investigators occasionally reference standards or previously used methodologies rather than detailing those methods used in their investigation.

because they know what to look for and where to find it in published studies. In this section, we will lightly touch on the most important characteristics of some of the components of clinical trials. Later in the chapter we will examine experimental methods and results in more depth.

One important recommendation with regard to the title, abstract, and conclusion sections of the clinical trial is that the reader should *never* base a clinical decision based solely on the information contained in these sections. The purpose of the abstract is to give the reader sufficient information to decide whether reading the entire study is appropriate to the reader's question.[8] When a publication's title, abstract, or conclusion suggests efficacy or superiority of a drug therapy, the reader must fully assess the study's internal and external validity prior to applying its conclusion to a specific patient or group of patients. Readers of a published investigation must ask themselves the following four questions before applying the findings of a clinical trial:

Internal validity:

1. Are the study design and methods appropriate?
2. Are the study results interpreted correctly?

External validity:

3. Are the study's subjects similar to my patients?
4. Is the disease treated in this study the same disease that I will be treating?

Study results are not always accurately reflected by the investigators in the discussion/conclusion section in the last pages or paragraphs of a clinical trial. This is where biases often show themselves. The above questions can only be answered through a thorough examination of the study's design and outcome assessment.

RESEARCH METHODS

Table 2-3 provides an overview of the most important and most common experimental methods utilized in clinical trials. Inclusion and exclusion criteria serve as the primary basis for one's judgment of a trial's external validity. These criteria help to ensure that the enrolled subjects represent the appropriate disease and prognosis, are able to adhere to the protocol, tolerate the study's interventions, and be accurately assessed on the study's outcomes. An investigator whose specialty practice focuses on women athletes conducts a large, well-designed RCT, which primarily enrolls middle-aged females with a diagnosis of osteoarthritis, to study a newly developed medication for arthritis. Is this a poor study due to its narrow external validity? No, it's not at all poor. However, it would certainly be inappropriate to apply the results of this study to young, male Veteran's Administration patients with rheumatoid arthritis. The above issues relate to the external validity of the study relative to the clinician's needs.

Randomization is discussed in several sections of this chapter and, for now, is adequately defined in Table 2-3. **Stratification** is a method for classifying subjects into secondary groups, or *strata*. Stratification is used to prospectively identify potential subject characteristics that may influence outcomes so that the effects of these "nuisance" characteristics can later be assessed. We will discuss these "nuisance" variables throughout this chapter and will frequently refer to them as *confounders* or *covariates*. Let's discuss an example: A multicenter study's primary objective might be to investigate the superiority of one antidepressant medication over another. The investigators also want to further evaluate differences in efficacy between males and females. In this case, subjects are first stratified as male or female and then randomly assigned to treatment with antidepressant A or antidepressant B. Now, the study will not only assess the overall difference in efficacy between drugs, it also can assess how gender affects the outcome of either drug therapy. But wait, that's not all! In multicenter trials, center stratification is usually identified so the investigators will be able to detect unexpected outlier outcomes from one or more centers. In fact, in a multicenter clinical trial, a commonly used stratification, or confounder, is the site or center at which patients are treated and evaluated.

The use of blinding and placebos in clinical trials are related techniques that reduce the potential for bias and add experimental control, respectively. **Blinding** is used to prevent the patient and/or the investigator from knowing which intervention a patient is receiving—active or placebo. A patient or an investigator with a preconceived (and sometimes subconscious) notion that one intervention is superior to another may falsely perceive or interpret an outcome in a biased fashion. Blinding takes on a more important role when an endpoint is subjective, such as pain, nausea, or various "global impressions of improvement."

Table 2-3 Experimental Methods

Method	Description
Inclusion/exclusion criteria	Inclusion and exclusion criteria directly impact the external validity of the study. *Common inclusion criteria:* • Subjects with signs, symptoms, or documentation of diagnosis of the relevant disease or condition being studied. • Subjects who are available and able to adhere to the study protocol. *Common exclusion criteria:* • Subjects at undue risk, such as subjects with a history of allergy to the study drug and subjects who might not tolerate the intervention. • Subjects with comorbidities potentially complicating the responsiveness of the disease or the investigator's ability to assess the treatment response.
Randomization and stratification	*Randomization* is a structured procedure for assigning subjects in a balanced and random fashion to the different interventions the study is comparing. *Stratification* is frequently used to prospectively identify subject characteristics that might influence the response to the intervention.
Blinding	To prevent inadvertent bias, investigators and subjects (double blinding) or subjects (single blinding) often are kept from knowing what intervention a subject is receiving.
Power and power analysis	A study must enroll and evaluate the outcomes of a sufficient number of subjects to ensure study validity, specifically when the results of the study indicate that no statistically significant difference could be found between the different interventions. An inadequately powered study might not be able to show superiority of an intervention when a superiority does, in fact, exist. The number of subjects needed for a study should not be confused with the *number needed to treat* (NNT). NNT is a pharmacoeconomic term unrelated to study requirements or study power.
Placebo control	A placebo is an inert dosage form or a sham treatment designed to look like an active therapy. Placebos are necessary for blinding and for bias control. Blinding sometimes involves a more complex use of placebos. An example would be when an investigator is comparing an intravenously administered medication to an orally administered medication. In this case, patients might receive an active oral medication and a placebo intravenous medication, or vice-versa. This is termed a *double-dummy*.
Intent to treat	When a patient is randomized and assigned to a treatment group, that patient's outcome data is included in the results of the treatment group, regardless of whether the patient adheres with the medication plan, whether the patient shows up for all of the assessment visits or misses several, or even drops out of the study.
Per-protocol	When a patient is randomized and assigned to a treatment group, that patient's outcome data might be discarded if the patient misses key assessments or drops out of the study.
Data interpolation, LOCF	When an intent-to-treat method is used and a patient misses one or more of a series of response evaluations, investigators are faced with the challenge of how to deal with the missing data. Many different approaches are used, including just leaving the missing data out. Another approach would be to reuse the same data from the patient's last evaluation. This is referred to as *last observation carried forward (LOCF)*.

