How to Read a Medical Research Article

MERCK

RESOURCE DOCUMENT
Printed material...not to be used in detailing.

Table of Contents

How to Read a Medical Research Article

LEARNING OBJECTIVE

To become familiar with the structure of a medical research article and some basic methodology of a research study, so that an article can be reviewed quickly and critically for key points.

PERFORMANCE CRITERIA

- Describe the components of a citation and explain the significance of the journal title
- Explain the components of a well-written abstract
- Know where the study objectives would most likely appear in the body of an article
- Explain what a protocol is and know where the protocol is described in an article
- Explain what statistical significance means, as for example using the P value
- Know which section of an article contains the author's interpretation of a study

INTRODUCTION

As a Professional Representative, you frequently will be talking to doctors about the results of studies that have been done using Merck products. The purpose of this medical backgrounder is to familiarize you with the way such studies are written for publication in medical journals, so that if a doctor requests details from an approved reprint, you will be able to respond effectively.

While information in this backgrounder should increase your comfort, reading the medical literature is the only way to become thoroughly familiar with it. Carefully reading approved reprints as they become available will make it easier to discuss important highlights with physicians.

Note about study types

Some clinical studies are conducted to collect and evaluate information about disease, such as pathology, diagnosis, or epidemiology. While these studies are important, the main focus of this backgrounder is on studies evaluating drugs in the treatment of disease.

DRUG DEVELOPMENT PROCESS

The research and development process for new drugs takes a great deal of time and effort. As you begin to read a research report, it may be helpful to know if the research is part of the formal research and development (R& D) process.

HOW NEW DRUGS REACH THE MARKETPLACE

It takes about 12 to 15 years for a drug to move from the lab to a bottle sitting in someone's medicine cabinet. Approximately six of these years are spent in clinical trials prior to approval by the Food and Drug Administration (FDA). How these drugs are approved and why their approval often takes so long are two questions we can answer by reviewing the drug approval process.

DRUG APPROVAL PROCESS

In order for a drug to be administered to humans it must undergo preclinical testing in animals. This is followed by three stages whereby it then undergoes evaluation in humans - Phases I, II, and III - before approval by the FDA. Through all of these stages of drug development, safety is continuously monitored. The first step for a company that wants to test a new drug in humans is to submit an **Investigational New Drug Application (IND)** for review by the FDA. This is filed at the conclusion of preclinical testing period in animal models. Let's take a closer look at this process.

Preclinical The earliest estimates of safety and efficacy are determined by experiments in animal models. During this process, the manufacturer must show that the drug is reasonably safe to introduce in humans and that the compound exhibits pharmacologic activity that justifies commercial development. Review of the IND application involves three key areas:

- Animal pharmacology and toxicology studies.
- Data which shows the drug's composition and stability and manufacturing controls
- Proposed protocols for clinical studies.

The proposed clinical studies are reviewed by the medical officer who will determine whether humans could be exposed to unnecessary risks during testing as well as if the study design will provide data regarding the safety and efficacy of the drug. The FDA's Center for

Drug Evaluation and Research (CDER) has 30 days to review the IND application and notify the sponsor to proceed with clinical testing in humans.

Phase I Administered to approximately 100 to 200 healthy volunteers. Phase I tests are designed to assess the safety, tolerance, metabolism, absorption, elimination, preferred route of administration and insight into safe dosage range.

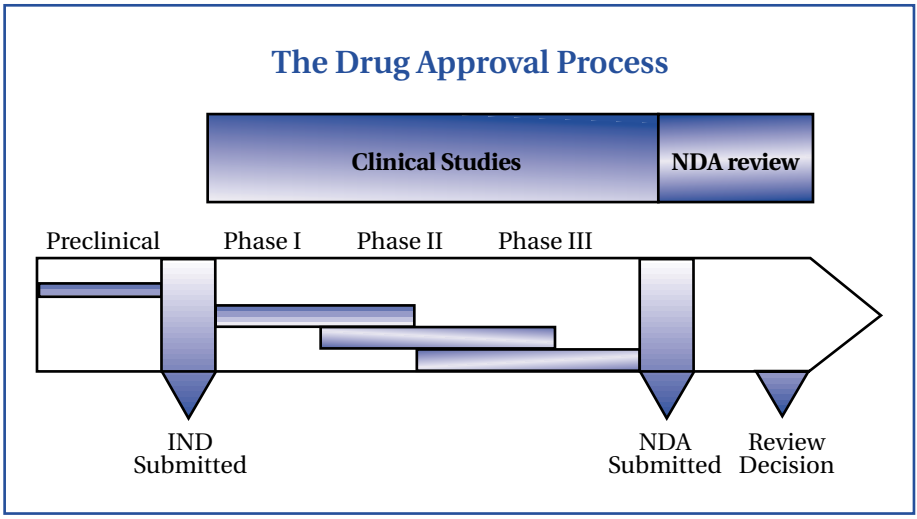
Phase II Phase II studies are designed to collect preliminary data on the effectiveness of a drug for a particular indication or indications in a small number of patients with a disease or condition. This phase lasts approximately two years allowing the researchers to learn more about the short-term side effects and risks of the drug.

Phase III Phase III studies are large-scale clinical trials designed to further determine the safety and effectiveness of a drug, identify adverse reactions and confirm optimum dosing regimen. Phase III studies involve 1,000 to 3,000 patients and last approximately three to four years.

Approval and Post-marketing

Throughout the entire drug approval process, CDER members are involved in an ongoing review of each phase of the study results, as noted above. Once Phase III clinical trials have been completed, the sponsor files a New Drug Application (NDA). The NDA is the way the sponsor will gain approval to market a new drug. Before the NDA approval is granted, extensive review of the data collected in Phases one through three is conducted by CDER who then re-analyzes the data collected and submitted by the sponsor. Review is often completed within 1 year from the time of the initial filing of the NDA. In some cases a "1P" review will be granted, which is an accelerated review and approval of an agent which is most often completed in 6 months from the date of the initial filing of the NDA.

Phase IV Phase IV trials continue clinical investigation of the drug in a large number of patients over a long period of time once the drug has been approved for marketing by the FDA. These studies provide additional data on current uses and the incidence of side effects.



The Citation

Before reading an unfamiliar research report, look critically at the citation: the authors, title, and journal where the article is published.

Authors

The authors may be listed as individuals, or as a group, as in the above citation. If individual authors are listed (as is usually the case), often the main investigator is the first author; if the study is from a university, the department head may appear as the last author. Additional information about the authors, such as government or other affiliation, is usually identified on the first page of the article at the bottom. Information about grants or sponsorships, as from a pharmaceutical company or medical society, is also provided, either on the first page or at the end of the article.

Title

The title of a scientific article is written to convey as much as possible about the study in the fewest possible words. You might think of it as a one-sentence abstract of the article! A well-written title can tell the reader at a glance which drugs were used, which disease was studied, and what type of study was done.

Journal

The journal name appears in abbreviated form, followed by the publication year, the volume, and the page numbers. If the article is from a journal supplement, the supplement issue appears immediately after the volume number.

It may be of value for you to know that, in the eyes of physicians (and the press), not all journals are created equal. An article in a well-respected journal, such as the *New England Journal of Medicine* or the *Journal of the American Medical Association,* usually carries more weight than does one in another journal.

One of the reasons that "not all journals are created equal" is the process by which manuscripts are reviewed and the criteria they must meet for acceptance. Some journals have little or no review policies. Rather, articles are accepted if a fee is paid to the journal. Journal supplements, which are often paid for by a pharmaceutical company or medical organization, may not have the same standards of review applied to them as are applied to articles in the regular monthly issues.

For the better medical journals, those serving on editorial boards are knowledgeable in research methodology. Statisticians, health economics experts and ethicists, in addition to clinicians in the subject field, help to ensure thorough review of the data submitted. Disclosure of any financial relationship between the author and the sponsor of the study is often important, as is the knowledge of whether an independent steering committee oversaw the study while it was in progress, and whether a manuscript must receive written approval from the study's sponsor prior to the author's submission of the manuscript to a journal. Journals which investigate these potential conflicts of interest and apply stringent acceptance criteria for manuscripts often gain the highest respect of their readership. Hence, studies published in such journals tend to have more credibility.

1. **True or False:** The editorial review process is basically the same for all articles published in medical journals, regardless of the journal.

Standard Format of Research Reports

If you review a medical research journal, most of the articles you see will fall into one of two easily distinguishable categories:

- Research reports
- Literature reviews or summaries

A **research report** deals with a single study or series of studies. The authors of the articles are the people who actually conducted and/or supervised the research.

A **literature review,** on the other hand, describes a number of studies conducted by many researchers. A literature review attempts to bring together much of the relevant research regarding a single treatment or disease. Such reviews are helpful for bringing the reader up to date on what is known about a particular medical problem or treatment.

Most of the articles you may use will be medical research reports, therefore this backgrounder will focus primarily on them. Literature reviews are generally easier to understand than research reports because they include less detail on the technical and statistical aspects of the research and provide more background information. If you can read research reports comfortably, the literature reviews that summarize them will be relatively easy to follow.

Common features of research reports

Nearly all research reports follow a common format of five sections:

- Summary or Abstract
- Introduction
- Materials and Methods
- Results
- Discussion

The names of these sections will vary according to each journal's style rules, but the format remains the same. By learning what each section contains, you will be able to find the information you need quickly.

The **Summary** or **Abstract** section describes the study in a few well-chosen sentences. This section usually appears at the beginning of the article. It is often a good way to quickly find out what is in the article; so it should be read thoroughly and carefully before proceeding to any other part of the study. You will find that, occasionally, the summary conclusions are not fully supported by the data in the article.

The **Introduction** section tells the reader what problem this research addressed. It usually describes the importance of the problem, briefly reviews previous research, and states the purpose of the study.

The **Materials and Methods** section describes in detail how the study was carried out. This includes a description of the patients and how they were selected, and a listing of the treatments and the diagnostic procedures that were used before, during and after treatment as well as the statistical procedures used to analyze the data.

The **Results** section tells what happened to the patients as a result of the treatment. This section is probably the most confusing to the lay person because of the heavy use of statistical tools and tables to describe the results. Due to space limitations, not all of the results available to the researchers can be published in this section. Questions often arise about the researchers' conclusions because readers of the article do not have access to the more extensive unpublished information.

The **Discussion** section summarizes the findings of the study and discusses their implications and limitations. This is the "bottom line" of the study from the point of view of the researcher, so it is one of the most important sections of the report. Reading a well-written discussion section can tell you what was found, what it means to medical practice, and what is still unknown.

2. Typically, a research article has 5 sections. Please name them.

The Summary or Abstract Section

After the citation, start reviewing an article by looking at the summary section. The summary is where one quickly learns what an article is all about and the conclusions drawn by the authors. In many cases, this is the only portion of the article that is read; if the reader can find the answer to a question or determine that the article is not relevant from the summary, it may not be necessary to read the entire article. As previously noted, however, occasionally the summary is not a completely accurate description of the study or its results.

Sample abstract below:

A multicenter, prospective, open-label, randomized trial compared drug A (treatment group with new therapy being tested) with drug B (control group or standard therapy). Efficacy and tolerability in the treatment of serious pelvic infections were evaluated in 94 female patients with acute salpingitis, pelvic abscess, or postoperative pelvic cellulitis. Duration of therapy averaged 5.4 days for treatment successes and ten days for treatment failures. The overall treatment success rate was 98% (43 of 44 patients) in the drug A treatment group, compared with 92% (46 of 50 patients) in the drug B control group (P=NS). Adjunct therapy for two treatment successes in the drug A treatment group included laparoscopy and surgical removal of a pelvic abscess without change in antibiotics. Both drug A and drug B were highly effective and generally well tolerated for the treatment of salpingitis, pelvic abscess, and post-operative pelvic cellulitis.

In a good summary, the study **design** is provided, the **number and initial condition of the patients** are given, the **treatment** is described, the **results** are tallied, and **final conclusions** are outlined. The summary thus condenses the entire article into a paragraph or two.

3. **True or False:** A well-written summary or abstract section provides the reader with information about study design, results, and conclusions.

The Introduction Section

Ideally, the introduction section should contain three things:

- A statement of the general problem area. (Which drug is being tested? Which disease?)
- A short review of previous research.
- A statement of the purpose or objectives of the study.

There is no standard format to observe when writing the introduction.

The purpose, if explicitly stated, is often found in the last sentence of the introduction section. If the purpose is not explicitly stated in the summary or introduction, you can probably get a good idea of the authors' intent by scanning the rest of the article.

4. **True or False:** The purpose or objectives of a study are usually found in the last sentence of the Introduction section.

The Materials and Methods Section

PROTOCOL OR STUDY DESIGN

The Materials and Methods section describes how the study was carried out. The methods followed may also be referred to as the **protocol** or **study design**. This section should contain enough details so that another researcher can read the article and conduct a similar study on another set of patients. The ability of other researchers to reproduce the results of a particular study refers to the reliability of that study. The **validity** of a study refers to the ability of a study to prove what the investigators set out to prove.

Four questions should be answered in this section:

- Who were the patients (and how were they selected)?
- What were the treatments?
- How were results measured?
- How was bias controlled?

NOTE: In developing a protocol, the researchers usually consult a statistician who can identify the appropriate statistical criteria needed to support their hypothesis and produce a clinically significant study. *On some occasions, when the results of the study do not support the original hypothesis, some researchers may attempt to rearrange the data into subgroups to support a new hypothesis or make other, unplanned analyses. Such unplanned analyses do not carry the same weight as pre-planned analyses. Also, if unplanned multiple comparisons are made additional statistical corrections will have to be made.*

5. Protocol, another name for _____ _____, describes the patient population, the treatments, how results were measured, and how bias was controlled.

Patient Selection

An experienced reader of medical literature who is critically evaluating a new study will look for the following types of information in the Materials and Methods section:

- How many patients were there?
- What conditions were selected for inclusion in the study?
- What conditions were excluded?
- Was the stage of their disease noted?
- For how long a period were the patients followed?
- What prior therapy did they receive?
- What concomitant therapy was allowed?
- What other data are given? (age, sex, race, etc.)

When comparing studies that appear initially to be similar, inconsistency in the criteria for selecting patients can lead to gross discrepancies in their results.

The study design is crucial to the reliability and validity of a study. Every study should be conducted in accordance with a predetermined protocol. Guidelines are set up for selection of patients, treatment, how tests are to be done, and what kinds of analyses are to be used in interpretation of results.

Treatments and Dosage Adjustment

The Materials and Methods section contains a description of what drugs were given and in what dosage, how they were administered, and how long they were given.

Baseline dosages of the agents being studied are established during Phase I studies in healthy persons. At that time any of several conditions may be identified as requiring adjustments from the baseline during later Phases. These include:

- age - prematurity, old age
- size - infants, children, adults and extreme variations from normal weight

- metabolism - state of the organ or organs where metabolism or processing takes place
- other diseases, states or conditions - e.g., diabetes, menopause, peptic ulcer
- concomitant medications - e.g., agents known to interact with the agent being studied.
- side effects - severity, e.g., determining point at which agent should be withdrawn.

Premature infants and very young children with developing metabolic systems as well as the elderly with reduced tolerance to exogenous substances should be considered in the protocol. Adjustments based on differing responses because of age usually involve decreases from the baseline dosage.

If children are included in the study, adjustments in dosage may be made by body surface area, measured in square meters (m^2) or by body weight, measured in kilograms (kg).

Allowances should also be made in the event that a patient has or develops a condition that inhibits or accelerates metabolism or processing of the drug for eventual elimination. For example, an agent that is metabolized primarily by the liver might accumulate and reach toxic levels in a patient with hepatic insufficiency. Unless the investigator took this eventuality into consideration when designing the study, the toxicity might be ascribed to the agent being studied.

In a similar manner, other disease states or conditions might have an adverse effect on the outcome of the study. Diabetes, for example, is one condition that is frequently found in the precautions section of direction circulars for agents cleared for use in patients. One reason for its importance relates to the altered microcirculation in the kidneys of diabetics; another is the delicate balance between hypo- and hyperglycemia that may be upset by agents that involve glucose metabolism.

Before a study is undertaken, the investigator needs to know how the study agent interacts with other agents or with the protein-binding of other agents so that he may be able to predict the responses of patients in his study who may be taking both agents. Occasionally, concomitant use of interacting agents is reason for exclusion of

otherwise appropriate patients from the study. In most cases, however, such use is covered in the protocol as a possible reason for dosage adjustment.

Side effects vary considerably in number and degree from patient to patient, and it is not uncommon for the dosage of a drug to be lowered to reduce side effects in a patient who is having problems tolerating a drug. In some studies, a specified method of dose reduction is defined in the protocol; in others, the drug may be withdrawn, or the attending physician may reduce the dosage at his discretion. In some cases, the patient may adjust the dosage without consulting the physician; if there are many of these cases in a particular study, the results may be invalidated by patient non-compliance.

Study Bias

Bias is any effect at any stage in a study tending to produce results that depart systematically from the true values (bias does not include simple "random" variation).

Before the results of a comparative study can be accepted, the reader needs to know how the researchers attempted to control bias. Many standard methods to reduce or eliminate bias are in common use, including blind and double-blind experiments, placebos, and most importantly, the use of randomization.

Bias can affect study results in two principal ways.

Patients being compared were different in some important way. In studying the data after the trial is completed, researchers may find some important prognostic factor that differs substantially between the treatment groups. When this occurs, the study groups are no longer comparable (this is especially common in non-randomized studies).

There may be differences in the way the groups were handled during the study, such as additional psychological support or exercise therapy for one group and not others. This is true especially in "open-label" studies.

Bias can make a sample or result different from what it otherwise would be. In "open" trials, which are common in the early stages of drug development, bias is almost unavoidable. Selection of patients may be unconsciously based on whether the patient is perceived as

likely to tolerate the drug or in need of the drug rather than by random allocation. Similarly, the physician or nurse may classify a very slight response in a patient on active drug as "real", while the same response in a patient known to be on placebo might be classified as "no response".

6. _____ is a statistical concept referring to effects that produce results different from the true values.

Minimizing Bias

Randomization

The most reliable way to control sources of bias is through **randomization**. In a carefully controlled clinical trial, patients are placed in a treatment or control group by some random method, such as through the use of a table of random numbers. Because the selection of the groups is out of the investigator's hands, there is a better chance that the groups will be comparable. Randomization attempts to ensure that extraneous factors that would effect the outcome of the trial are evenly distributed in all treatment groups so that such factors have no apparent effect.

Sometimes, a procedure known as **stratified random assignment** is used. In stratified randomization, patients are first categorized into groups or **strata** according to the seriousness of their disease, by age, or by other conditions they have. Patients within each stratum are then randomly assigned to treatments by the procedures already described. This approach may be important if great differences are expected in results among the various strata. Stratification should be part of the trial design; if patients are categorized only when the results of treatment are analyzed, questions may arise about the validity of the study.

Blinding

Another way to control bias is to make the trial **blind** or **double-blind,** a method which is often used in conjunction with randomization. In a double-blind trial, neither the patients nor the investigators know which patients receive the active drug(s) and which receive placebo. In a **random-double-blind trial,** random numbers are used to identify

the bottles of medication (active drug or placebo) for each patient. The contents of the medication bottles are known only to a third party (such as a pharmacist) who has no contact with the patients. The medical personnel who see the patients and dispense the medication bottles should not know at any time whether the patient is receiving active drug or placebo. (It is not always possible to conceal this information; for example, in a study comparing a beta-blocker to a placebo or to another type of medication, the pulse rates are often so much lower in the beta-blocker group that a physician or nurse can guess which patients are receiving the beta-blocker.)

In **single-blind trials,** the patient is unaware of the identity of the medication, but the medical personnel know. This may produce bias through their attitudes or expectations.

7. In a single-blind study the _____ does not know whether the medication is drug or placebo; in a double-blind study both the patient and the _____ do not know whether the medication is drug or placebo.

Sample Size and Power

Power is the ability of a study to detect a statistically significant difference in some endpoint between the treatment groups. The greater the number of endpoints or subjects, depending on the type of endpoint in each group, e.g., MI, death, etc., the greater the 'power' of the study to detect such differences. As the number of subjects increases in the treatment groups, the number of 'endpoints' (e.g., MI, death) should also increase. Therefore, the power of a study to detect significant differences increases as the number of subjects and the duration of the study (and consequently endpoints) increases.

To demonstrate the interplay between power and sample size, consider that drug X reduced mortality in two studies. In study #1, 2 of 4 patients in the control group died (50%) vs. 1 of 4 (25%) in the treatment group. In study #2, 200 of 400 patients died in the control group vs 100 of 400 in the treatment group. Obviously, the results of study #1 are not statistically significant because of the small numbers of patients and "endpoints" (deaths in this case), whereas the results of study #2 are highly significant because of the greater numbers of both patients and endpoints (deaths). Thus, study #2

had far greater "power" to detect a statistically significant difference than did study #1.

The following example demonstrates a study with sufficient sample size and power to have total mortality as the primary endpoint.

> During the double-blind study period 438 patients died, 256 (12%) in the placebo group and 182 (8%) in the treatment group; the relative risk was 0.70 (95% CI 0.58-0.85, p =0.0003) with active treatment. The Kaplan-Meier 6-year (70 months) probability of survival was 87.7% in the placebo group and 91.3% in the treatment group. Adjustment for the baseline covariates made no difference to the results for survival or the other endpoints. There were 189 coronary deaths in the placebo group (74% of all deaths in this group), compared with 111 in the treatment group. The relative risk of coronary death was 0.58 (95% CI 0.46-0.73) with active treatment. This 42% reduction in the risk of coronary death accounts for the improvement in survival. There was no statistically significant difference between the two groups in the number of deaths from non-cardiovascular causes. There were similar numbers of violent deaths (suicide plus trauma) in the two groups, 7 versus 6. Of the fatal cancers, 12/35 in the placebo group and 9/33 in the treatment group arose in the gastrointestinal system. There were similar numbers of cerebrovascular deaths in the two groups, and the difference (6 vs 11) in deaths from other cardiovascular diseases is not significant.

8. **True or False:** Power refers to the ability of a study to detect differences in the occurrence of a given endpoint, such as myocardial infarction or death in the treatment group. The larger the sample size, the greater the power of a study.

DATA ISSUES

Per Protocol vs. All Patients Analysis

When final data are being compiled for a clinical study, the researcher must decide whether to include only that data which is from subjects who conformed to the protocol, i.e. perform a **per protocol analysis,** or include data from all subjects, i.e., an **all patients or intention-to-treat analysis.**

Not all patients who are enrolled in a study necessarily complete the study. Serious adverse events may require discontinuation, or the patient may withdraw for personal reasons (e.g. moving away from a monitoring site). Even if a patient remains in the study for the duration, a deviation from the study design or "protocol violation" may be discovered which could invalidate data from a particular subject. Protocol violations can occur for many reasons: for example, a patient may be non-compliant, or perhaps it may be discovered, after the study begins, that an individual meets one of the exclusion criteria. Sometimes researcher error, for example, accidental "unblinding" of the treatment, can result in protocol violation.

Because sample size determines statistical significance of data, eliminating subjects from the final data analysis can have a great effect on whether or not the objectives of the study will be met. For example, if one were to discount three subjects in a study of 1,000 patients, it would not have nearly the same statistical effect as discounting three subjects in a study of 50 patients. Similarly, if one were to include data from those three subjects in each case, depending on the nature of the protocol violation, or at what point study withdrawal occurred, etc., the relative impact of those data would differ based on the sample size. Discounting subjects could also introduce bias into the analysis.

When protocols are developed and studies are designed, researchers depend on protocol compliance from all participants, both patients and investigators alike. If data from a subject or subjects is not included in the final analysis, it is important that the investigator state the reason(s) why. Failure to do so might compromise the credibility of the study.

9. True or False: Per Protocol analysis involves analyzing data from all patients entered into a study protocol.

Sensitivity and Specificity

In some studies, it is very important to be able to accurately detect the presence or absence of a certain disease. Sensitivity and specificity of a certain test are the criteria by which the accuracy of a test is measured.

Sensitivity is the ability of a test to detect the presence of given disease. The more 'sensitive' the test, the greater its ability to detect all of those who have that disease. A highly sensitive test will often be falsely positive in people who do not have that disease.

Specificity is the ability of a test to correctly label as 'negative' those who do not have the disease being screened for.

The Results Section

The Results section of a research article on a drug study typically presents the efficacy and safety data. To a layperson, the results section of a medical research article may be intimidating and perplexing. This section often contains statistics, tables and graphs with little explanation for those not familiar with the field. A short review of statistical concepts and reading tables and graphs is presented below.

Basic Statistical Concepts for Evaluating Results

Medical studies are conducted on samples or small groups selected to be representatives of a population. If a sample is truly representative of a population, then results from that study should be similar to results of such a study in the entire population. Also, the results would have a high probability of being reasonable and repeatable.

Data are often analyzed for:

- summary information, for example, to find the mean value of a measurement, and to determine whether the data follow a normal distribution
- variability of data around a central value using the range, standard deviation, and variance
- statistical significance, when the data from two populations can be compared, for example, a treatment group and a control group

Summarizing Groups of Values & Normal Distribution

One of the first steps in organizing data is to determine whether there is one number which can be used to summarize a group of values. The mean, median, and mode are used as summary statistics, i.e., they are used in an attempt to summarize a group of values. These represent a "typical" value among a set of values. Which of these, i.e., mean, median or mode, is the most appropriate summary of a group of values depends on whether the values are skewed in some way.

The **mean** is the average of the set of values, that is, the sum of the values divided by the number of values. This is a good summary statistic when the group values are "normally" distributed.

The **median** is the middle value of the set; it splits the set so that half the values in the set are greater than the median, and half the values are less than the median. It may not have a value similar to the mean, particularly if there are a few very high or very low values that skew the mean. Calculating both the mean and the median can point out this kind of skew very quickly.

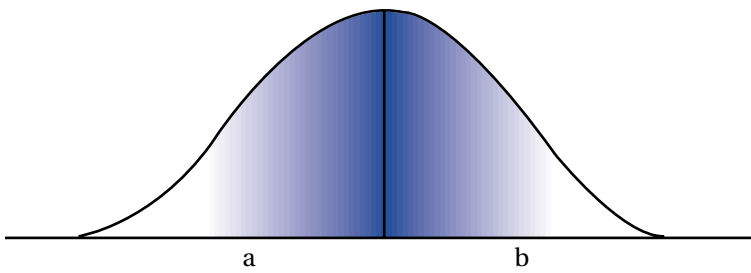
The median is a very useful summary statistic when the distribution of values within a group is skewed. For example, if we were to describe the "average" family income, the median would be more useful since the mean would be affected by the few families having very high incomes.

The **mode** is the most frequently occurring value in a set. It is not often discussed as a value, but comes from a frequency table.

The mode shows on a graph as a peak. If the data cluster is such that there are two peaks, then the data are **bimodal.** In a distribution which is VERY skewed, the mode may be the best single number to summarize the "average" value for the whole group of values.

In the ideal case, the sample will be **normally distributed** that is, the majority of the sample will fall in the middle and taper off on both extremes. A normal distribution is represented by a bell-shaped curve that matches the bell-shaped curve of the population.

Example of Normal Distribution or Bell Curve



The shaded area represents the majority of the population which falls in the normal range.

10. True or False: Calculations of mean, median, and mode are ways that data can be summarized to represent a group of values.

TREATMENT EFFECT

Evidence about the impact of a drug's treatment effect can be expressed as absolute risk reduction, relative risk, relative risk reduction, and number needed to treat." (Note, the treatment group = Y and the control group = X).

- ***Absolute risk reduction*** (ARR) is the absolute difference between the event rate in the treated group and the event rate in the control group. For example, if the event rates were 20% in the treatment group (Y) and 25% in the control group (X), the ***absolute risk reduction*** would be 5%.

- **Relative risk (RR) or risk ratio** is defined as the ratio of the incidence of some event in an "exposed" group (e.g., a group taking a drug) divided by the incidence of that event in a "non-exposed" (e.g., placebo) group. For example, if the event rates were 20% in the treatment group (Y) and 25% in the control group (X), the relative risk (also referred to as risk ratio) would be 0.20/0.25 = 0.80.

- **Relative risk reduction** (RRR) is the relative difference between the event rate in the treatment group (Y) and the event rate in the control group (X). More simply stated, RRR describes the percent reduction of adverse events if treatment is used, relative to the rate of adverse events in the control group. For example, if the event rates were 20% in the treatment group (Y) and 25% in the control group (X), the ***relative risk*** or ***risk ratio*** would be 0.20/0.25 = 0.80, and the ***relative risk reduction*** would be 100 x (1 - 0.80) = 20%. An RRR of 20% means that the new treatment reduced the risk of the adverse event by 20% relative to that occurring in the control group. In general, the greater the RRR, the more effective the therapy. You may see these results displayed like the figure shown on next page.

- **Number needed to treat** (NNT) is also a useful measure of the clinical impact of a drug and indicates the number of people who must be treated with the drug over a period of time in order to prevent one event. NNT provides healthcare professionals with more clinically relevant data to determine whether the treatment benefits are worth the risk. NNT is the reciprocal of the ARR and can be calculated as follows:

NNT = 1/absolute risk reduction

For example: if the ARR = 0.05 or 5% then the NNT = 1/.05 = 20. For every 20 patients treated with the drug being evaluated, one event is prevented, such as a heart attack.

Table 1. Summary of Measures of Effects of Therapy

Absolute Risk Reduction **X-Y**	0.25-0.20 = .05 or 5%
Relative Risk **Y/X**	0.20/0.25 = 0.8 or 80%
Relative Risk Reduction **[(X-Y)/X] x 100%**	[(0.25-0.20)/0.25] x 100 = 20%
Numbers Needed to Treat **1/ARR**	1/.05 = 20

To apply all of these statistical concepts, let's take a look at the results of the 4S Study:

HOW TO INTERPRET THE RESULTS OF 4S

The results of 4S are expressed in terms of the **'relative risk'** of death and various other endpoints. As a reminder, **relative risk** is simply a way of comparing the risk of some event in one group to the risk of that same event in another group, and is expressed as a **ratio of the risk** of the event occurring in one group as compared to a reference group.

In 4S, the relative risk of coronary death in the simvastatin group was 0.58 compared to the placebo group. That is, the ratio of the coronary death rate in the simvastatin group compared to the placebo group was 0.58. This means that the risk of coronary death was 42% less in the simvastatin group than in the placebo group. That 42% was derived as follows: assuming (as noted above) that the risk of coronary death is 1.0 in the placebo group, the risk of coronary death in the study was found to be 0.58 in the simvastatin group. Therefore, the risk reduction of coronary death in the simvastatin group is 1.0 (the risk in the placebo group) - 0.58 (the risk in the simvastatin group) = 0.42. Thus, a relative risk of 0.58 means that persons in the simvastatin group were 42% less likely to experience coronary death than those in the placebo group over the period of the study. Note, in the medical literature 'risk reduction' and 'percent reduction' are used interchangeably.

Can these numbers be calculated directly from just knowing the numbers of persons experiencing a particular event in each group, such as death due to cardiac event?

No, and here is why.