A placebo effect is a common component of many therapeutic outcomes. A patient being treated for pain may respond well to a investigational analgesic, but some portion of this favorable outcome often is due to the caring and empathic attention that the caregiver provides. Symptoms of many conditions wax and wane over time and some conditions spontaneously resolve on their own. A study group receiving a nonpharmacologically active placebo and the same caring, empathy, attention, and patterns of waxing, waning, and spontaneous remission serves as a control for these as well as other phenomena.

How many patients are needed in a clinical trial? This issue presents itself most often when the results of a trial have failed to show a statistical difference in effect between interventions. Does this failure to show a difference mean that there is no difference, or does it mean that the study was not powerful enough to statistically demonstrate a real difference? When a study concludes that there is no difference between two interventions, a type II error is being committed if there is, in fact, a difference. These issues are related to the study's **power**. When a study is being designed, an investigator will frequently conduct a power analysis. Students are most familiar with a study's *n,* which is the number of subjects studied, as determining a study's power. Note that other factors affect power, sometimes to a greater degree than just the *n.* These factors include, but are not limited to, the investigator's estimate of the variability of the results that will be found, the smallness of the difference that the study wishes to detect, and the planned alpha (α) value of the study (see the alpha in Table 2-7).

Consider the following example of power: An investigator calculates that a minimum of 100 patients must be enrolled, studied, and assessed for a study power of 80%. A power of 80% indicates that if no statistically significant difference is found between interventions for the primary endpoint, then there is only a 20% chance that a difference really does exist but that the study did not have a sufficient power to detect this difference. In other words, this study has a beta (β) of 0.2, or a 20% chance of a type II error. A commonly accepted power for studies is 80%; however, many studies are designed to attain a 90% or better power.

When a study has been conducted and the results are being evaluated, how does one handle data from patients who have either dropped out of the study or who have not adhered to the study's requirements? This is a decision that must be made prospectively prior to the initiation of the study. Data evaluation based on an **intent-to-treat (ITT) population** requires that the data on all of the patients who are randomized and initiated on a particular intervention will be included in the evaluation of that intervention, regardless of whether the patients adhered to the protocol or whether they dropped out of the study prematurely. In other words, patients intended to be evaluated within an intervention are evaluated within that intervention even if they have not adhered to the intervention. Hollis and Campbell[9] provide an explanation of ITT and examine a survey of published literature. The ways in which ITT methods are described and applied and the ways in which missing data is handled vary considerably between studies. The reader should not merely take the claim of ITT use at face value but must critically evaluate the application of ITT in each study.

When performing a study, it is realistic to expect that some patients will miss one or more visits in a series of assessments. When this occurs, how does the investigator deal with missing pieces of data? Many methods of data interpolation are available. With the **last observation carried forward (LOCF) method**, the data from the patient's last evaluation is reused to populate the current missing data item. The assumption in this case is that the patient's parameters did not change since the last time he or she was evaluated. Alternative approaches include interpolating with other estimation methods, leaving the data blank, or assuming the worst evaluation.

An alternative to ITT, study data are evaluated based on a **per-protocol population**. With this method, data from patients who have missed key assessments or who have dropped out of the study prematurely are not included in the final evaluation of the intervention's efficacy.

Which of the two methods—ITT or per-protocol—is better? The answer is that both methods have advantages as well as disadvantages. The ITT methodology may underestimate treatment efficacy and the per-protocol methodology may underestimate treatment failures. Additionally, when patients are dropped from the evaluation under a per-protocol method, the lower number of patients might jeopardize the anticipated power of a study. Let's face it. It would not be appropriate to consider only the outcomes of patients completing a trial when many patients may have dropped out of the trial because of death, lack of relief, or intolerable adverse effects.

EVALUATION OF RESEARCH RESULTS

Although not always appreciated, a study's demographics table appears in the results section of the paper. This table is the result of the study's inclusion and exclusion criteria and the randomization procedure. It is important to review this table for both potential internal and external validity issues.

A well-designed and executed process of randomization should result in an equal distribution of patient characteristics between all of the study arms; however, the reader should confirm this. Characteristics such as gender, race, nationality, and so on are nominal data. Characteristics such as mean age, duration of illness, and mean blood pressure are continuous data. Statistical tests often are, but not always, used to assess the "likeness" of characteristics between study arms.

The most important subject characteristics will depend on the study's objectives and endpoints. If a trial is designed to investigate the efficacy of an antibiotic, it might not do well to have an imbalance of patients infected with resistant organisms between study arms. These are issues of internal validity, or the overall reliability of the study itself and its conclusions. Regarding external validity, the reader must understand the patient types and disease characteristics being studied to determine if the results and conclusions of the study are applicable to his or her patients.

To appreciate the meaning of the research results, it is imperative to be able to distinguish between the *study objectives* and the *endpoints*. Further, it is important to appreciate the differences between different types of endpoints, including clinical endpoints, surrogate endpoints, and composite endpoints. **Table 2-4** includes a brief description of some of these terms.

Table 2-4 Study Results: Types of Outcomes and Endpoints

Type of Outcome or Endpoint	Description
Study objectives	A relatively nonspecific description of what a study is designed to find out. *Example:* "The objective of this study is to investigate the efficacy, safety, and tolerability of supercillin compared to placebo." The primary objective here might be the drug's efficacy, but this statement does not specifically define efficacy, nor does it state how efficacy will be described.
Endpoints	Endpoints are very specific and tell the reader precisely what the measure is. For example, for a lipid-lowering medication, the endpoint might be the mean reduction in LDL cholesterol attained with therapy, the proportion of patients who attain a normal LDL cholesterol blood level while on the drug, or the proportion of patients who do not have a cardiovascular event after long-term therapy with the investigated agent. Endpoints also are designated as primary or secondary. In most cases, a study has one primary endpoint and multiple secondary endpoints. Occasionally, a study will designate two endpoints as being co-primary. The primary endpoint is usually the most important and is the endpoint that the study's power analysis is usually based upon.
Clinical endpoints	In the above examples of endpoints, "the proportion of patients who do not have a cardiovascular event" is a clinical endpoint. As another example, a clinical endpoint might be the reduction in pain as measured by a visual analog scale, cure rate, or duration of remission. These endpoints have a common thread in that they measure a direct and clinical outcome, which represents a "bottom-line" purpose for the therapy.
Surrogate endpoints	Oftentimes the measurement of clinical endpoints is difficult or requires a long follow-up period. In the example above, investigators might have to follow their patients on an investigational lipid-lowering drug for many years before they could determine if the drug prevents subjects from having heart attacks. In such cases, surrogate endpoints are used instead. In this case, reduction or normalization of serum LDL cholesterol levels is a surrogate marker. Some surrogate markers have a better correlation to a clinical endpoint than others.
Composite endpoints	Oftentimes an endpoint used in a clinical trial is actually a combination of several endpoints. For example, consider the following endpoint: "Patients having a myocardial infarction, being admitted to a hospital for a cardiovascular event, or dying from a cardiovascular cause." In this case, a patient having one or more of these endpoint components is counted as one endpoint. Endpoints combined into a composite endpoint should be of similar severity and importance.