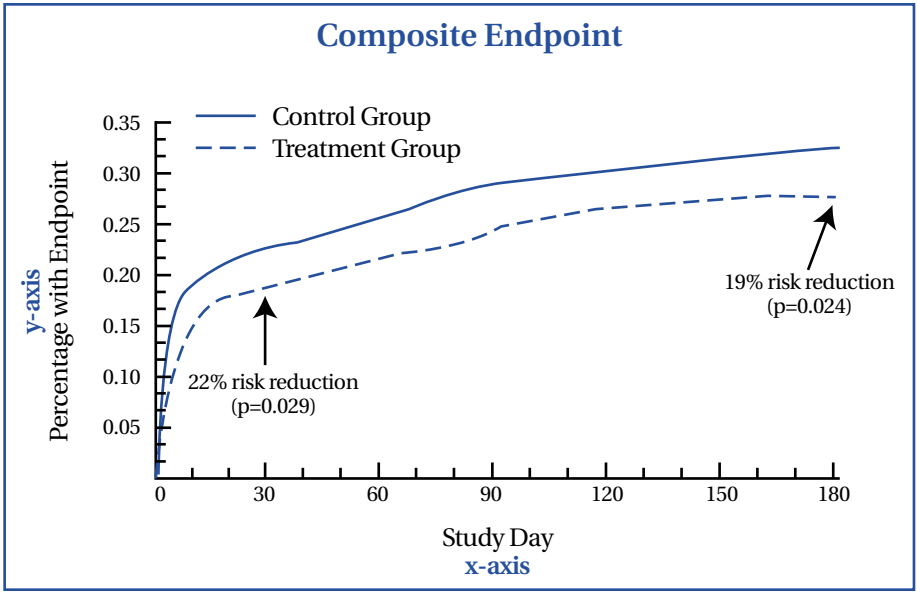
As you recall, there were 111 coronary deaths in the simvastatin group vs. 189 in the placebo group, with the direct calculation of the coronary death rate in each group shown below:

Simvastatin Group n=2221	**Placebo Group** n=2223
111/2221 = 5.00%	189/2223 = 8.50%

- The difference in the directly calculated coronary death rates between the two groups is 41.2% [(8.50 - 5.00) / 8.50], not 42% as quoted above! Why?
- This difference 41.2% vs. 42%, arises because all of the **'risk reductions'** reported in 4S were calculated using a very sophisticated statistical technique called **Cox regression analysis.** Cox regression analysis takes into account both when the deaths occurred as well as the total length of follow-up for all subjects. Many other methods of analysis usually just assume that a death (or other endpoint) occurred at the midpoint between two points in time when the subjects were studied. Therefore, Cox regression analysis produces more accurate estimates of the differences between the simvastatin and placebo groups than would a direct calculation of event rates between the two groups, as shown above.

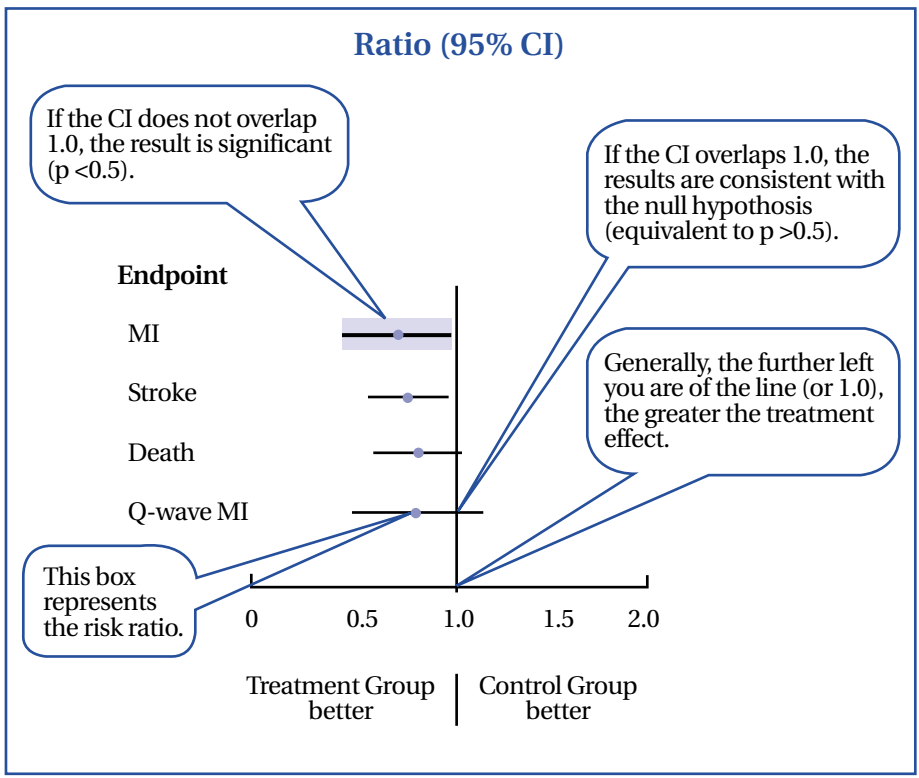
Again, Cox regression analysis is a statistical technique that accounts for the time of the event in order to get more valid measures of the ***risk ratio*** and ***relative risk reduction.*** Unfortunately, the calculation is complex, and can't be reproduced from the summary data typically reported in a journal article. It's for this reason that the relative risk reductions reported in studies using this technique differ somewhat from the simple calculations described above.

The ***Kaplan-Meier curves*** are estimates of the event rates in the treatment groups as functions of time. For any point in time represented on the x-axis, the y-axis gives the cumulative percentage of patients who have had a clinical event prior to that time.



Confidence Intervals

Confidence intervals (CI) tell us the "neighborhood" within which the true effect is likely to occur. Most often, the 95% confidence interval (CI) is provided. Basically, this range implies the true treatment effect of a drug, or RRR, has a 95% probability of lying within this range. Generally, the larger the sample size, the narrower the CI, since we have more confidence in our estimate of the true effect of the treatment.



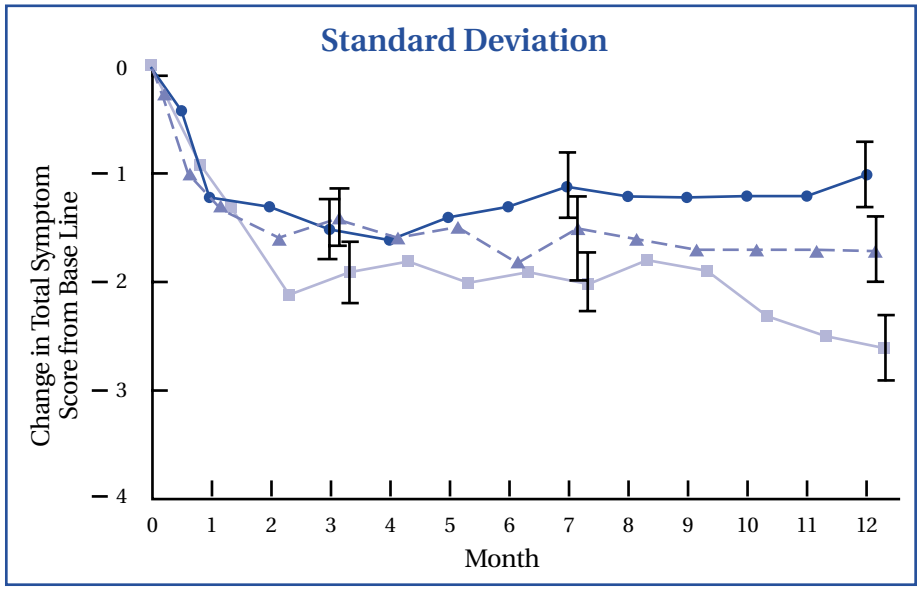
Variability of Data Around a Central Value and Standard Deviation

Once the pattern of clustering is determined, the variability of the data from the central value is examined using the range, the standard deviation, and the variance.

The **range** is a statement of the highest and lowest values of the set, and gives a quick analysis of how representative that mean or median is. For example, for a group for which the mean age is 50 years, if the range is 45-55 years, the group is all middle-aged, but if the range is 20-85 years, one cannot assume that the group is mostly middle-aged, or that a middle-aged group would be well-represented by the study group, even though the mean would tend to suggest that this might be the case.

The **standard deviation** is a specialized form of averaging the differences between each value and the mean. A small standard deviation indicates that the values are clustered fairly closely; a large standard deviation indicates that the values are widely scattered. Thus, standard deviation gives an impression of how much variability one might reasonably expect in a value.

Sometimes standard deviation results are illustrated visually. For example, in the figure below the standard deviations for mean change in total symptom score in men with benign prostatic hyperplasia are shown with "vertical bars" at three points during the course of therapy. In some cases these "vertical bars" may represent **standard error** and not standard deviation. Standard error is a measure of the precision with which the population mean is estimated.



The **variance** is the average of the squared deviations between the values and the means, and provides the same information as the standard deviation.

11. The standard deviation refers to the variability of a measurement for a given population. A (large, small) standard deviation indicates that values are clustered fairly closely.

t-Test, P Values, and Statistical Significance

The t-test is often used to determine whether the difference between means for two (or more), normally-distributed populations, such as the active-drug group and placebo-group, is significant. The t-test is (as are all statistical tests) an attempt to determine whether the difference or an even greater difference between two groups could have occurred due to chance alone. The "P-Value" which the t-test (or other statistical test) produces, is a calculation of the probability that such differences could have occurred due to chance alone. When "P" is less than 0.05, this means that there is less than a 5% chance that the differences observed between the study groups occurred due to chance alone. Therefore, we conclude that something other than chance (presumably the drug) may have caused the observed differences between groups.

T-tests are described as being one-tailed or two-tailed. One-tailed tests apply only to a change in one direction, while two-tailed tests can validate changes in either direction. Because it must be decided before the study is conducted whether the analysis will be one-tailed or two-tailed, the two-tailed analysis is commonly used so that changes in either direction can be tested for statistical significance.

12. True or False: The t-test, which produces a P value, is used to compare the difference between means for two or more populations, such as a treatment group and a placebo group.

Chance events can greatly affect the outcomes of research. The concept of **statistical significance** deals with the interpretation of the effects of chance on the results of a study. Statistical significance should not be confused with other uses of the word "significant"; an author referring to a "significant breakthrough" is probably not discussing statistical significance at that time.

A significance level of 0.05 is most commonly used in clinical studies to denote "statistical significance." However, it is important to understand that there is nothing "magical" about the 0.05 level for the P value. Again, when P=0.05, there is still a 5% chance that such differences could have arisen due to chance alone. Also, a researcher may decide that his/her results are "significant" if P=0.10 or some other number. The level selected for the P value will be determined by how sure the researcher wants to be that the differences observed in an experiment are unlikely to have occurred by chance alone.

13. What does P < 0.05 mean?

Clinical Significance

Clinical significance implies that the effects of treatment are large enough to have practical and useful implications. For example, assume that a study finds that a new agent produces a statistically significant increase in the median life-span of patients when compared to the traditional agent, but that this increase amounts to only one week. If the traditional agent is less expensive to administer and results in fewer side effects, it is unlikely that such a difference would be large enough to justify switching agents. Ideally, research results should show both statistical significance and clinical significance.

It is also important to understand that the reader can determine how much of a change is needed to be considered clinically significant, e.g., how much of a blood pressure change is needed for that change to be called "clinically significant." The reader determines the level of change needed for clinical significance using his/her own clinical experience, i.e., *statistics are NOT used to determine the magnitude of change needed for "clinical significance".*

14. True or False: Statistically significant and clinically significant are interchangeable terms.

Other Statistical Tests (For reference only)

Sign test

The sign test analyzes data in order to determine whether there is any change (positive or negative), in the variable of interest from time A to time B. The sign test does not take into account the magnitude of change, only whether there is a positive or negative change in a particular variable.

For example, if one were to study the effects of an antihypertensive agent on blood pressure at the time of dosing and 15 minutes, 30 minutes, 60 minutes and 90 minutes post-dose, one might apply the sign test to determine whether the blood pressure had increased or decreased at each time point.

Signed rank test

The signed rank test is similar to the sign test, but in addition to determining whether there is any change (positive or negative) in the variable, it takes into account the *magnitude* of the change in the variable of interest from time A to time B.

Using the above example, the signed rank test could be applied to measure *by how much* the blood pressure had increased or decreased at each time point.

F-test (Analysis of Variance, ANOVA)

The F-test is a ratio of variances, used when comparing several groups which cannot be compared with the t-test. Because it compares between-group differences with within-group differences, a value very close to 1.0 indicates that there is no real difference between groups, since the same magnitude of difference is seen within the groups. Like the t-test, this test also requires normally-distributed groups and cannot be applied when group selection has produced a badly skewed sample.

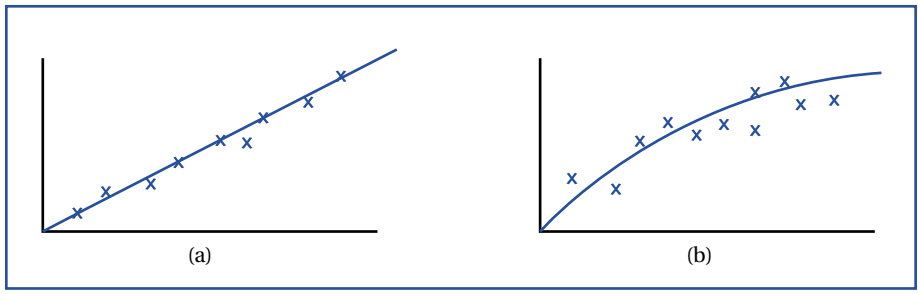
Chi Square Test (χ^2)

References to the chi-square are frequently made in the medical literature because this is the test that can be used when the sample is not normally distributed. The chi-square test is a test of distribution; it separates the results into categories and then compares the categories. It does not require a normally-distributed sample because the mean for the sample is not calculated or taken into

consideration at any time. This test is also good for taking a quick look at raw data to see if any pattern appears to be present. It is frequently used to analyze data from "2x2" tables, i.e., tables with two columns and two rows. This test is not considered as powerful as either the t-test or the F-test, so if the sample is normally-distributed, the chi-square test is not usually used.

Regression

Regression is a method of finding a relationship among data points by plotting the points on a graph, drawing a line which seems to fit them best, and using the equation for that line as the model for the problem's solution. Regression is useful for predicting values that fall outside the area studied, provided the right kind of line is chosen. Linear relationships provide fairly good predictions, but if the regression line is curved, predictions will be best when close to the area studied.



(In studies where there are large numbers of points, the regression line may be calculated by computer rather than be plotting all the points. The result is the same, an equation used to predict the probable location of other points.)

Specialized Types of Analyses

Subgroup Analysis

Subgroup analysis is the examination of certain subsets of a study population. While sometimes very helpful, subgroup analysis can also be misleading.

When patients are grouped for sub-analysis, it is important to ensure that their similarity, upon which the sub-analysis is based, is a characteristic which was present before randomization, for example,

age, sex, baseline ejection fraction of <30% *before* initiation of therapy with an ACE-inhibitor, etc. Potentially misleading data may arise when the patients are grouped according to a variable which was determined after randomization and could possibly be affected by the treatment being studied. An example of an improper subgroup analysis would be evaluation of the effects of lovastatin on atherosclerotic progression in patients whose degree of arterial stenosis was first determined *after* initiation of study drug therapy.

Very large-scale, well designed, randomized trials usually have a sample size with enough power (or ability to ensure a high probability) to discover clinically important differences among patient groups; however, smaller trials which are not randomized would not be likely candidates for subgroup analysis. It is usually best for the investigator to identify, in advance, the criteria for any subgroup analysis that might be performed upon completion of the main study. Consistency of the data subgroup analysis with data from other studies is important, as is the suggestion of a difference coming from patients *within* a particular study and not from patients in different studies.