A clinical trial will describe its objectives within the paper's introduction. **Study objectives** are relatively nonspecific descriptions of the aims of the research; for example, "The main objective of this trial is to study the efficacy of drug X in the treatment of disease Y." Endpoints, in contrast, should be very specific and should clearly denote the detail of the measurable result; for example, "The primary endpoint of this study is to determine the proportion of patients treated with drug X who attain adequate pain control as defined by an 80% or greater reduction in their visual analog pain scale scores."

The majority of clinical trials will identify a single primary endpoint. Occasionally a study will identify a dual primary endpoint. The primary endpoint is the most important endpoint of the trial and is usually the endpoint that is included in a study's power analysis. Therefore, the predetermined estimate of a type II error risk in studies having a power analysis applies only to the primary endpoint. Clinical trials usually have multiple secondary endpoints. Officially, there is no such thing as a tertiary point, but you may occasionally see one identified in a clinical study. It's a reasonably good bet that the investigator of such a study did not read this chapter.

Some endpoints are termed *clinical* and others *surrogate*. **Clinical endpoints** directly measure outcomes related to a patient's symptoms, morbidity, and mortality. Examples include cure rates, symptom improvement, speed of response, duration of remission, death rates, and so on. **Surrogate endpoints** are indirect measures of a clinical outcome and are generally easier and quicker to assess.[10] Surrogate endpoints include laboratory values, physiologic changes, and other values that relate, to varying degrees, to an actual desired clinical outcome for a patient. For example, when patients are treated for hyperlipidemia, a true clinical goal is the avoidance of long-term cardiovascular events. However, measuring this outcome requires lengthy trials and a large number of study subjects. A surrogate endpoint, such as the reduction of serum low-density lipoprotein (LDL) levels, requires smaller and less lengthy trials. In this case, reduction of LDL levels in patients is considered to be a reasonably good surrogate that corresponds reasonably well with clinical long-term cardiovascular risk.

In the above clinical endpoint example, cardiovascular events may include acute coronary events, acute cerebrovascular events, hospitalization for a cardiovascular event, death from any cardiovascular cause, and so on. A cardiovascular study's **composite endpoint** would typically include several specific endpoints. Any single patient can be counted as having reached the composite endpoint only once. For example, a patient who suffers an acute coronary event, is hospitalized, and then subsequently dies would be counted as only one endpoint incident. Combining endpoints into a composite has the advantage of requiring shorter clinical trials and fewer study subjects compared to trials with only a single endpoint, such as an acute coronary event. It is mandatory that each component of the composite endpoint be similarly important, similarly severe, and clinically related to the composite.[11]

The four basic types of data are listed and described in **Table 2-5**. **Continuous data** have a predictable mathematical relationship between data points. Blood pressure, distance walked, and weight gained or lost are examples of continuous data. A patient with peripheral artery disease who can walk 20 meters without pain is able to walk twice as far as a patient who could only walk 10 meters. A patient who has lost 30 pounds on a weight-reducing diet has lost twice as much weight compared to a patient who has lost only 15 pounds. Continuous data can be parametric or nonparametric. The term **parametric** denotes that the data points are evenly distributed around a mean, or average. Data that are skewed, containing significant outliers, are considered to be **nonparametric**. An example of nonparametric, or skewed, data might be a set of final examination scores taken by a small class of 10 students. Nine of these students received an examination grade ranging from 89 to 94 percentage points. The 10th student had a very bad day and received a score of zero. Here, a very small proportion of the subjects introduced a value that was so much different from the majority that it significantly altered, or skewed, the average. In general, the class did very well as a whole; however, the average class score of 80% would inappropriately give the impression of a mediocre class result. It is inappropriate to describe nonparametric continuous data with statistical terms, such as the mean, and it is equally inappropriate to assess it with parametric statistical tests, such as the Student's *t*-test and others. In this example, it would be better to describe the class performance in terms of a median score of 90% rather than the mean of 80%.

Table 2-5 Data Types with Examples

Data Type	Description
Continuous, parametric	A set of numeric data points that have a mathematical relationship and are evenly distributed around a mean. *Example:* Class grades on an examination where the average grade (85) is close to the center of all grades, which range from 80 to 90 points. A student earning an 88 has done about 10% better than a student earning an 80.
Continuous, nonparametric (skewed)	A set of numeric data points that are skewed (i.e., not normally distributed around the mean). Either abnormally high or low outliers cause the mean to be a poor descriptor of the data as a whole. *Example:* The same examination results as above, but in this case a small number of students earned 0, causing the mean to underestimate the overall good performance of the class as a whole. *Examples of commonly skewed clinical data:* length of hospital stay (LOHS), serum triglyceride levels
Ordinal	An ordered set of data where there is no mathematical relationship between the values. *Example:* Results of a Likert-type scale evaluating a course instructor. Each student could rank the instructor as: 1 = terrible, 2 = OK, 3 = pretty good, or 4 = wonderful. It would be inappropriate to report the results as a mean score because there is no mathematical relationship between the numbers 1 through 4. A mean score of 4 does not mean that the instructor is twice as good as another faculty member with a mean score of 2. This is where medians are appropriately used. *Examples of ordinal clinical data:* Likert scales measuring pain intensity, depression, or anxiety.
Nominal (categorical)	Simply descriptive categories or names. Most commonly reported as proportions. *Examples of nominal clinical outcomes:* proportion of patients cured, improved, worsened; dead, alive; and in remission vs. stabilization vs. progression.

In clinical trials, length of hospital stay[9] and serum triglyceride levels[10] often are skewed. There almost always seems to be a small number of patients who have unusually long hospital stays and unusually high serum triglyceride concentrations. In both of these examples, you will find that statistical tests used to assess these data are designed for nonparametric data. Skewed continuous data is transformed into ordinal data and assessed as ordinal data.

CLINICAL PEARL

A number of outcomes seen in clinical trials are frequently skewed and therefore nonparametric. Statistical tests for assessing nonparametric data must be used to analyze such data.

With **ordinal data**, the data are ranked but there is no consistent quantitative relationship between the data points. Likert scales generate ordinal data. Several Likert scales measure the severity of conditions such as anxiety and depression. An 11-point scale is commonly used clinically to quantify pain severity. It assigns a point value from 0 to 10, with each value having a pain severity description (e.g., 0 points for "no pain" and up to 10 points for "the worst pain imaginable"). This data is ordinal and not continuous because there is not a consistent quantitative relationship between scores; that is, a pain score of 4 cannot be assumed to be twice as severe as a pain score of 2.

CLINICAL PEARL

Few clinical trials use ordinal data for their outcomes. If you think that an outcome represents ordinal data, you might be correct but think again.

Consider the following regarding ordinal endpoints:

1. Relatively few clinical trials use ordinal data for their outcomes.

2. Oftentimes in clinical research the results of ordinal and Likert scales, such as those used to assess pain, depression, and anxiety, will be assessed statistically as though they were continuous data. This practice, right or wrong, is a little more acceptable in robust trials evaluating large numbers of patients.
3. The pain scale commonly referred to as VAS, or the visual analog scale, uses a continuous horizontal strip that does not have pain severity descriptions or rating numbers. Patients are told to indicate their pain severity from left-most (no pain) to right-most (worst pain). A clinician then measures the point's position in millimeters and rates the patient's pain severity on a continuous scale. VAS data are usually considered to be continuous rather than ordinal data.

The word *nominal* is derived from the Latin word *nominum*, meaning "name." **Nominal data**, also known as **categorical data**, do not have a mathematical value. Nominal data can be dichotomous (dead versus alive, cured versus not cured), but more than two possible outcomes are possible (cured versus improved versus worsened). The key to recognizing nominal data is that it is reported as a proportion or a percentage.

CLINICAL PEARL

When a study defines its endpoint as "the proportion of patients who" or "the percentage of patients who," you can be pretty confident that the data are nominal.

Descriptive and summary statistics are listed and described in **Table 2-6**. As a general statement, only parametric and continuous data should be reported as a mean. Skewed continuous data and ordinal data are most appropriately described using the median, range, interquartile ranges, and percentiles.

Figure 2-1 illustrates the meaning and utility of a **standard deviation (SD)**. When a mean is reported, it is common practice to report a standard deviation. The standard deviation describes how the results are distributed around the mean. For example, say that exam results in a course are equally, or normally, distributed around the average grade for the exam. The median grade on the exam is generally very close to the mean grade for that exam. The results are reported as a class mean, say 85%, with a standard deviation of 2%. These two summary statistics offer a good bit of information on how the class as a whole performed on the exam. The mean score of 85% ±1 SD represents all of the grades between 83% and 87%. Because the mean ±1 SD includes approximately 68% of the results, you can assume that 68% of the class earned an exam grade between 83% and 87%. Likewise, the mean ±2 SD represents grades between 81% and 89%, which describes 95% of the students in the class. Virtually all of the student grades are included in the mean ±3 SD, so all of the exam grades probably fell between 79% and 91%. When a standard deviation is large relative to its mean, one can assume that the results are either skewed and/or excessively variable. Say that the students took a poorly written examination and the mean grade was 80% and the S.D. was 40%. If you go through the math, you see that we run into trouble. You would estimate that 68% of the students earned an exam grade between 40% and 120% and 95% of the students earned a grade between 0% and 160%. This doesn't make a lot deal of sense. If this were to happen, the examination was either not valid or the wrong Scantron® examination key was used.

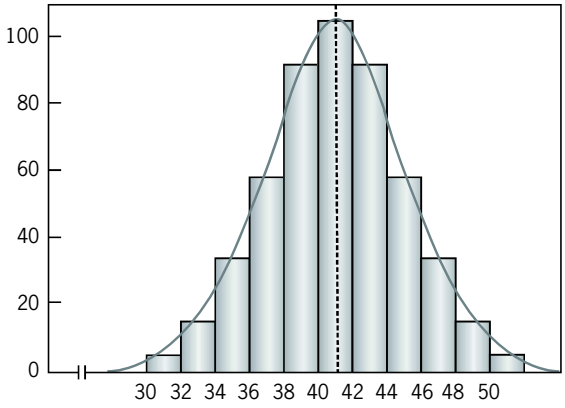

Figure 2-1 Normal distribution curve and standard deviation.

CONFIDENCE INTERVALS

Confidence intervals reveal important information regarding the statistical significance of results and often are used in published reports.[12,13] A commonly used explanation of what a 95% confidence interval (95% CI) is: "If you repeat a clinical trial 100 times, the mean result of 95 of these studies will fall within the 95% confident interval." Well, this explanation is correct, but who would waste their time and money repeating a trial that many times? Let's try to approach