Meta-Analysis

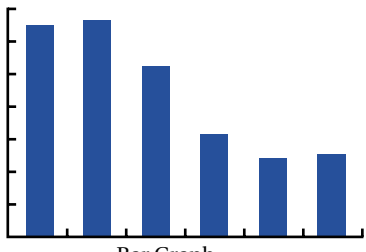
A **meta-analysis** combines and analyzes data from existing (and usually published) clinical studies. The purpose of most meta-analyses is to either evaluate the therapeutic effectiveness of a particular drug or procedure, or to develop plans for new studies which might be performed on a particular topic. Before performing a meta-analysis, the researcher must decide the means by which data will be collected (computer or manual search, help of a specialist in medical information), the inclusion criteria for the analysis, i.e., study sample size, study design (e.g. only randomized trials), patient population, etc. If a researcher wanted to do a meta-analysis on the safety of a particular drug, one of the inclusion criteria would be studies which specifically reported adverse clinical and laboratory effects. The compilation and analysis of the combined data must be performed with the same care as one would take when performing an original study, as potential problems such as bias can taint the conclusion. As with an original research study, numerous statistical tests may be applied to a meta-analysis to test its results.

Reading Tables

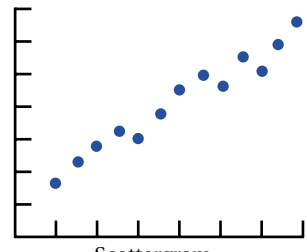
In many cases, much of the information in the Results section of a research article is presented in the form of tables. Tables have the advantage of providing a large amount of information in a small space. While it is easy to think of the data, or body, as being the "meat" of the table, careful reading of the information in the various labels (title, columns, rows, and notes) is very important to understanding what is being conveyed by the table. *(It should be noted that in some cases, this is not obvious, and a quick glance at the table may leave one with a mistaken impression of its contents.)*

Reading Graphs

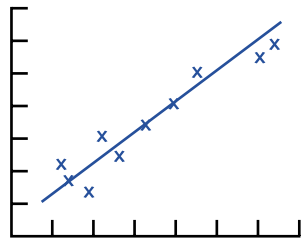
Graphs are another efficient means for communicating pertinent information from a study. They may describe the distribution of data collected in an experiment, illustrate the changes in a measurement across time, and portray differences or relationships existing between treatment groups, which might otherwise be difficult to communicate effectively. Different kinds of graphs are used to display different kinds of information. Some data are displayed to best advantage in bar graphs; other data lend themselves to scattergrams or line graphs. Examples of different types of graphs follow.



Bar Graph



Scattergram



Line Graph

As with a table,

- Title or caption
- Labels for x-axis, y-axis
- Legend

are important in identifying the information conveyed by the curve.

Generally, the **title** or captions of a graph will tell what the variables are and the relationship represented in the graph.

Most graphs consist of two perpendicular axes, each one representing the values of a different variable. The vertical axis (the ordinate, or **y-axis**) is used most frequently to represent the variable that the researcher is interested in observing (results), while the horizontal axis (the abscissa, or **x-axis**) generally represents the variable that is being controlled by the researcher, or else may represent a variable such as the passage of time. The units of measurement associated with each variable should be included on the axes.

The **legend** explains the labeling of points, or curves, on the graph; it may appear as a small box in an open space on the graph, or it may be included in the caption.

Clinical Efficacy

Efficacy of a treatment regimen is assessed statistically to determine whether the results are likely or unlikely to be due to chance alone. If the study presents clinically significant findings and the patient population is small, it may be repeated by other investigators to determine its validity. If it is a large, multicenter study, there may not be the resources available to repeat it. Nevertheless, if the study was well-researched, results from the study should be similar to results of such a study in the entire population.

Clinical Safety

Often, the end of the Results section includes information about the side effects experienced by the subjects. Depending on the purpose of the research, the side effects may be described in great detail or provided as a casual list. They are commonly classified by the body system affected, such as cardiac, gastrointestinal, hepatic, etc., and are measured as the total number of side effects/adverse experiences,

and/or the number or percentage of patients with a particular adverse event. The length of time the patient(s) took the drug before the onset of the adverse event, the severity, and whether the investigator assessed the adverse event as not drug related or as possibly, probably or definitely drug related are also described. The researchers usually note whether an adverse event required withdrawal from the study, dosage adjustment, or whether the event resolved over time (either spontaneously or due to withdrawal).

Adverse events classified as "serious" are usually those requiring hospitalization or prolonged hospitalization; those which are immediately life-threatening, cause either permanent substantial disability, cancer, congenital anomaly, or death; or those which are the result of an overdose.

Although the descriptions of side effects vary greatly from article to article, the kind of information typically covered includes: total number of side effects, the number (or percentage) of patients in whom they occurred, the length of time the patients took the drug before adverse reactions occurred, and the type and severity of side effects. With a new drug, the discussion of side effects may be a fairly general listing because experience has not been sufficient to indicate which side effects may be common problems. With a more established drug, the researchers may carefully look for and comment on side effects which have been frequently noted in previous studies.

15. Efficacy data typically appears in the _____ section of an article. Safety data typically appears at the end of the _____ section of a research article.

The Discussion Section

There are six kinds of information to look for in the Discussion section, although they will not necessarily all appear in a given article.

- A **summary** of the entire study, with the major purposes, methods, and results briefly retold.
- Suggestions for **future investigation,** such as which avenues the authors feel are promising, and what is still unknown or unclear about this line of research/treatment.
- **Comparisons** of this study to other studies, whether it confirmed or differed with the results of other studies.
- **Limitations** of the present study, which make it possible to generalize the conclusions only to certain groups or conditions.
- **Unplanned findings,** which cannot be confirmed by this study but which appear to be worthy of follow-up in other studies.
- The **bottom line,** or the authors' interpretation of the results. (Occasionally, authors interpret results to mean more than they do; other researchers may judge for themselves whether the conclusions are warranted.)

All six points may not appear in any one discussion section, and a single sentence can summarize, compare, and give the bottom line.

16. True or False: The discussion section of an article may contain a summary of the study and the main conclusions, but would not include the author's interpretation of the study.

Summary

- The **Citation** of a research article should be examined critically for information about the authors, title, and journal.
- **A research article typically has 5 sections** (Summary or Abstract, Introduction, Materials and Methods, Results, Discussion).
- The **Summary or Abstract** should be read first; a well-written summary will include the study design, number of patients, disease being treated, treatment, results, and conclusion.
- The **Introduction** often describes the problem, that is, why the researchers conducted the study. The last sentence of the introduction often contains the study objective.
- The **Materials and Methods** section contains the protocol or study design. A protocol identifies the patient population, the treatment, how results were measured, and how bias was controlled. Bias can be minimized by procedures such as randomization and double-blind distribution of medication. The larger the patient population, the less chance for bias. An especially large patient population is necessary in some studies that have events such as death or myocardial infarction as endpoints. A study with a large number of patients generally has high statistical power. Protocol also involves issues of data analysis, including whether to use a *per protocol* approach (patients conforming to protocol) or *all patients* or *intention to treat* approach (all patients entered into the study).
- The **Results** section typically describes the efficacy and safety of the treatment. Statistical analysis of the collected data often includes a summary of values, e.g , the mean; the distribution of values, e.g., normal distribution; and the variability of values, e.g., the standard deviation, with a smaller standard deviation indicating less variability around a central value. The difference between two study populations, e.g., drug-treated and placebo-treated, is often analyzed using the t-test or other statistical test to derive P values and determine statistical significance. A P value of 0.05 is often chosen as a cutoff to determine significance. A P value of less than 0.05 is interpreted as, "The likelihood is less than 5% that the difference in results could have occurred by chance alone." A study may show statistical significance, but not have clinical significance, if the data don't have practical implications.

- The **Discussion** may include a summary of the study, comparisons to other studies, the author's interpretation of the findings, and the conclusion. The **conclusion** often appears as the last sentence of the Discussion section.

Review Questions

1. **True or False:** The editorial review process is basically the same for all articles published in medical journals, regardless of the journal.
2. Typically, a research article has 5 sections. Please name them.
3. **True or False:** A well-written summary or abstract section provides the reader with information about study design, results, and conclusions.
4. **True or False:** The purpose or objectives of a study are usually found in the last sentence of the Introduction section.
5. Protocol, another name for _____ _____ , describes the patient population, the treatments, how results were measured, and how bias was controlled.
6. _____ is a statistical concept referring to effects that produce results different from the true values.
7. In a single-blind study the _____ does not know whether the medication is drug or placebo; in a double-blind study both the patient and the _____ do not know whether the medication is drug or placebo
8. **True or False:** Power refers to the ability of a study to detect differences in the occurrence of a given endpoint, such as myocardial infarction or death in the treatment group. The larger the sample size, the greater the power of a study.
9. **True or False:** Per Protocol analysis involves analyzing data from all patients entered into a study protocol.
10. **True or False:** Calculations of mean, median, and mode are ways that data can be summarized to represent a group of values.
11. The standard deviation refers to the variability of a measurement for a given population. A (large, small) standard deviation indicates that values are clustered fairly closely.

12. **True or False:** The t-test, which produces a P value, is used to compare the difference between means for two or more populations, such as a treatment group and a placebo group.

13. What does $P < 0.05$ mean?

14. **True or False:** Statistically significant and clinically significant are interchangeable terms.

15. Efficacy data typically appears in the _____ section of an article. Safety data typically appears at the end of the _____ section of a research article.

16. **True or False:** The discussion section of an article may contain a summary of the study and the main conclusions, but would not include the author's interpretation of the study.

Answers to Review Questions

1. **True or False:** The editorial review process is basically the same for all articles published in medical journals, regardless of the journal. *(False)*
2. Typically, a research article has 5 sections. Please name them. *(Summary or abstract, Introduction, Materials and Methods, Results, Discussion)*
3. **True or False:** A well-written summary or abstract section provides the reader with information about study design, results, and conclusions. *(True)*
4. **True or False:** The purpose or objectives of a study are usually found in the last sentence of the Introduction section. *(True)*
5. Protocol, another name for _____ _____ , describes the patient population, the treatments, how results were measured, and how bias was controlled. *(study design)*
6. _____ is a statistical concept referring to effects that produce results different from the true values. *(Bias)*
7. In a single-blind study the _____ does not know whether the medication is drug or placebo; in a double-blind study both the patient and the _____ do not know whether the medication is drug or placebo *(patient)* *(physician)*
8. **True or False:** Power refers to the ability of a study to detect differences in the occurrence of a given endpoint, such as myocardial infarction or death in the treatment group. The larger the sample size, the greater the power of a study. *(True)*
9. **True or False:** Per Protocol analysis involves analyzing data from all patients entered into a study protocol. *(False)*
10. **True or False:** Calculations of mean, median, and mode are ways that data can be summarized to represent a group of values. *(True)*
11. The standard deviation refers to the variability of a measurement for a given population. A (large, small) standard deviation indicates that values are clustered fairly closely. *(small)*

12. **True or False:** The t-test, which produces a P value, is used to compare the difference between means for two or more populations, such as a treatment group and a placebo group. *(True)*

13. What does P < 0.05 mean? *(The likelihood is less than 5% that the difference in results could have occurred by chance alone.)*

14. **True or False:** Statistically significant and clinically significant are interchangeable terms. *(False)*

15. Efficacy data typically appears in the _____ section of an article. Safety data typically appears at the end of the _____ section of a research article. *(results) (results)*

16. **True or False:** The discussion section of an article may contain a summary of the study and the main conclusions, but would not include the author's interpretation of the study. *(False)*

Glossary of Statistical Terms*

abscissa: horizontal or x-axis of a graph.

absolute risk: probability of an event occurring during a specified period. The absolute risk equals the relative risk times the average probability of the event during the same time, if the risk factor is absent.[23(p327)] See `absolute risk reduction`.

absolute risk reduction: proportion in the control group experiencing an event minus the proportion in the intervention group experiencing an event. The inverse of the absolute risk reduction is the `number needed to treat`. See `absolute risk`.

accuracy: ability of a test to produce results that are close to the true measure of the phenomenon.[23(p327)] Generally, assessing accuracy of a test requires that there be a `criterion standard` with which to compare the test results. Accuracy encompasses a number of measures including `reliability`, `validity`, and lack of `bias`.

actuarial life-table method: see `life table`, `Cutler-Ederer method`.

adjustment: techniques used after the collection of data to control for the effect of known or potential confounding variables.[23(p327)] A typical example is adjusting a result for the independent effect of age of the subjects (age is the `independent variable`).

aggregate data: data accumulated from disparate sources.

agreement: statistical test performed to determine the equivalence of the results obtained by 2 tests when 1 test is compared with another (of which 1 is usually but not always a `criterion standard`).

⇒Agreement should not be confused with `correlation`. Correlation is used to test whether 2 variables are interdependent, whereas agreement tests whether 2 variables are equivalent. For example, an investigator compares results obtained by 2 methods of measuring hematocrit. Method A gives a result that is exactly twice that of method B. The correlation between A and B is perfect since A is always twice B, but the agreement is very poor; method A is not equivalent to method B (written communication, George W. Brown, MD, September 1993). One appropriate way to assess agreement has been described by Bland and Altman.[24]

algorithm: systematic process that consists of an ordered sequence of steps; each step depends on the previous step.[25(p6)] An algorithm may be used clinically to guide treatment decisions for an individual patient on the basis of the patient's clinical outcome or result.

α (alpha), α level: size of the likelihood acceptable to the investigators that a relationship observed between 2 variables is due to chance (the probability of a `type I error`); usually $\alpha = .05$. If $\alpha = .05$, $P \leq .05$ will be considered significant.

analysis: process of mathematically summarizing and comparing data to confirm or refute a `hypothesis`. Analysis serves 3 functions: (1) to test hypotheses regarding differences in large populations based on samples of the populations, (2)

to control for confounding variables, and (3) to measure the size of differences between groups or the strength of the relationship between variables in the study.[23(p25)]

analysis of covariance (ANCOVA): statistical test used to examine data that include both continuous and nominal independent variables and a continuous dependent variable. It is basically a hybrid of multiple regression (used for continuous independent variables) and analysis of variance (used for nominal independent variables).[23(p299)]

analysis of residuals: see linear regression.

analysis of variance (ANOVA): statistical method used to compare a continuous dependent variable and more than one nominal independent variable. The null hypothesis in ANOVA is tested by means of the F test.

The most commonly used type of ANOVA is the 1-way ANOVA, in which there are more than 2 mutually exclusive categories for a nominal independent variable (eg, systolic blood pressure for the continuous variable and race/ethnicity as the nominal variable, categorized as non-Hispanic black, non-Hispanic white, or Hispanic). If there are 2 mutually exclusive categories for the nominal independent variable (eg, Hispanic or non-Hispanic), the 1-way ANOVA is equivalent to the *t* test.

A 2-way ANOVA is used if there are 2 categorical variables with a continuous variable (eg, systolic blood pressure for the continuous variable, with race/ethnicity categorized as non-Hispanic white, non-Hispanic black, or Hispanic, and age categorized as 20-40 years, 40-60 years, and 60 years and older as the 2 categorical variables). Three- and 4-way ANOVAs may also be performed.

If more than 1 nonexclusive independent variable is analyzed (eg, race and sex in addition to systolic blood pressure), the process is called factorial ANOVA. (An analysis of main effects in this analysis would assess the independent effects of either race or sex; an association between female sex and systolic blood pressure that exists in one race but not another would mean that an interaction between race and sex exists.)

If repeated measures are made on an individual (such as measuring blood pressure over time) so that a matched form of analysis is appropriate, but potentially confounding factors (such as age) are to be controlled for simultaneously, repeated-measures ANOVA is used. Randomized-block ANOVA is used if treatments are assigned by means of block randomization.[23(pp291-295)]

⇒ANOVA can establish only whether a significant difference exists among groups, not which groups are significantly different from each other. To determine which groups differ significantly, a pairwise analysis of a continuous dependent variable and more than 1 nominal variable is performed by the Newman-Keuls test or Tukey test. These multiple comparisons procedures avoid the potential of a type I error that might occur if the *t* test were applied at this stage.

⇒The F ratio is the statistical result of ANOVA and is a number between 1 and infinity. The F ratio is compared with tables of the F distribution, taking into account the α level and degrees of freedom (df) for the numerator and denominator, to determine the P value.