Table 2-6 Descriptive Statistics

Statistic	Description
Mean	The most commonly used descriptive statistic. Also called the *average* or *arithmetic mean*, it is calculated by simply totaling all the values of the dataset and dividing this value by the number of data points. The mean is used only for continuous data and is best used as a descriptor of nonskewed continuous data. The standard deviation is a statistic that describes the dispersion or spread of data around the mean of the sample (see Figure 2-1).
Median	The median is the centermost value of a dataset. The median is a better descriptor of ordinal or significantly skewed continuous data. *Example:* Housing values in a small neighborhood of seven houses are as follows: $100,000, $120,000, $125,000, $150,000, $180,000, $190,000, $600,000. The median value of $150,000 better describes the value of the houses in this neighborhood than the mean of $210,000.
Modes	The mode is the value that occurs most frequently in a dataset. It often is not reported but it is useful when describing things such as the incidences of an occurrence related to time, age, or season. *Examples:* The incidence of snowstorms in this area is highest in February. Or, seizure disorders are most frequently diagnosed during two age ranges, less than 6 years and over 65 years. The seizure example is bimodal.
Risk/hazard ratio	A simple ratio describing the likelihood of an event. *Example:* The risk of a developing cervical cancer in a certain population of unvaccinated women is approximately 1 in 30 over 5 years, a risk ratio of 0.033, or 3.3%. In a similar population that received a series of HPV vaccines, the risk ratio is 0.016, or 1.6%.
Absolute risk reduction	In the example above, one can estimate, simply by subtraction, that being vaccinated would reduce the *absolute* risk of developing cervical cancer by 0.017, or 1.7%.
Relative risk and relative risk reduction	Using the same above example, the *relative* risk for cervical cancer with vaccination is estimated, by simple division, to be only 0.48, or 48%, of the unvaccinated population's risk, a relative risk reduction of 0.52, or 52%.
Odds ratio	We will not go into detail about odds ratios, but the odds ratio will usually be similar to the risk ratio in similar situations. However, because odds are not the same as risks, the odds ratio can deviate from the risk ratio and be much higher as the risks of an event becomes greater.

this from a very slightly different angle using the following fictitious clinical trial.

A very well-designed and executed clinical trial compared the lipid-lowering effect of a new agent, Newlip, to an older agent, Olelip, in a sample of 200 hyperlipidemia patients. The primary endpoint was the mean difference in mean LDL reduction after 6 months of therapy with Newlip versus Olelip. An LDL reduction of 35% was found with Newlip and a 30% reduction with Olelip. The mean difference was 5%, and the 95% CI for this mean difference was determined to be 1% to 7%.

Based on this information, consider the following questions:

1. Is the apparent superiority of Newlip statistically significant at a p value of 0.05 or less?
2. Should I expect that when Newlip is available for widespread use that it will outperform Olelip by 5%, on average?

This 5% superiority for Newlip is the mean result of this single study based only on a sample of patients selected from potentially millions of patients with hyperlipidemia. Therefore, the 5% result can only be, at

best, an estimate of its real effect. If the study included millions of patients, then one could be relatively secure that the 5% better response is precise and reproducible.

The following are the answers to the above questions:

1. Newlip was shown to be statistically significantly better than Olelip. The 95% CI tells us that the possible differences (with 95% confidence) between the drugs is somewhere between 1% better (the worst case) and 7% better (the best case). The worst case of being 1% better is still better, and therefore the drug is statistically superior with a p value of 0.05 (i.e., 5%) or less.
2. The 95% CI suggests that the probable real superiority of Newlip may be as little as 1% or as high as 7% when used in the larger population. Which is more likely to be correct? The 1% estimate is exactly as probable and valid as the 7% estimate, or, for that matter, any result between 1% and 7%.

But what if the mean difference was 5% and the 95% CI was –1% to 7%? In this case, the difference between the two drugs is *not* statistically significant. In fact, the possibility exists that Newlip is actually 1% worse than Olelip. This possibility of inferiority is no less likely than the possibility that it is 7% better.

Note that the 95% CI (–1% to 7%) suggests nonsignificance because a zero is found within the interval. This is true when the data being assessed are absolute data or nonratio data. In other words, zero superiority is a valid possibility.

If we were looking at ratios (hazard ratios, odds ratios, or risk ratios) in a 95% CI, then finding the value of 1.0 in the interval would suggest statistical nonsignificance. In other words, a ratio of 1.0 within the confidence interval suggests that no change in the risk, odds, or hazard is a valid possibility.

When using 95% CIs for determining noninferiority, the investigator looks within the confidence interval for the worst case (usually the lower bound of the CI) and compares this value to a predetermined threshold for inferiority. For example, imagine that this fictitious lipid study was designed as a noninferiority trial. To determine if Newlip is noninferior to Olelip, the investigator set a –5% inferiority threshold for Newlip. The mean difference in lipid reduction with Newlip (35% reduction) compared to Olelip (30% reduction) was +5% better reduction if the 95% CI was found to be –7% to +11%. Based on these results, Newlip was *not* found to be noninferior to Olelip because it might be more inferior than the inferiority threshold of –5% (i.e., possibly inferior by as much as –7%).

STATISTICAL TESTS AND APPLICATION

The key to mastering this chapter's elusive skill of applying the appropriate statistical tests to study data is Table 2-8, the "Application of Statistical Tests." The effective use of this table depends on your having a good understanding of the data types just discussed and a working understanding of the study designs and confounder types listed. Before discussing the application of statistical tests, however, let's use **Table 2-7** to review some fundamental statistical concepts and definitions.

When reviewing published studies, you will frequently encounter the term *variable*. The two types of variables are independent variables and dependent variables. An **independent variable** is a causative variable and is either the primary intervention that is being studied in a trial or a secondary factor that has the potential for influencing the research results. The primary independent variable in a drug study is simply the administration of the study drug to study subjects. **Dependent variables** are the outcomes caused by the independent variables. In plain English, the independent variable, the drug, is expected to cause outcomes, or dependent variables, such as therapeutic effects and adverse effects.

In most clinical situations, a variety of conditions can influence the expected results of an intervention, or independent variable. For example, say that an antihypertensive drug therapy (the independent variable) is expected to result in a mean blood pressure lowering (the dependent variable) in a group of patients with hypertension. A number of subjects in the trial, however, might have characteristics

CLINICAL PEARL

It is important to remember that type I errors are only possible when the results are determined to be statistically different and that type II errors are only possible when no statistical difference is found between the results.