The *df* are provided along with the F statistic. The first subscript (2) is the *df* for the numerator; the second subscript (63) is the *df* for the denominator. The *P* value is obtained from an F statistic table that provides the *P* value that corresponds to a given F and *df*. Because ANOVA does not determine which groups are significantly different from each other, this example would normally be accompanied by the results of the multiple comparisons procedure.[26] Other models such as Latin square may also be used.

ANCOVA: see abbreviation for analysis of covariance.

ANOVA: see abbreviation for analysis of variance.

Ansari-Bradley dispersion test: rank test to determine whether 2 distributions known to be of identical shape (but not necessarily of normal distribution) have equal parameters of scale.[22(p6)]

area under the curve (AUC): technique used to measure the performance of a test plotted on a receiver operating characteristic (ROC) curve or to measure drug clearance in pharmacokinetic studies.[27(p12)] When measuring test performance, the larger the AUC, the better the test performance. When measuring drug clearance, the AUC assesses the total exposure of the individual, as measured by levels of the drug in blood or urine, to a drug over time. The curve of drug clearance used to calculate the AUC is also used to calculate the drug half-life.

⇒ The method used to determine the AUC should be specified (eg, the trapezoidal rule).

artifact: difference or change in measure of occurrence of a condition that results from the way the disease or condition is measured, sought, or defined.[23(p327)]

> *Example:* An artifactual increase in the incidence of acquired immunodeficiency syndrome (AIDS) was expected because the definition of AIDS was changed to include a larger number of AIDS-defining illnesses.

assessment: in the statistical sense, evaluating the outcome(s) of the study and control groups.

assignment: process of distributing individuals to study and control groups. See also randomization.

association: statistically significant relationship between 2 variables in which one does not necessarily cause the other. When 2 variables are measured simultaneously, association rather than causation generally is all that can be assessed.

> *Example:* After controlling for confounding factors by means of multivariate regression, a significant association remained between age and disease prevalence.

attributable risk: disease that can be attributed to a given risk factor; conversely, if the risk factor were eliminated entirely, the amount of the disease that could be eliminated.[23(pp327-328)] Attributable risk assumes a causal relationship (ie, the factor to be eliminated is a cause of the disease and not merely associated with the disease). An attributable risk of 1 indicates that the factor does not contribute, an

attributable risk of less than 1 indicates that the factor reduces risk, and an attributable risk of greater than 1 indicates that the factor increases risk. See attributable risk percentage and attributable risk reduction.

attributable risk percentage: the percentage of risk associated with a given factor among those with the risk factor.[23(pp327-328)] For example, risk of stroke in an older person who smokes and has hypertension and no other risk factors can be divided among the risks attributable to smoking, hypertension, and age. Attributable risk percentage is often determined for a population and is the percentage of the disease related to the risk factor. See population attributable risk percentage.

attributable risk reduction: the number of events that can be prevented by eliminating a particular risk factor from the population. Attributable risk reduction is a function of 2 factors: the strength of the association between the risk factor and the disease (ie, how often the risk factor causes the disease) and the frequency of the risk factor in the population (ie, a common risk factor may have a lower attributable risk in an individual than a less common risk factor, but could have a higher attributable risk reduction because of the risk factor's high prevalence in the population). Attributable risk reduction is a useful concept for public health decisions. The inverse of attributable risk reduction is the number needed to treat, a number more useful for clinical practice. See also attributable risk.

average: sum of all measurements divided by the total number of measurements. Mathematically synonymous with mean, but mean is preferred since the term *average* is often used loosely.

Bayesian analysis: theory of statistics involving the concept of prior probability, conditional probability or likelihood, and posterior probability.[22(p16)] For interpreting studies, the prior probability is based on previous studies and may be informative, or, if none exists or those that exist are not useful, one may assume a uniform prior. The study results are then incorporated with the prior probability to obtain a posterior probability. Bayesian analysis can be used to interpret how likely it is that a positive result indicates presence of a disease, by incorporating the prevalence of the disease in the population under study and the sensitivity and specificity of the test in the calculation.

⇒Bayesian analysis has been criticized because the weight a particular study is given when prior probability is calculated can be a subjective decision, but the process most closely approximates how studies are considered when they are incorporated into clinical practice. When Bayesian analysis is used to assess posterior probability for an individual patient in a clinic population, the process may be less subjective than usual practice because the prior probability, equal to the prevalence of the disease in the clinic population, is more accurate than if the prevalence for the population at large were used.[20]

β (beta), β level: probability of showing no significant difference when a true difference exists; a false acceptance of the null hypothesis.[25(p57)] $1 - \beta$ is the statistical power of the test to detect a true difference, so the smaller the β, the greater the power. A value of .2 for β is equal to .8 or 80% power. A β of .1 or .2 is most frequently used in power calculations. The β error is synonymous with type II error.[26]

bias: situation or condition that causes a result to depart from the true value in a consistent direction. Bias refers to defects in study design (often selection bias) or measurement.[23(p328)] One method to reduce measurement bias is to ensure that the investigator measuring outcomes for a subject is unaware of the group to which the subject belongs (ie, blinded assessment).

bimodal distribution: nonnormal distribution with 2 peaks, or modes. The mean and median may be equivalent, but neither will describe the data accurately.

binary variable: variable that has 2 mutually exclusive subgroups, such as male/female or pregnant/not pregnant; synonym for dichotomous variable.[27(p75)]

binomial distribution: probability with 2 possible mutually exclusive outcomes; used for modeling cumulative incidence and prevalence rates[25(p17)] (for example, the probability of a person having a stroke in a given population over a given period; the outcome must be stroke or no stroke).

biological plausibility: evidence that an independent variable can be expected to exert a biological effect on a dependent variable with which it is associated. For example, studies in animals were used to establish the biological plausibility of adverse effects of passive smoking.

bivariable analysis: see bivariate analysis.

bivariate analysis: used when 1 dependent and 1 independent variable are to be assessed.[23(p263)] Common examples include the t test for 1 continuous variable and 1 binary variable and χ^2 test for 2 binary variables. Bivariate analyses can be used for hypothesis testing in which only 1 independent variable is taken into account, to compare baseline characteristics of 2 groups, or to develop a model for multivariate regression. See also univariate and multivariate analysis.

⇒Bivariate analysis is the simplest form of hypothesis testing but is often used incorrectly, either because it is used too frequently, resulting in an increased likelihood of a type I error, or because tests that assume a normal distribution (eg, the t test) are applied to nonnormally distributed data.

Bland-Altman plot: a method to assess agreement (eg, between 2 tests) developed by Bland and Altman.[24]

blinded (masked) assessment: evaluation or categorization of an outcome in which the person assessing the outcome is unaware of the treatment assignment. Masked assessment is the term preferred by some investigators and journals, particularly those in ophthalmology.

⇒ Blinded assessment is important to prevent bias on the part of the investigator performing the assessment, who may be influenced by the study question and consciously or unconsciously expect a certain test result.

blinded (masked) assignment: assignment of individuals participating in a prospective study (usually random) to a study group and a control group without the investigator or the subjects being aware of the group to which they are assigned. Studies may be single-blind, in which either the subject or the investigator does not know the treatment assignment, or double-blind, in which neither knows the

treatment assignment. The term masked assignment is preferred by some investigators and journals, particularly those in ophthalmology.

block randomization: type of randomization in which the unit of randomization is not the individual but a larger group, sometimes stratified on particular variables such as age or severity of illness to ensure even distribution of the variable between randomized groups.

Bonferroni adjustment: statistical adjustment applied when multiple comparisons are made. The α level (usually .05) is divided by the number of comparisons to determine the α level that will be considered statistically significant. Thus if 10 comparisons are made, and α of .05 would become $\alpha = .005$ for the study. Alternatively, the P value may be multiplied by the number of comparisons, while retaining the α of .05.[27(pp31-32)] Alternatively, the P value may be multiplied by the number of comparisons, while retaining the α of .05. For example, a P value of .02 obtained for 1 of 10 comparisons would be multiplied by 10 to get the final result of $P = .20$, a nonsignificant result.

⇒The Bonferroni test is a conservative adjustment for large numbers of comparisons (ie, less likely than other methods to give a significant result) but is simple and used frequently.

bootstrap method: statistical method for validating a new diagnostic parameter in the same group from which the parameter was derived. Thus, the validation of the method is based on a simulated sample, rather than a new sample. The parameter is first derived from the entire group, then applied sequentially to subsegments of the group to see if the parameter performs as well for the subgroups as it does for the entire group (derived from "pulling oneself up by one's own bootstraps").[27(p32)]

For example, a number of prognostic indicators are measured in a cohort of hospitalized patients to predict mortality. To determine if the model using the indicators is equally predictive of mortality for subsegments of the group, the bootstrap method is applied to the subsegments and confidence intervals are calculated to determine the predictive ability of the model. The jackknife dispersion test also uses the same sample for both derivation and validation.

⇒ Although the preferable means for validating a model is to apply the model to a new sample (eg, a new cohort of hospitalized patients in the example listed), the bootstrap method can be used to reduce the time, effort, and expense necessary to complete the study. However, the bootstrap method provides less assurance than validation in a new sample that the model is generalizable to another population.

Brown-Mood procedure: test used with a regression model that does not assume normally distributed data or common variance of the errors.[22(p26)] It is an extension of the median test.

case: individual with the outcome or disease of interest.

case-control study: retrospective study in which subjects with the disease (cases) are compared with those who do not have the disease (controls). Cases and controls are identified without knowledge of exposure to the risk factors under study. Cases and controls are matched on certain important variables, such as age, sex, and year in which the individual was treated or identified. A case-

control study conducted withinin a cohort study is referred to as a *nested case-control study*.[25(p111)] This type of case-control study may be an especially strong study design if characteristics of the cohort have been carefully ascertained. See also 17.2.4, Case-Control Study.

⇒ Cases and controls should be selected from the same population to minimize confounding by factors other than those under study. Matching cases and controls on too many characteristics may obscure the association of interest, since if cases and controls are too similar, their exposures may be too similar to detect a difference (see overmatching).

case fatality rate: probability of death among people diagnosed as having a disease. The rate is calculated as the number of deaths during a specific period divided by the number of persons with the disease at the beginning of the period.[27(p38)]

case series: retrospective descriptive study in which clinical experience with a number of patients is described. See 17.2.6, Case Series.

categorical data: counts of members of a category or class; for the analysis each member or item should fit into only 1 category or class[22(p29)] (eg, sex or race/ethnicity). The categories have no numerical significance. Categorical data are summarized by means of proportions, percentages, fractions, or simple counts. Categorical data is synonymous with nominal data.

cause, causation: something that brings about an effect or result. To be distinguished from association, especially in cohort studies. To establish something as a cause it must be known to precede the effect. The concept of causation includes the contributing cause, the direct cause, and the indirect cause.

censored data: for continuous data, defining a cutoff for measuring or reporting data. Censoring data reduces the problem of extreme outliers skewing distribution of the data and is also used for individuals for whom the final outcome is not known, such as in survival analyses for individuals who have not experienced the outcome (usually death) at the time the analysis is conducted. The term left-censored data means that data were censored from the low end or left of the distribution; right-censored data come from the high end or right of the distribution[25(p26)] (eg, in survival analyses). For example, if data for falls are categorized as individuals who have 0, 1, or 2 or more falls, falls exceeding 2 have been right-censored.

central limit theorem: theorem that states that the mean of a number of samples with variances that are not large relative to the entire sample will increasingly approximate a normal distribution as the sample size increases. This is the basis for the importance of the normal distribution in statistical testing.[22(p30)]

central tendency: property of the distribution of data, usually measured by mean, median, or mode.[25(p41)]

χ^2 test (chi-square test): a test of significance based on the χ^2 statistic, usually used for categorical data. The observed values are compared with the expected values under the assumption of no association. The χ^2 goodness-of-

fit test compares the observed with expected frequencies. The χ^2 test can also compare an observed variance with hypothetical variance in normally distributed samples.[22(p33)] In the case of a continuous independent variable and a nominal dependent variable, the χ^2 test for trend can be used to determine whether a linear relationship exists (for example, the relationship between systolic blood pressure and stroke).[23(p284-285)]

⇒ The *P* value is determined from χ^2 tables with the use of the specified α level and the *df* calculated from the number of cells in the χ^2 table. The χ^2 statistic should be reported to no more than 1 decimal place; if Yates correction was used, that should be specified. See also contingency table.

> *Example:* The exercise intervention group was least likely to have suffered a fall in the previous month ($\chi^2_3 = 17.7$, $P = .02$).

Note that the *df* for χ^2 (subscript 3) is specified; it is derived from the number of cells in the χ^2 table (for this example, 4 cells in a 2×2 table). The value 17.7 is the χ^2 value. The *P* value is determined from the χ^2 value and *df*.

chloroplethic map: map of a region or country that uses shading to display quantitative data.[25(p28)] See also 2.14, Manuscript Preparation, Figures.

chunk sample: subset of a population selected for convenience without regard to whether the sample is random or representative of the population.[22(p32)] A synonym is convenience sample.

Cochran *Q* test: method used to compare percentage results in matched samples, often used to test whether the observations made by 2 observers vary in a systematic manner. The analysis results in a *Q* statistic, which, with the *df*, determines the *P* value; if significant, the variation between the 2 observers cannot be explained by chance alone.[22(p25)] See also interobserver bias.

coefficient of determination: square of the correlation coefficient, used in linear or multiple regression analysis. This statistic indicates the proportion of the variation of the dependent variable that is explained by the independent variable.[23(p328)] If the analysis is bivariate, the correlation coefficient is *r* and the coefficient of determination is r^2. If the correlation coefficient is derived from multivariate analysis, the correlation coefficient is *R* and the coefficient of determination is R^2. See also correlation coefficient.

> *Example:* The sum of the R^2 values for age and body mass index was 0.23. [Twenty-three percent of the variance could be explained by those 2 variables.]

⇒ When R^2 values of the same dependent variable total more than 1.0 or 100%, then the independent variables have an interactive effect on the dependent variable.

coefficient of variation: ratio of the standard deviation [SD] to the mean. The coefficient of variation is expressed as a percentage and is used to compare dispersions of different samples. The smaller the coefficient of variation, the greater the precision.[26] The coefficient of variation is also used when the SD is dependent on the mean; eg, the increase in height with age is accompanied by an increasing SD of height in the population.

cohort: term to describe a group of individuals who share a common exposure, experience, or characteristic, or a group of individuals followed up or traced over time in a cohort study.[25(p31)]

cohort effect: change in rates that can be explained by the common experience or characteristic of a group or cohort of individuals. A cohort effect implies that a current pattern of variables may not be generalizable to a different cohort.[23(p328)]

Example: The decline in socioeconomic status with age was a cohort effect explained by fewer years of education among the older individuals.

cohort study: study of a group of individuals, some of whom are exposed to a variable of interest (eg, a drug treatment or environmental exposure), in which subjects are followed up over time to determine who develops the outcome of interest and whether the outcome is associated with the exposure. Cohort studies may be concurrent (prospective) or nonconcurrent (retrospective).[23(pp328-329)] See also 17.2.3, Cohort Study.

⇒ Whenever possible, a subject's outcome should be assessed by an individual(s) without knowledge of whether the subject was exposed (see `blinded assessment`).

concordant pair: pair in which both individuals have the same trait or outcome (as opposed to `discordant pair`). Used frequently in twin studies.[25(p35)]

conditional probability: `probability` that an event E will occur given the occurrence of F, called the conditional probability of E given F. The reciprocal is not necessarily true: the probability of E given F may not be equal to F given E.[27(p55)]

confidence interval (CI): range of numerical expressions within which one can be confident (usually 95% confident, to correspond to an α `level` of .05) the population value the study is intended to estimate lies.[23(p329)] The CI is an indication of the precision of an estimated population value.

⇒ Confidence intervals used to estimate a population value usually are symmetric or nearly symmetric around a value, but CIs used for relative risks and odds ratios may not be. Confidence intervals are preferable to `P values` since they convey information about precision as well as statistical significance. If the CI does not overlap 1, the result is significant ($P < .05$); if the CI overlaps 1, the results are consistent with the null hypothesis (equivalent to $P > .05$). If a CI value equals 1, then generally $P = .05$. In all cases, the point estimate should be contained within the CI (although if the CIs are very close to the point estimate, the rounded-off CI may be identical to the point estimate).

⇒ Confidence intervals are expressed with *to* or a hyphen separating the 2 values. To avoid confusion, hyphens are not used if 1 of the values is a negative number. Units that are closed up with the numeral are repeated for each CI; those not closed up are repeated only with the last numeral. See also 17.3, Significant Digits and Rounding Numbers, and 16.4, Numbers and Percentages, Use of Digit Spans and Hyphens.

Example: The odds ratio was 3.1 (95% CI, 2.2-4.8). The prevalence of disease in the population was 1.2% (95% CI, 0.8%-1.6%).

confidence limits (CLs): upper and lower boundaries of the `confidence interval`, expressed with a comma separating the 2 values.[25(p35)]

Example: The mean (95% confidence limits) was 30% (28%, 32%).

confounding: confounding has 3 possible meanings when used in a statistical sense: (1) the apparent effect of an exposure on risk is caused by an association

with other factors that can influence the outcome; (2) the effects of 2 or more causal factors as observed by a set of data cannot be separated to identify the cause of any single causal factor; (3) the measure of the effect of an exposure on risk is distorted because of the association of exposure with another factor(s) that influences the outcome under study.[25(p35)] See also confounding variable.

confounding variable: variable that can cause or prevent the outcome of interest, is not an intermediate variable and is associated with the factor under investigation. Unless it is possible to adjust for confounding variables, their effects cannot be distinguished from those of the factors being studied. Bias can occur when adjustment is made for any factor that is caused in part by the exposure and also is correlated with the outcome.[25(p35)] Multivariate analysis is used to control the effects of confounding variables that have been measured.

contingency coefficient: the coefficient, *C*, is used to measure the strength of association between 2 characteristics in a contingency table.[27(pp56-57)]

contingency table: table created when categorical variables are used to calculate expected frequencies in an analysis and to present data, especially for a χ^2 test (2-dimensional data) or log-linear models (data with at least 3 dimensions). A 2×3 contingency table has 2 rows and 3 columns. The *df* are calculated as (number of rows − 1)(number of columns − 1). Thus, a 2×3 contingency table has 6 cells and 2 *df*.

continuous data: data with an unlimited number of equally spaced values[23(p329)] (eg, weight, systolic blood pressure, cholesterol), as opposed to categorical, nominal, ordinal, or dichotomous data. Parametric statistics require that continuous data have a normal distribution.

contributory cause: independent variable (cause) that is thought to contribute to the occurrence of the dependent variable (effect). That a cause is contributory should not be assumed unless all of the following have been established: (1) an association exists between the putative cause and effect, (2) the cause precedes the effect in time, and (3) altering the cause alters the probability of occurrence of the effect.[23(p329)] Other factors that may contribute to establishing a contributory cause include the concept of biological plausibility, the existence of a dose-response relationship, and consistency of the relationship when evaluated in different settings.

control: in a case-control study, the designation for an individual without the disease or outcome of interest; in a cohort study, the individuals not exposed to the independent variable of interest; in a randomized controlled trial, the group receiving a placebo or standard treatment rather than the intervention under study.

controlled clinical trial: study in which a group receiving an experimental treatment is compared with a control group receiving a placebo or an active treatment. See also 17.2.1, Randomized Controlled Trial.

convenience sample: sample of subjects selected because they were available for the researchers to study, not because they are necessarily representative of a particular population.

⇒Use of a convenience sample limits generalizability and can confound the analysis depending on the source of the sample. For instance, cardiac auscultation are compared with the results obtained from echocardiography and cardiac catheterization in a group of patients who have undergone all 3 procedures. The patients studied, simply by virtue of their having undergone cardiac catheterization and echocardiography, likely are not comparable to an unselected population.

correlation: description of the strength of an association among 2 or more variables, each of which has been sampled by means of a representative or naturalistic method from a population of interest.[23(p329)] The association is described by the `correlation coefficient`. See also `agreement`.

⇒ The `Kendall τ rank correlation` test is used when testing 2 ordinal variables: the `Pearson product moment correlation` is used when testing 2 normally distributed continuous variables, and the `Spearman rank correlation` is used when testing 2 nonnormally distributed continuous variables.[26]

⇒ Correlation is often depicted graphically by means of a scatterplot of the data (see 2.14, Manuscript Preparation, Figures). The more circular a scatter plot, the smaller the correlation; the more linear a scatterplot, the greater the correlation.

correlation coefficient: measure of the association between 2 variables. The coefficient falls between −1 and 1; the sign indicates the direction of the relationship and the number the magnitude of the relationship. A positive sign indicates that the 2 variables increase or decrease together; a negative sign indicates that one increases while the other decreases. A value of 1 or −1 indicates that the sample values fall in a straight line, while a value of 0 indicates no relationship.[19(p38)] The correlation coefficient should be followed by a measure of the significance of the correlation, and the statistical test used to measure correlation should be specified.

> *Example:* Body mass index increased with age (Pearson $r = 0.61$; $P < .001$); years of education decreased with age (Pearson $r = -0.48$; $P = .01$).

⇒ When 2 variables are compared, the correlation coefficient is expressed by r; when more than 2 variables are compared by multivariate analysis, the correlation coefficient is expressed by R. The symbol r^2 or R^2 is termed the `coefficient of determination` and indicates the amount of variation in the `dependent variable` that can be explained by knowledge of the `independent variable`.

cost-benefit analysis: economic analysis that compares the costs accruing to an individual for some treatment, process, or procedure and the ensuing medical consequences, with the benefits of reduced loss of earnings resulting from prevention of death or premature disability. The cost-benefit ratio is the ratio of marginal benefit (financial benefit of preventing 1 case) to marginal cost (cost of preventing 1 case).[25(p38)] See also 17.2.8, Cost-effectiveness Analysis, Cost-benefit Analysis.

cost-effectiveness analysis: comparison of interventions to determine which provides the most clinical value for the cost.[26] The preferred intervention is the one that will cost the least for a given result or be the most effective for a given cost.[25(pp38-39)] Outcomes are expressed by the cost-effectiveness ratio, such as cost per year of life saved. See also 17.2.8, Cost-effectiveness Analysis, Cost-benefit Analysis.

cost-utility analysis: form of economic evaluation in which the outcomes of alternative procedures are expressed in terms of a single utility-based measurement, most often the quality-adjusted life-year (QALY).[25(p39)]

Cox-Mantel test: method for comparing 2 survival curves that does not assume a particular distribution of data,[27(p63)] similar to the log-rank test.[28(p113)]

Cox proportional hazards regression model (Cox proportional hazards model): in survival analysis, a procedure used to determine relationships between survival time and treatment and prognostic independent variables such as age.[21(p290)] The hazard function is modeled on the set of independent variables without making assumptions that the hazard function is dependent on time. Estimates depend only on the order in which events occur, not on the times they occur.[27(p64)]

criterion standard: test considered to be the diagnostic standard for a particular disease or condition, used as a basis of comparison for other (usually noninvasive) tests. Ideally, the sensitivity and specificity of the criterion standard for the disease should be 100%. (A commonly used synonym, gold standard, is considered jargon,[25(p70)] and thus criterion standard is preferred.) See also diagnostic discrimination.

Cronbach α: index of the internal consistency of a test,[27(p65)] which assesses the correlation between the total score across a series of items and the comparable score that would have been obtained had a different series of items been used.[25(p39)] The Cronbach α is often used for psychological tests.

cross-design synthesis: method for evaluating outcomes of medical interventions, developed by the US General Accounting Office, that pools results from databases of randomized controlled trials and other study designs. It is a form of meta-analysis (see 17.2.7, Meta-analysis).[25(p39)]

crossover design: method of comparing 2 or more treatments or interventions. Individuals initially are randomized to 1 treatment or the other; after completing 1 treatment they are crossed over to the other randomization arm and undergo the other course of treatment. Advantages are that a smaller sample size is needed to detect a difference between treatments, since a paired analysis is used to compare the treatments in each individual, but the disadvantage is that an adequate washout period is needed after the initial course of treatment to avoid carryover effect from the first to the second treatment. Order of treatments should be randomized to avoid potential bias.[27(pp65-66)] See 17.2.2, Crossover Trial.

cross-sectional study: study that identifies subjects with and without the condition or disease under study and the characteristic or exposure of interest at the same point in time.[23(p329)] See 17.2.5, Cross-sectional Study.

⇒ Causality is difficult to establish in a cross-sectional study because the outcome of interest and associated factors are assessed simultaneously.

crude death rate: total deaths during a year divided by the midyear population. Deaths are usually expressed per 100 000 persons.[27(p66)]

cumulative incidence: number of people who experience onset of a disease or outcome of interest during a specified period; may also be expressed as a rate or ratio.[25(p40)]

Cutler-Ederer method: form of life-table analysis that uses actuarial techniques. The method assumes that the times at which follow-up ended (because of death or the outcome of interest) are uniformly distributed during the time period, as opposed to the Kaplan-Meier method, which assumes that termination of follow-up occurs at the end of the time block. Therefore, Cutler-Ederer estimates of risk tend to be slightly higher than Kaplan-Meier estimates.[23(p308)] Often an intervention and control group are depicted on 1 graph and the curves are compared by means of a log-rank test. This is also known as the actuarial method.

cut point: in testing, the arbitrary level at which "normal" values are separated from "abnormal" values, often selected at the point 2 SDs from the mean. See also receiver operating characteristic curve.[25(p40)]

data: collection of items of information.[25(p42)] (Datum, the singular form of this word, is rarely used.)

data dredging (aka "fishing expedition"): jargon meaning post hoc analysis, with no a priori hypothesis, of several variables collected in a study to identify which have a statistically significant association for purposes of publication.

⇒ Although post hoc analyses occasionally can be useful to generate hypotheses, data dredging increases the likelihood of a type I error and should be avoided. If post hoc analyses are performed, they should be declared as such and the number of post hoc comparisons performed specified.

decision analysis: process of identifying all possible choices and outcomes for a particular set of decisions to be made regarding patient care. Decision analysis uses epidemiologic data to estimate the likelihood of occurrence of each outcome. The process is displayed as a decision tree, with each node depicting a branch point representing a decision in treatment or intervention to be made (usually represented by a square at the branch point) or possible outcomes (usually represented by a circle at the branch point). The relative worth of each outcome may be expressed as a utility, such as the quality-adjusted life-year[25(p44)] (Figure 2).

degrees of freedom (*df*): see *df*.

dependent variable: outcome variable of interest in any study; the outcome that one intends to explain or estimate[23(p329)] (for example, death, myocardial infarction, or reduction in blood pressure). When investigating the dependent variable, independent variables that might modify the occurrence of the dependent variable (eg, age, sex, and other medical diseases or risk factors) are controlled for using multivariate analysis.

descriptive statistics: method used to summarize or describe data with the use of the mean, median, SD, SE, or range, or to convey in tabular form (eg, by using a histogram, shown in 2.14, Manuscript Preparation, Figures) for purposes of data presentation and analysis.[27(p73)]

***df* (degrees of freedom)** (*df* is not expanded at first mention): the number of independent comparisons that can be made among members of a sample. In a

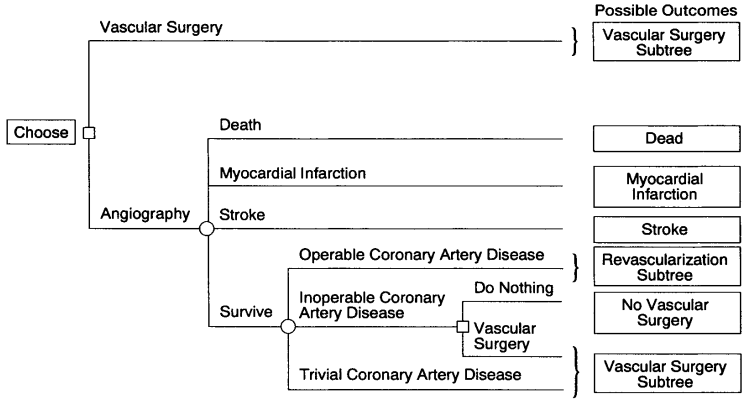


FIGURE 2 Decision tree showing decision nodes (squares) and chance outcomes (circles). End branches are labeled with outcome states. The subtrees to which the decision tree refers are depicted in a separate figure for simplicity. Adapted from Mason JJ, Owens DK, Harris RA, Cooke JP, Hlatky MA. The role of coronary angiography and coronary revascularization before noncardiac vascular surgery. *JAMA.* 1995;273:1919-1925.

`contingency table`, *df* is calculated as (number of rows − 1)(number of columns − 1).

⇒ The *df* should be reported as a subscript after the related statistic, such as the *t* test, analysis of variance, and χ^2 test (eg, $\chi^2_3 = 17.7$, $P = .02$; in this example, the subscript 3 is the number of *df*).

diagnostic discrimination: statistical assessment of how the performance of a clinical diagnostic test compares with the `criterion standard`. To assess a test's ability to distinguish an individual with a particular condition from one without the condition, the researcher must (1) determine the variability of the test, (2) define a population free of the disease or condition and determine the normal range of values for that population for the test (usually the central 95% of values, but in tests that are quantitative rather than qualitative, a `receiver operating characteristic curve` may be created to determine the optimal cut point for defining normal and abnormal), and (3) determine the criterion standard for a disease (by definition, the criterion standard should have 100% sensitivity and specificity for the disease) with which to compare the test. Diagnostic discrimination is reported with the performance measures `sensitivity`, `specificity`, `positive predictive value`, and `negative predictive value`, `false-positive rate`, and the `likelihood ratio`.[23(pp151-163)] See Table 2.

⇒ Because the values used to report diagnostic discrimination are ratios, they can be expressed either as the ratio, using the decimal form, or as the percentage, by multiplying the ratio by 100.

> *Example:* The test had a sensitivity of 0.80 and a specificity of 0.95; the false-positive rate was 0.05.
>
> *Or:* The test had a sensitivity of 80% and a specificity of 95%; the false-positive rate was 5%.

TABLE 2. DIAGNOSTIC DISCRIMINATION

Test Result	Disease by Criterion Standard	Disease-Free by Criterion Standard
Positive	a (true positives)	b (false positives)
Negative	c (false negatives)	d (true negatives)
	$a + c$ = total number of persons with disease	$b + d$ = total number of persons without disease
	Sensitivity $= \frac{a}{a+c}$	Specificity $= \frac{d}{b+d}$
	Positive predictive value $= \frac{a}{a+b}$	Negative predictive value $= \frac{d}{c+d}$

⇒When the diagnostic discrimination of a test is defined, the individuals tested should represent the full spectrum of the disease and reflect the population on whom the test will be used. For example, if a test is proposed as a screening tool, it should be assessed in the general population.

dichotomous variable: a variable with 2 possible answers (eg, male/female); synonym for `binary variable`.[27(p75)]

⇒ The variable may have a `continuous` distribution during data collection but is made dichotomous for purposes of analysis (eg, age < 65 years/age ≥ 65 years). This is done most often for nonnormally distributed data.

direct cause: `contributory cause` that is believed to be the most direct cause of a disease. The direct cause is dependent on the current state of knowledge and may change as more immediate mechanisms are discovered.[23(p330)]

> *Example:* Although several other causes were suggested when the disease was first described, the human immunodeficiency virus is the direct cause of acquired immunodeficiency syndrome.

discordant pair: pair in which the individuals have different outcomes. In twin studies, only the discordant pairs are informative about the association between exposure and disease.[25(pp47-48)] Antonym is concordant pair.

discrete variable: variable that is counted as an integer; no fractions are possible.[27(p77)] Examples are numbers of pregnancies or surgical procedures.

discriminant analysis: analytic technique used to classify subjects according to their characteristics (eg, the `independent variables`, signs, symptoms, and diagnostic test results) to the appropriate outcome or `dependent variable`[27(pp77-78)] This is also referred to as discriminatory analysis[22(pp59-60)] and tests the ability of the `independent variable` model to correctly classify an individual in terms of outcome.

dispersion: degree of scatter shown by observations; may be measured by `SD`, `quantile`, or `range`.[22(p60)]

distribution: group of ordered values; the frequencies or relative frequencies of all possible values of a characteristic.[23(p330)] Distributions may have a `normal`

distribution (bell-shaped curve) or a nonnormal distribution (eg, binomial or Poisson distribution).

dose-response relationship: relationship in which changes in levels of exposure are associated with changes in the frequency of an outcome in a consistent direction. This supports the idea that the agent of exposure (most often a drug) is responsible for the effect seen.[23(p330)] May be tested statistically by using a χ^2 test for trend.