Table 2-7 Fundamental Statistical Terms and Concepts

Term/Concept	Description
Independent vs. dependent variables	Variables are simply a fancy way of saying things that are done or things that result when a study is performed. *Independent variables* are interventions that are not affected by other things in the study. Simply put, the main independent variables are the interventions that are being studied (e.g., administration of drug X and a placebo in a trial comparing the efficacy of drug X to placebo). *Dependent variables* are affects that are influenced by the independent variables and are measurable and variable. Simply put, dependent variables are the outcome measurements of the study. *Example:* The independent variable, administration of antihypertensive drug A, results in a dependent variable, the mean blood pressure change of the subjects.
Confounders and covariates	Confounders and covariates also are independent variables. They might affect the outcome of the study results, but they are usually not the main intervention being studied. Confounders often are categorical or nominal in nature, as the example illustrates. Covariates, in contrast, are continuous data confounders that might possibly alter the outcome, such as baseline (pretreatment) blood pressure. *Example:* Drug X is expected to reduce the subjects' blood pressure, but confounders such as obesity, the clinic in which the patient is being treated and studied, race/ethnicity, and so on, might alter the degree of blood pressure lowering.
Type I error and alpha (α)	If an investigator looks at the results of a clinical trial and concludes that one treatment is better than (or different than) another and the difference in apparent efficacy was due to chance rather than the differences in the efficacy of the drugs, the investigator has made a type I error. The risk of a type I error is almost never zero. Researchers generally accept up to, but not greater than, a 5% risk of making a type I error. When a study is being designed, an investigator will prospectively choose the level of type I error risk he or she will accept once the study is completed. This is called the *alpha*, or α, and it will always be 5% (0.05) or less.
Type II error, beta (β), and power	If an investigator looks at the results of a clinical trial and concludes that one treatment is no better than (or is not different than) another and the apparent lack of difference is false and was due to chance or a weakly powered study, the investigator has made a type II error. Beta, or β, is the risk of erroneously concluding "no difference" when a difference really exists. Generally, a beta, or chance of a type II error, of 20% (0.2) or less is acceptable. The corresponding power value in this case is 80% (0.8) or more.
p-values	A *p*-value is generated from the result of a statistical test of outcome data, and it simply provides the chance that a type I error has occurred. *Example:* The results of a trial report that 60% of patients treated with "supercillin" for a skin infection are cured while 65% of patients treated with "super-duper-cillin" are cured. The result of a Chi-square statistical analysis of the data is reported as a *p*-value of 0.06. A *p*-value of 0.06 means that there is a 6% chance that a type I error would occur if the investigator concluded that "super-duper-cillin" was *statistically* superior to "supercillin."

such as obesity or might be members of a racial or cultural group that has a different response to blood pressure medications. As discussed previously, these secondary, or nuisance, independent variables are **confounders**. When a confounder represents a continuous data type, such as baseline blood pressure or body mass index, it is called a **covariate.**

Type I and type II errors are relatively easy to understand and are described in Table 2-7. Remember which type of error is which and that **type I errors** are only possible when results are concluded as being statistically different and **type II errors** are only possible when no statistical difference is found between results. Two of the most important concepts regarding *p*-values are what they tell us and what they don't tell us. Note that ***p*-values** simply provide a measure of the risk of a type I error if it is concluded that two results are different. When a statistical test indicates

that the mean result of one intervention is statistically significantly different from the mean result of another intervention with the *p*-value of 0.02, then there is a 2% chance that concluding such a difference is incorrect and that the difference was a result of chance rather than the studied intervention (a type I error). Likewise, a *p*-value of 0.06 would suggest a type I error risk of 6%, which is above the universally accepted limit of alpha of 5% and is generally thought of as being statistically nonsignificant.[14]

It is important to note that statistical significance is quite different from clinical significance. A statistically highly significant difference between two results is not evidence for a clinically significant difference. A statistical significance defined with a *p*-value of 0.001 compared to a *p*-value of 0.04 does not indicate a difference in superiority or effect size of either result. It only indicates that in the first case the chance of a type I error is only 0.1% and the chance of a type I error in the second case is 4%. When reporting results in a drug information response, and stating that an intervention produced a statistically significant better result than another intervention, nothing has been said about how superior one intervention is compared to the other. This is still true even when the *p*-value is given. It is necessary to include the actual difference in the results, such as drug A resulted in a mean blood pressure reduction that was 10 mm Hg greater than that produced by drug B.

CLINICAL PEARL

Statistical significance is not the same as clinical significance. A statistically highly significant difference between two results is not evidence for a clinically significant difference. A statistical significance defined with a *p*-value of 0.001 compared to a *p*-value of 0.04 does not indicate a difference in superiority or effect size of either result.

Let's now move on to **Table 2-8**. The steps involved in utilizing this table are as follows:

1. Identify the study's design. (column 1)
2. Determine if the investigators intend to evaluate confounders or covariates. (column 2)
3. Identify the data type of the results being statistically assessed. (row 1)
4. See the appropriate statistical test provided for that study design, confounder(s), and data type. (intersection of column 2 and row 1)

Depending on the answers to these questions the user would select the most appropriate column, row, and subrow. The statistical test found at the intersection of a column and row of the table is generally the most appropriate statistical test designed for that result.

Let's work through a couple of examples.

- *Example 1:* A clinical trial compares the ability of an inhaled long-acting beta-adrenergic agonist to reduce nighttime awakenings due to shortness of breath compared to the use of a placebo inhaler. Eighty patients are recruited and randomized in a 1:1 ratio to receive active therapy or placebo. The primary endpoint is defined as the difference between the mean number of nighttime awakenings per week in the active versus the placebo study groups. The investigators are not interested in assessing the effects of any potential confounders or covariates in this study. The trial results indicate that the use of the active beta-adrenergic agent is associated with a mean frequency of one nighttime awakening per week, and the placebo therapy is associated with a mean frequency of three nighttime awakenings per week. Which statistical test is most appropriate to determine if this difference is statistically significant?

  Answer: The correct statistical test is Student's *t*-test. The study features two independent groups, parallel design arms, no cofounders, and continuous parametric data.

- *Example 2:* This study is practically identical to that described in Example 1, but in this trial the results of three interventions are compared: weekly nighttime awakenings on beta-adrenergic agonist A versus beta-adrenergic agonist B versus inhaled placebo. Which statistical test is now the most appropriate to determine if there is a statistical difference between any of the three results?

  Answer: The correct statistical test is ANOVA (one-way analysis of variance). The study features more than 2 independent groups, parallel design, no cofounder, and continuous parametric data.