Duncan multiple range test: modified form of the Newman-Keuls test for multiple comparisons.[27(p82)]

Dunnett test: multiple comparisons procedure intended for comparing each of a number of treatments with a single control.[27(p82)]

Dunn test: multiple comparisons procedure based on the Bonferroni adjustment.[27(p84)]

Durbin-Watson test: test to determine whether the residuals from linear regression or multiple regression are independent or, alternatively, are serially correlated.[27(p84)]

ecological fallacy: error that occurs when the existence of a group association is used to imply, incorrectly, the existence of a relationship at the individual level.[23(p330)]

effectiveness: extent to which a treatment is beneficial when implemented under the usual conditions of clinical care for a group of patients,[23(p330)] as distinguished from efficacy (the degree of beneficial effect seen in a clinical trial) and efficiency (the treatment effect achieved relative to the effort expended in time, money, and resources).

effect of observation: bias that results when the process of observation alters the outcome of the study.[23(p330)] See also Hawthorne effect.

effect size: observed or expected change in outcome as a result of an intervention. Expected effect size is used when the sample size necessary to achieve a given power is estimated, since, given a similar amount of variability, a large effect size will require a smaller sample size to detect a difference than will a smaller effect size.

efficacy: degree to which a treatment produces a beneficial result under the ideal conditions of an investigation,[23(p330)] usually in a randomized controlled trial; to be distinguished from effectiveness.

efficiency: effects achieved in relation to the effort expended in money, time, and resources. Statistically, the precision with which a study design will estimate a parameter of interest.[25(pp52-53)]

effort-to-yield measures: amount of resources needed to produce a unit change in outcome, such as number needed to treat[26]; used in cost-effectiveness

and cost-benefit analyses. See 17.2.8, Cost-effectiveness Analysis, Cost-benefit Analysis.

error: difference between a measured or calculated value and the true value. Three types are seen in scientific research: a false or mistaken result obtained in a study; measurement error, a random form of error; and systematic error that skews results in a particular direction.[25(pp56-57)]

estimate: value or values calculated from sample observations that are used to approximate the corresponding value for the population.[23(p330)]

event: end point or outcome of a study; usually the dependent variable. The event should be defined before the study is conducted and assessed by an individual blinded to the intervention or exposure category of the study subject.

exclusion criteria: characteristics of potential study subjects that will exclude them from the study sample (such as being younger than 65 years, history of cardiovascular disease, expected to move within 6 months of the beginning of the study). Exclusion criteria are defined before subjects are enrolled.

explanatory variable: synonymous with independent variable, but preferred by some since *independent* in this context does not refer to statistical independence.[22(p98)]

extrapolation: conclusions drawn about the meaning of a study for a target population that includes types of individuals or data not represented in the study sample.[23(p330)]

factor analysis: procedure used to group related variables to reduce the number of variables needed to represent the data. This analysis is used to explain correlations among groups of variables or factors,[26] especially for developing scoring systems for rating scales and questionnaires.

false negative: negative test result in an individual who has the disease or condition as determined by the criterion standard.[23(p330)] See also diagnostic discrimination.

false-negative rate: proportion of test results found or expected to yield a false-negative result; equal to 1 − sensitivity.[26] See also diagnostic discrimination.

false positive: positive test result in an individual who does not have the disease or condition as determined by the criterion standard.[23(p330)] See also diagnostic discrimination.

false-positive rate: proportion of tests found to or expected to yield a false-positive result; equal to 1 − specificity.[26] See also diagnostic discrimination.

F distribution: ratio of the distribution of 2 normally distributed independent variables; synonymous with variance ratio distribution.[25(p61)]

Fisher exact test: assesses the independence of 2 variables by means of a 2×2 contingency table, used when the frequency in at least 1 cell is small[27(p96)] (usually <6). This test is also known as the Fisher-Yates test and the Fisher-Irwin test.[22(p77)]

fixed-effects model: model used in meta-analysis that assumes that differences in treatment effect in each study all estimate the same true difference. This is not often the case, but the model assumes that it is close enough to the truth that the results will not be misleading.[29(p349)] Antonym is random-effects model.

Friedman test: a nonparametric test for a design with 2 factors that uses the ranks rather than the values of the observations.[22(p80)] Nonparametric analog to analysis of variance.

Friedman urn model: an alternative to random allocation of patients in a clinical trial to avoid imbalance in the number of patients in each group when the number of subjects is small. The model considers an urn filled with balls; each color of ball represents a treatment arm of the study, and the number of balls of each color is proportional to the number of patients to be enrolled in that treatment arm. If the number of balls greatly exceeds the number of patients to be enrolled, selecting a ball to determine which treatment arm the patient will be enrolled in will approach random assignment. However, if the number of balls is similar to the total number of patients, after most of the balls have been selected the remaining balls will tend to even out the total number of patients in each treatment arm and treatment will not be based on random assignment.[27(p101)]

F test (score): alternative name for the variance ratio test,[25(p74)] which results in the F score. Often encountered in analysis of variance.[27(p101)]

> *Example:* There were differences by academic status in perceptions of the quality of both primary care training ($F_{1,682} = 6.71$, $P = .01$) and specialty training ($F_{1,682} = 6.71$, $P = .01$). [The numbers set as subscripts are the *df* for the analysis.]

Gaussian distribution: see normal distribution.

gold standard: see criterion standard.

goodness of fit: agreement between an observed set of values and a second set that is derived wholly or partly on a hypothetical basis.[22(p86)] The Kolmogorov-Smirnov test is one example.

group association: situation in which a characteristic and a disease both occur more frequently in 1 group of individuals than another. The association does not mean that individuals with the characteristic also have the disease.[23(p331)]

group matching: process of matching during assignment in a study to ensure that the groups have a nearly equal distribution of particular variables; also known as frequency matching.[23(p331)]

Hartley test: test for the equality of variances of a number of populations that are normally distributed, based on the ratio between the largest and smallest sample variations.[22(p90)]

Hawthorne effect: effect produced in an experiment because of the awareness of the study subjects that they are participating in a study. The term usually refers to an effect on the control group that changes the group in the direction of the outcome, resulting in a smaller effect size.[27(p115)] A related concept is effect of observation.

hazard rate, hazard function: theoretical measure of the likelihood that an individual will experience an event within a given period.[25(p73)] A number of hazard rates for specific intervals of time can be combined to create a hazard function.

heterogeneity: inequality of a quantity of interest (such as variance) in a number of groups or populations. Antonym is homogeneity.

histogram: graphical representation of data in which the frequency (quantity) within each class or category is represented by the area of a rectangle centered on the class interval. The heights of the rectangles are proportional to the observed frequencies. See also 2.14, Manuscript Preparation, Figures.

Hoeffding independence test: bivariate test of nonnormally distributed continuous data to determine whether the elements of the 2 groups are independent of each other.[25(p93)]

Hollander parallelism test: determines whether 2 regression lines for 2 independent variables plotted against a dependent variable are parallel. The test does not require a normal distribution, but there must be an equal and even number of observations corresponding to each line. If the lines are parallel, then both independent variables predict the dependent variable equally well. The Hollander parallelism test is a special case of the signed rank test.[22(p94)]

homogeneity: equality of a quantity of interest (such as variance) specifically in a number of groups or populations.[22(p94)] Antonym is heterogeneity.

homoscedasticity: statistical determination that the variance of the different variables under study is equal.[25(p78)] See also heterogeneity.

Hosmer-Lemeshow goodness-of-fit-test: a series of statistical steps used to assess goodness of fit; approximates the χ^2 statistic.[30]

Hotelling *T* statistic: generalization of the *t* test for use with multivariate data; results in a *T* statistic. Significance can be tested with the variance ratio distribution.[22(p94)]

hypothesis: supposition that leads to a prediction that can be tested to be either supported or refuted.[25(p80)] The null hypothesis is that no such relationship exists, and any association is based strictly on chance. Hypothesis testing includes (1) generating the study hypothesis and defining the null hypothesis, (2) determining the level below which results are considered statistically significant, or α level (usually $\alpha = .05$), and (3) identifying and applying the appropriate statistical test to accept or reject the null hypothesis.

incidence: number of new cases of disease that occur over time,[25(p82)] as contrasted with prevalence, which is the total number of persons with the disease at any

given time. Incidence is usually expressed as a percentage of individuals who will be affected during a year, or as a rate calculated as the number of individuals who develop the disease during a period divided by the number of person-years at risk.

Example: The incidence rate for the disease was 1.2 cases per 100 000 per year.

inclusion criteria: characteristics a study subject must possess to be included in the study population (such as age 65 years or older at the time of study enrollment and willing and able to provide informed consent). Inclusion criteria are defined before subjects are enrolled.

independence, assumption of: assumption that the occurrence of 1 event is in no way linked to another event. Many statistical tests depend on the assumption that each outcome is independent.[25(p83)] This may not be a valid assumption if repeated tests are performed on 1 individual (eg, blood pressure is measured sequentially over time), if more than 1 outcome is measured for a given individual (eg, myocardial infarction and death or all hospital admissions), or if more than 1 intervention is made on the same individual (eg, blood pressure is measured during 3 different drug treatments). Tests for `repeated measures` may be used in those circumstances.

independent variable: variable postulated to influence the `dependent variable` within the defined area of relationships under study.[25(p83)] The term does not refer to statistical independence, so some use the term `explanatory variable` instead.[22(p98)]

Example: Age, sex, systolic blood pressure, and cholesterol were the independent variables entered into the multiple logistic regression.

indirect cause: contributory cause that acts through the biological mechanism that is the direct cause.[23(p331)]

Example: Overcrowding in the cities facilitated transmission of the tubercle bacillus and precipitated the tuberculosis epidemic. [Overcrowding is an indirect cause; the tubercle bacillus is the direct cause.]

inference: process of passing from observations to generalizations, usually with calculated degrees of uncertainty.[25(p85)]

Example: Intake of a high-fat diet was significantly associated with cardiovascular mortality; therefore, we infer that eating a high-fat diet increases the risk of cardiovascular death.

instrument error: error introduced in a study when the testing instrument is not appropriate for the conditions of the study or is not accurate enough to measure the study outcome[23(p331)] (may be due to deficiencies in such factors as calibration, `accuracy`, and `precision`).

intention-to-treat analysis, intent-to-treat analysis: analysis of outcomes for individuals based on the treatment arm to which they were randomized, rather than which treatment they actually received and whether they completed the study. The intention-to-treat analysis preserves the process of randomization and should be the main analysis of a randomized trial.[27(p125)] See 17.2.1, Randomized Controlled Trials.

⇒Although other analyses, such as evaluable patient analysis, are often performed to evaluate outcomes based on treatment actually received, the intention-to-treat analysis should be presented regardless of other analyses because the intervention may influence whether treatment was changed and whether subjects dropped out.

interaction: see interactive effect.

interaction term: variable used in analysis of covariance in which 2 independent variables interact with each other (for example, when assessing the effect of energy expenditure on cardiac output, the increase in cardiac output per unit increase in energy expenditure might differ between men and women; the interaction term would enable the analysis to take this difference into account).[23(p301)]

interactive effect: effect of 2 or more independent variables on a dependent variable in which the effect of an independent variable is influenced by the presence of another.[22(p101)] The interactive effect may be additive (ie, equal to the sum of the 2 effects present separately), synergistic (ie, the 2 effects together have a greater effect than the sum of the effects present separately), or antagonistic (ie, the 2 effects together have a smaller effect than the sum of the effects present separately).

interim analysis: data analysis carried out during a clinical trial to monitor treatment effects. Interim analysis should be determined as part of the study protocol prior to patient enrollment and specify the stopping rules if a particular treatment effect is reached.[3(p130)]

interobserver bias: likelihood that one observer is more likely to give a particular response than another observer because of factors unique to the observer or instrument. For example, one physician may be more likely than another to identify a particular set of signs and symptoms as indicative of religious preoccupation on the basis of his or her beliefs, or a physician may be less likely than another physician to diagnose alcoholism in a patient because of the physician's expectations.[28(p25)] The Cochran *Q* test is used to assess interobserver bias.[28(p25)]

interobserver reliability: test used to measure agreement among observers about a particular measure or outcome.

⇒Although the proportion of times that 2 observers agree can be reported, this does not take into account the number of times they would have agreed by chance alone. For example, if 2 observers must decide whether a factor is present or absent, they should agree 50% of the time according to chance. The κ statistic assesses agreement while taking chance into account and is described by the equation [(observed agreement) − (agreement expected by chance)]/(1 − agreement expected by chance). The value of κ may range from 0 (poor agreement) to 1 (perfect agreement) and may be classified by various descriptive terms, such as slight (0-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and near perfect (0.81-0.99).[28(pp27-29)]

⇒ In cases in which disagreement may render especially grave consequences, such as one pathologist rating a slide "negative" and another rating a slide "invasive carcinoma," a weighted κ may be used to grade disagreement according to the

severity of the consequences.[28(p29)] See also Pearson product moment correlation.

interobserver variation: see interobserver reliability.

interquartile range: range used to describe the dispersion of values; describes the variate distance between the upper and lower quartiles. This and other quantiles are used to describe nonnormally distributed data, since SD does not accurately describe such data. The interquartile range describes the inner 50% of values; the interquintile range describes the inner 60% of values; the interdecile range describes the inner 80% of values.[22(pp102-103)]

interrater reliability: reproducibility among raters or observers; synonymous with interobserver reliability.

interval estimate: see confidence interval.[23(p331)]

intraobserver reliability (or variation): reliability (or, conversely, variation) in measurements by the same person at different times.[23(p331)] Similar to interobserver reliability, intraobserver reliability is the agreement between measurements by 1 individual beyond that expected by chance and can be measured by means of the κ statistic or the Pearson product moment correlation.

intrarater reliability: synonym for intraobserver reliability.

jackknife dispersion test: technique for estimating the variance and bias of an estimator, applied to a predictive model derived from a study sample to determine whether the model fits subsamples from the model equally well. The estimator or model is applied to subsamples of the whole, and the differences in the results obtained from the subsample compared with the whole are analyzed as a jackknife estimate of variance. This method uses a single data set to derive and validate the model.[27(p131)]

⇒ Although validating a model in a new sample is preferable, investigators often use techniques such as jackknife dispersion or the bootstrap method to validate a model to save the time and expense of obtaining an entirely new sample for purposes of validation.

Kaplan-Meier method: nonparametric method of compiling life tables. Unlike the Cutler-Ederer method, the Kaplan-Meier method assumes that termination of follow-up occurs at the end of the time block. Therefore, Kaplan-Meier estimates of risk tend to be slightly lower than Cutler-Ederer estimates.[23(p308)] Often an intervention and control group are depicted on one graph and the curves are compared by a log-rank test. This method is also known as the product-limit method.

κ (kappa) statistic: statistic used to measure nonrandom agreement between observers or measurements.[25(p94)] See interobserver and intraobserver reliability.