I will offer a few general pearls about data types and study design and then will comment on a few special considerations related to selected statistical tests discussed in Table 2-8.

Table 2-8 Application of Statistical Tests

Study Design	Confounder (Data Type)	Continuous Parametric Data	Ordinal or Skewed Continuous Data	Nominal or Categorical Data
Independent Groups				
Parallel design, 2 arms	No confounder	Student's *t*-test	Wilcoxon rank sum[d]	Chi-square or Fisher's exact test
	1 confounder (categorical)	2-way ANOVA	2-way ANOVA ranks	Mantel-Haenszel chi-square[c] or Cochran-Mantel-Haenszel chi-square
	≥ 2 confounders (categorical) or ≥ 1 covariates (continuous)	ANCOVA	ANCOVA ranks	Logistic regression
Parallel design, > 2 arms	No confounder	1-way ANOVA	Kruskal-Wallis[a]	Chi-square
	1 confounder (categorical)	2-way ANOVA[a]	2-way ANOVA ranks[a]	Mantel-Haenszel chi-square or Cochran-Mantel-Haenszel chi-square
	≥ 2 confounders (categorical)/ ≥ 1 covariates (continuous)	ANCOVA[a]	ANCOVA ranks[a]	Logistic regression
Related Samples				
Crossover, 2 arms	Not applicable	Paired (Student's) *t*-test	Wilcoxon signed rank	McNemar chi-square
Crossover, > 2 arms	Not applicable	Repeated measures ANOVA[a,b]	Friedman ANOVA[a]	Cochran Q
Pre-post	No confounder	Paired (Student's) *t*-test	Wilcoxon signed rank[d]	McNemar chi-square[d]
	1 confounder (categorical)	Repeated ANOVA	2-way ANOVA repeated ranks	
	≥ 2 confounders (categorical)/ ≥ 1 covariates (continuous)	Repeated measures regression	Repeated measures regression	

[a] ANOVA and ANCOVA tests, when used to assess more than two comparisons, will only indicate that at least one result is statistically different. A multicomparison procedure (MCP) is then performed to determine which result is different.
[b] Repeated measures ANOVA also can be used in independent group trials when repeated measurements are taken from the same subject. *Example:* Blood pressure measurements at 1, 2, and 3 hours following treatment.
[c] Mantel-Haenszel chi-square or Cochran-Mantel-Haenszel tests are commonly used when the study uses stratification in the methodology. *Example:* The center in multicenter studies.
[d] Do not confuse Wilcoxon rank sum, Wilcoxon signed rank, and Wilcoxon tests. They are not the same. The Wilcoxon test is used to evaluate time to an event (e.g., survival).

The statistical tests in column 4, "Ordinal or Skewed Continuous Data," are specifically designed to assess ordinal data. Clinical trial readers will find that these tests are frequently employed when data are continuous but not parametric. As noted earlier, skewed continuous data cannot be adequately described using the statistical terms meant for parametric data and means of results. Likewise, statistical tests designed for parametric data are inappropriate when the data being evaluated are skewed and nonparametric. When an investigator is faced with the need to assess skewed continuous data, it might be necessary to convert the data from continuous to ordinal data and then assess the "ordinalized" data using statistical tests found in the ordinal statistical column.

Regarding the influence of confounders or covariates on the appropriateness of a statistical test, the reader should appreciate that investigators are not required to consider them when statistically assessing their data, but failure to do so might constitute a study limitation in some instances.

The Student's *t*-test is the simplest and most straightforward statistical test for continuous parametric data. When an investigator uses more complex statistical assessments such as ANOVA or ANCOVA, there is usually a good reason to do so. Likewise, the same holds true when investigators use statistical tests that are more complex than the Wilcoxon rank sum for ordinal data or the chi-square (χ^2) test for nominal data.

There are two reasons to use ANOVA tests. The first is that the two-way ANOVA is utilized similarly to the Student's *t*-test (continuous data, comparing two mean results) except that in this case the effect of a confounder needs to be controlled within the statistical assessment. In essence, a two-way ANOVA evaluates the effects of two independent variables: the primary variable (the drug) and the nuisance variable (the confounder) and thus can control for the effects of such nuisances and allow a more selective assessment of the effect of the primary variable being studied. The second reason for using an ANOVA test is when a comparison involves more than two study arms. In such cases, the one-way ANOVA is used to avoid "multiple-measurement bias."[15] If a Student's *t*-test was used multiple times to evaluate the differences among three or more treatments, each repeated comparison would accumulate additive risks for making a type I error. The result of a one-way ANOVA will only tell us that either no difference exists between any of the interventions results or that at least one result of the three or more interventions is statistically different from the others. It will not identify which result is the different one. In order to identify which result is different, an additional post-hoc statistical test or multicomparison procedure (MCP) is required. There are several MCP procedures, including Tukey, Scheffe, Duncan, and the Bonferroni correction.

CLINICAL PEARL

The student's *t*-test is the simplest and most straightforward statistical test for continuous parametric data. If a more complex test is used, such as ANOVA or ANCOVA, there is usually a good reason why. The same is true when investigators use statistical tests that are more complex than the Wilcoxon rank sum for ordinal data or the chi-square (χ^2) test for nominal data.

Although ANCOVA is used in situations with multiple confounders, its main purpose is to assess data that includes a covariate (a continuous data confounder such as baseline blood pressure or baseline depression scores). Just like the ANOVA, when evaluating the differences between three or more results with ANCOVA an additional post-hoc statistical test or MCP will be required to identify which result is statistically different.

Paired and repeated-measurement statistical tests are used when data are obtained from nonindependent groups of subjects, such as those used in crossover and pre-post study designs. Additionally, they are used when studies repeatedly take and evaluate measurements from an individual subject. The terms *paired* and *repeated* often are part of these statistical tests' names; however, you need to recognize that statistical tests such as the Wilcoxon signed rank, McNemar, Friedman ANOVA, and Cochran Q tests also are designed to evaluate data derived from nonindependent sources.

Regarding statistical tests for nominal (categorical) data, many of the statistical tests in column 5 of Table 2-8 are only variations of the chi-square (χ^2) test. The Fisher's exact test is one example. The Fisher's exact test is, essentially, a chi-square test designed to handle smaller data values. If you needed to statistically evaluate the difference in adverse reaction frequencies between two medication therapies, you might find that one or more types of adverse reaction occurs infrequently (e.g., fewer than four or five incidences). The Fisher's exact test is designed to handle smaller nominal values with more precision. Some investigators always use the Fisher's exact test instead of chi-square regardless of the size of the values in the data. This practice is fine.

Be careful with look-a-like, sound-a-like statistical names. The Wilcoxon rank sum test, the Wilcoxon signed rank test, and Wilcoxon test are three different statistical tests used for entirely different reasons. The Wilcoxon test is not included in Table 2-8, but it is one of several tests used to assess time to an event or survival.

Table 2-9 lists the major categories of regression analyses used in clinical research. This table describes the type and number of explanatory or causal variables and the type of outcome or result variables associated with each type of regression test. Additionally, the common ways in which the results of regression analyses are reported also are included. In general, regression analyses explore, model, and quantify the predictive ability or the contribution of one or more explanatory variables on an outcome. An example of a simple linear regression analysis might be a study investigating the impact of patients' daily dosing frequency on medication adherence. In this fictitious regression analysis, it was found that a higher frequency of doses required each day correlated with a greater degree of poor medication adherence (e.g., QID dosing results in poorer compliance than TID dosing; TID dosing results in poorer compliance than BID dosing). The study reports an r^2 statistic of 0.5, which means that 50% of a patient's nonadherence can be explained by the number of dosing intervals that a patient is expected to take daily.

Previously, we discussed statistical tests that are related to regression analysis, such as 2-way ANOVA, ANCOVA, and logistic regression. These tests are used to control secondary variables (confounders and covariates) to more specifically focus on the effects of the primary causative variable being studied (e.g., the effect of the drug therapy). Multiple regression analyses, on the other hand, serve to quantitate the influence of confounders and covariates on the outcome. We are familiar with and have used the Cockcroft-Gault equation to estimate a patient's creatinine clearance given his or her age, weight, gender, and serum creatinine. From where did this equation come? Yes, you guessed it, a multiple linear regression analysis that quantified the influences of age, weight, gender, and serum creatinine relative to a patient's measured renal function.

The Cox proportional-hazard regression and the log rank regression are two of the more commonly used regression analyses used for comparing two or more "survival" curves, such as a Kaplan-Meier Curve, and for performing time-to-event analyses.

Table 2-9 Regression Analyses

Predicting Responses from Variables	Number of Explanatory Variables (Data Type): Confounders and Covariates	Number of Outcome Variables (Data Type): Result Data	Reporting of Statistical Assessment
Simple linear regression	1 (continuous)	1 (continuous)	Regression coefficient and its 95% CI, r^2, test statistic, *p*-value
Multiple linear regression	> 1 (continuous or categorical)	1 (continuous)	Regression coefficients and their 95% CI, r^2, test statistic, *p*-value
Simple logistic regression	1 (continuous or categorical)	1 (categorical)	Regression coefficient, OR, and its 95% CI, test statistic, *p*-value
Multiple logistic regression	> 1 (continuous or categorical)	1 (categorical)	Regression coefficients, ORs, and their 95% CIs, test statistic, *p*-value
Cox proportional-hazard regression (CPHR)*	> 1 (continuous or categorical)	1 (continuous time to event)	Regression coefficients, HRs, and their 95% CI, test statistic, *p*-value, "time-to-event" analysis
Log rank*	Compares the time to event from 2 or more groups; it does not evaluate multiple explanatory variables	1 (continuous time to event)	Proportion of event-free patients at a given time; test-statistic; *p*-value

*Both the Cox proportional-hazard regression test and the log rank test are used to compare survival or time-to-event data from Kaplan-Meier curves.

Abbreviations: HR, hazard ratio; OR, odds ratio; r^2, coefficient of determination, R^2, coefficient of multiple determination for multiple regression models; test statistic, t or χ^2 statistic used to determine the *p*-value.

DISCUSSION QUESTIONS

1. What are some of the limitations of observational studies as compared to interventional studies? What are the advantages or rationales to utilize data from this type of design?
2. In what instances should there be great concern if a study is not powered?
3. What is the difference between statistical and clinical significance?

REFERENCES

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it isn't. BMJ. 1996;312:71–2.
2. Vermeulen L. Gain in popularity of noninferiority trial design: caveat lector. Pharmacotherapy. 2011;31(9):831–2.
3. d'Amato CC, Pizza V, Marmolo T, Giordano E, Alfano V, Nasta A. Fluoxetine for migraine prophylaxis: a double-blind trial. Headache. 1999;39:716–9.
4. Chan AT, Manson JE, Feskanich D, Stampfer MJ, Colditz GA, Fuchs CS. Long-term aspirin use and mortality in women. Arch Intern Med. 2007 Mar 26;167(6):562–72.
5. Rich S, Rubin L, Walker AM, Schneeweiss S, Abenhaim L. Anorexigens and pulmonary hypertension in the United States: results from the surveillance of North American pulmonary hypertension. Chest. 2000;117:870–4.
6. CDC. *Pneumocystis pneumonia*—Los Angeles. MMWR. 1981;30:250–2.
7. Jones T. Dewey defeats Truman: well, everyone makes mistakes. Chicago Tribune. Available from: http://www.chicagotribune.com/news/politics/chi-chicagodays-deweydefeats-story,0,6484067.story.
8. Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles. JAMA. 1999; 281:1110–1.
9. Hollis S, Campbell F. What is meant by intention to treat analysis? survey of published randomized controlled trials. BMJ. 1999;319:670–4.
10. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125: 605–13.
11. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003;289:2554–9.
12. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med. 1986;105:429–35.
13. Gardner MJ, Altman D. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. BMJ. 1986;292:746–50.
14. Wears RL. What is necessary for proof? is 95% sure unrealistic? [Letter]. JAMA. 1994;271:272.
15. Smith DG, Clemens J, Crede W, Harvey M, Gracely EJ. Impact of multiple comparisons in randomized clinical trials. Am J Med. 1987;83:545–50.