Kendall τ (tau) rank correlation: rank correlation coefficient for ordinal data. The coefficient is τ.[27(p134)]

Kolmogorov-Smirnov test: comparison of 2 independent samples of continuous data without requiring that the data be normally distributed[27(p136)]; may be used to test goodness of fit.[26]

Kruskal-Wallis test: comparison of 3 or more groups of nonnormally distributed data to determine whether they differ significantly.[27(p137)] The Kruskal-Wallis test is a nonparametric analog of analysis of variance and generalizes the 2-sample Wilcoxon rank sum test to the multiple-sample case.[22(p111)]

kurtosis: the way in which a unimodal curve deviates from a normal distribution; may be more peaked (leptokurtic) or more flat (platykurtic) than a normal distribution.[27(p137)]

Latin square: form of complete treatment crossover design used for crossover drug trials that eliminates the effect of treatment order. Each patient receives each drug, but each drug is followed by another drug only once in the array. For example, in the following 4×4 array, letters A through D correspond to each of 4 drugs, each row corresponds to a patient, and each column corresponds to the order in which the drugs are given[3(p142)]:

	First Drug	*Second Drug*	*Third Drug*	*Fourth Drug*
Patient 1	C	D	A	B
Patient 2	A	C	B	D
Patient 3	D	B	C	A
Patient 4	B	A	D	C

See also 17.2.2, Crossover Trial.

lead-time bias: artifactual increase in survival time that results from earlier detection of a disease, usually cancer, during a time when the disease is asymptomatic. Lead-time bias produces longer survival from the time of diagnosis but not longer survival from the time of onset of the disease.[23(p331)] See also length-time bias.

⇒Lead-time bias may give the appearance of a survival benefit from screening, when in fact the increased survival is only artifactual. Lead-time bias is used more generally to indicate a systematic error arising when follow-up of groups does not begin at comparable stages in the natural course of the condition.

least significant difference test: test for comparing mean values arising in analysis of variance. An extension of the *t* test.[22(p115)]

least squares method: method of estimation, particularly in regression analysis, that minimizes the differences between the observed response and the values predicted by the model.[27(p140)] The regression line is created so that the sum of the squares of the residuals is as small as possible.

left-censored data: see censored data.

length-time bias: bias that arises when a sampling scheme is based on patient visits, because patients with more frequent clinic visits are more likely to be

selected than those with less frequent visits. In a screening study of cancer, for example, screening patients with frequent visits is more likely to detect slow-growing tumors than would sampling patients who visit a physician only when symptoms arise.[27(p140)] See also `lead-time bias`.

life table: method of organizing data that allows examination of the experience of 1 or more groups of individuals over time with varying periods of follow-up. For each increment of the follow-up period, the number entering, the number leaving, and the number dying of disease or developing disease can be calculated. An assumption of the life-table method is that an individual not completing follow-up is exposed for half the incremental follow-up period.[27(p143)] (The `Kaplan-Meier method` and the `Cutler-Ederer method` are also forms of life-table analysis but make different assumptions about the length of exposure.) See Figure 3.

⇒ The *clinical life table* describes the outcomes of a cohort of individuals classified according to their exposure or treatment history. The *cohort life table* is used for a cohort of individuals born at approximately the same time and followed up until death. The *current life table* is a summary of mortality of the population over a brief (1- to 3-year) period, classified by age, often used to estimate life expectancy for the population at a given age.[25(p97)]

likelihood ratio: probability of getting a certain test result if the patient has the condition compared with the probability of getting the result if the patient does not have the condition. Calculated as `sensitivity/(1 − specificity)`. The greater the likelihood ratio, the more likely that a positive test result will occur in a patient who has the disease. A ratio of 2 means a person with the disease is twice as likely to have a positive test result as a person without the disease.[26] The

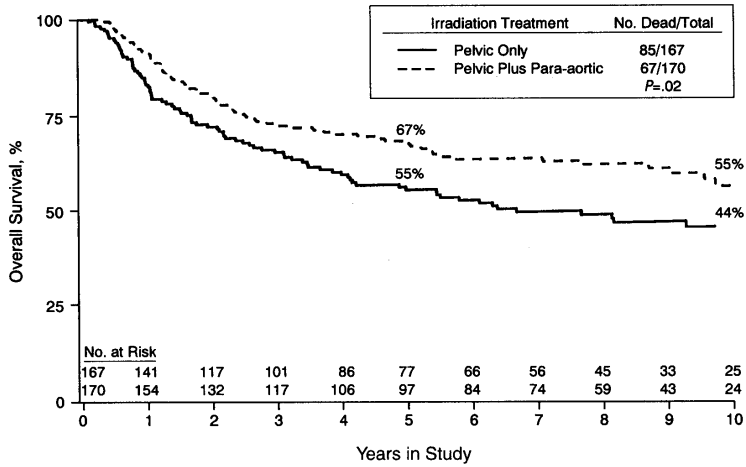


FIGURE 3 Survival curve showing outcomes for 2 treatments groups with number at risk at each time point. While numbers at risk are not essential to include in a survival analysis figure, this presentation conveys more information than the curve alone would. Adapted from Rotman M, Pajak TF, Choi K, et al. Prophylactic extended-field irradiation of para-aortic lymph nodes in stages IIB and bulky IB and IIA cervical carcinomas. *JAMA*. 1995;274:387-393.

likelihood ratio test is based on the ratio of 2 likelihood functions.[22(p118)] See also diagnostic discrimination.

Likert scale: scale often used to assess opinion or attitude, ranked by attaching a number to each response such as 1, strongly approve; 2, approve; 3, undecided or neutral; 4, disapprove; 5, strongly disapprove. The score is a sum of the numerical responses to each question.[27(p144)]

Lilliefors test: test of normality (using the Kolmogorov-Smirnov test statistic) in which mean and variance are estimated from the data.[22(p118)]

linear regression: statistical method used to compare continuous dependent and independent variables. When the data are depicted on a graph as a regression line, the independent variable is plotted on the x-axis and the dependent variable on the y-axis. The residual is the vertical distance from the data point to the regression line[26(p110)]; analysis of residuals is a commonly used procedure for linear regression. (See 2.14, Manuscript Preparation, Figures, Example F13.) This method is frequently performed using least squares regression.[21(pp202-203)]

⇒ The description of a linear regression model should include the equation of the fitted line with the slope and 95% confidence interval if possible, the fraction of variation in y explained by the x variables (correlation), and the variances of the fitted coefficients a and b (and their SDs).[21(p227)]

> *Example:* The regression model identified a significant positive relationship between the dependent variable weight and height (slope = 0.25; 95% CI, 0.19-0.31; $y = 12.6 + 0.25x$; $t_{451} = 8.3$, $P < .001$; $r^2 = 0.67$).[26]

[In this example, the slope is positive, indicating that as one variable increases the other increases; the t test with 451 df is significant; the regression line is described by the equation and includes the slope 0.25 and the constant 12.6, and the coefficient of determination r^2 demonstrates that 67% of the variance in weight is explained by the height.][26]

⇒ Four important assumptions are made when linear regression is conducted: the dependent variable is sampled randomly from the population, the spread or dispersion of the dependent variable is the same regardless of the value of the independent variable (this equality is referred to as homogeneity of variances or homoscedasticity), the relationship between the 2 variables is linear, and the independent variable is measured with complete precision.[23(pp273-274)]

location: central tendency of a normal distribution, as distinguished from dispersion. The location of 2 curves may be identical (means are the same) but the kurtosis may vary (one may be peaked and the other flat, producing small and large SDs, respectively).[31(p28)]

logistic regression: type of regression model used to analyze the relationship between a binary dependent variable (expressed as a natural log after a *logit transformation*) and 1 or more independent variables. Often used to determine the independent effect of each of several explanatory variables by controlling for several factors simultaneously in a multiple logistic regression analysis. Results are usually expressed by odds ratios or relative risks and 95% confidence intervals.[23(pp311-312)] (The multiple logistic regression equation may also be provided but is substantially more complicated than the linear regression equation.

Therefore, in AMA publications, the equation is generally not published but can be made available on request from authors. Alternatively, it may be placed in NAPS [see 2.10.6, Manuscript Preparation, National Auxiliary Publications Service (NAPS)] or on the World Wide Web.)

⇒To be valid, a multiple regression model must have an adequate sample size for the number of variables examined. A rough rule of thumb is to have at least 25 individuals in the study for each explanatory variable examined.

log-linear model: linear models used in the analysis of categorical data.[22(p122)]

log-rank test: method of using the relative death rates in subgroups to compare overall differences between survival curves for different treatments; same as the Mantel-Haenszel test.[22(pp122,124)]

main effect: estimate of the independent effect of an explanatory (independent) variable on a dependent variable in analysis of variance.[27(p153)]

Mann-Whitney test: nonparametric equivalent of the *t* test, used to compare ordinal dependent variables with either nominal independent variables or continuous independent variables converted to an ordinal scale.[25(p100)] Alternative name for Wilcoxon rank sum test.

MANOVA: multivariate analysis of variance.

Mantel-Haenszel test: another name for the log-rank test.

Markov process: process of modeling possible events or conditions over time that assumes that the probability that a given state or condition will be present depends only on the state or condition immediately preceding it and that no additional information about previous states or conditions would create a more accurate estimate.[27(p155)] If the assumptions are appropriate, it can be used instead of decision analysis.

masked assessment: synonymous with blinded assessment, preferred by some investigators and journals to the term ***blinded***, especially in ophthalmology.

masked assignment: synonymous with blinded assignment, preferred by some investigators and journals to the term ***blinded***, especially in ophthalmology.

matching: process of making study and control groups comparable with respect to factors other than the factors under study, generally as part of a case-control study. Matching can be done in several ways, including frequency matching (matching on frequency distributions of the matched variable[s]), category (matching in broad groups such as young and old), individual (matching on individual rather than group characteristics), and pair matching (matching each study individual with a control individual).[25(p101)]

McNemar test: form of the χ^2 test for binary responses in comparisons of matched pairs.[25(p103)] The ratio of discordant to concordant pairs is determined; the greater the number of discordant pairs with the better outcome

being associated with the treatment intervention, the greater the effect of the intervention.[27(p158)]

mean: sum of values measured for a given variable divided by the number of values; a measure of central tendency appropriate for normally distributed data.[31(p29)]

⇒ If the data are not normally distributed, the median is preferred. See also average.

measurement error: estimate of the variability of a measurement. Variability of a given parameter (eg, weight) is the sum of the true variability of what is measured (eg, day-to-day weight fluctuations) plus the variability of the instrument or observer measurement, or variability caused by measurement error (error variability, eg, the scale used for weighing). The intraclass correlation coefficient R measures the relationship of these 2 types of variability: as the error variability declines with respect to true variability, R increases, up to 1 when error variance is 0. If all variability is a result of error variability, then $R = 0$.[28(p30)]

median: midpoint of a distribution chosen so that half the values for a given variable appear above and half occur below.[23(p332)] For nonnormally distributed data, the median provides a better measure of central tendency than does the mean, since it is less influenced by outlying values.[30(p29)]

median test: nonparametric rank-order test for 2 groups.[22(p128)]

meta-analysis: See 17.2.7, Meta-analysis.

mode: in a series of values of a given variable, the number that occurs most frequently; used most often when a distribution has 2 peaks (bimodal distribution).[31(p29)]

mortality rate: death rate described by the following equation: [(number of deaths during period) × (period of observation)]/(number of individuals observed). For values such as the crude mortality rate, the denominator is the number of individuals observed at the midpoint of observation. See also crude death rate.[27(p66)]

⇒ Mortality rate is often expressed in terms of a standard ratio, such as deaths per 100 000 persons per year.

Moses ranklike dispersion test: rank test of the equality of scale of 2 identically shaped populations, applicable when the population medians are not known.[22(p134)]

multiple analyses problem: problem that occurs when several statistical tests are performed on one group of data because of the potential to introduce a type I error. The problem is particularly an issue when the analyses were not specified as primary outcome measures. Multiple analyses can be appropriately adjusted for by means of a Bonferroni adjustment or any of several multiple comparisons procedures.

multiple comparisons procedures: any of several tests used to determine which groups differ significantly after another more general test has identified that a significant difference exists but not between which groups. These tests are in-

tended to avoid the problem of a type I error caused by sequentially applying tests such as the *t* test not intended for repeated use.

⇒Some tests result in more conservative estimates (less likely to be significant) than others. More conservative tests include the Tukey test and the Bonferroni adjustment; the Duncan multiple range test is less conservative. Other tests include the Scheffé test, the Newman-Keuls test, and the Gabriel test.[22(p137)]

multiple regression: general term for multivariate analysis procedures used to estimate values of the dependent variable for all measured independent variables that are found to be associated. The procedure used depends on whether the variables are continuous or nominal. When all variables are continuous variables, multiple linear regression is used and the mean of the dependent variable is expressed using the equation $Y = \alpha + \beta_1\chi_1 + \beta_2\chi_2 + \ldots + \beta_k\chi_k$. When independent variables may be either nominal or continuous and the dependent variable is continuous, analysis of covariance is used. (Analysis of covariance often requires an interaction term to account for differences in the relationship between the independent and dependent variables.) When all variables are nominal and the dependent variable is time-dependent, life-table methods are used. When the independent variables may be either continuous or nominal and the dependent variable is nominal and time-dependent (such as incidence of death), the Cox proportional hazards model may be used. Nominal dependent variables that are not time-dependent are analyzed using logistic regression or discriminant analysis.[23(pp296-312)]

multivariable analysis: another name for multivariate analysis.

multivariate analysis: any statistical test that deals with 1 dependent variable and at least 2 independent variables. It may include nominal or continuous variables, but ordinal data must be converted to a nominal scale for analysis. The multivariate approach has 3 advantages over bivariate analysis: (1) it allows for investigation of the relationship between the dependent and independent variables while controlling for the effects of other independent variables, (2) it allows several comparisons to be made statistically without increasing the likelihood of a type I error, and (3) it can be used to compare how well several independent variables individually can estimate values of the dependent variable.[23(pp289-291)] Examples include analysis of variance, multiple (logistic or linear) regression, analysis of covariance, Kruskal-Wallis test, Friedman test, life table, and Cox proportional hazards model.

N: entire population under study.

> *Example:* We assessed the diagnoses of admission all patients admitted from the emergency department during a 1-month period (N = 127).

n: sample of the population under study.

> *Example:* Of the patients admitted from the emergency department (N = 127), the most frequent admission diagnosis was unstable angina (n = 38).

natural experiment: investigation in which a change in a risk factor or exposure occurs in one group of individuals but not in another. The distribution of individuals

into a particular group is nonrandom and, as opposed to controlled clinical trials, the change is not brought about by the investigator.[23(p332)] The natural experiment is often used to study effects that cannot be studied in a controlled trial, such as the incidence of medical illness immediately after an earthquake. This is also referred to as a "found" experiment.

naturalistic sample: set of observations obtained from a sample of the population in such a way that the distribution of `independent variables` in the sample is representative of the distribution in the population.[23(p332)]

necessary cause: characteristic whose presence is required to bring about or cause the disease or outcome under study.[32(p332)]

negative predictive value: the probability that an individual does not have the disease (as determined by the `criterion standard`) if the test result is negative.[23(p334)] This measure takes into account the `prevalence` of the condition or the disease. A more general term is `posttest probability`. See `diagnostic discrimination`.

nested case-control study: `case-control study` in which cases and controls are drawn from a cohort study. The advantages of a nested case-control study over a case-control study are that the controls are selected from subjects at risk at the time of occurrence of each case that arises in a cohort, thus avoiding the confounding effect of time in the analysis, and that cases and controls are by definition drawn from the same population.[25(p111)] See also 17.2.4, Case-Control Study, and 17.2.3, Cohort Study.

Newman-Keuls test: type of `multiple comparisons procedure`, used to compare more than 2 groups, that first compares the 2 groups that have the highest and lowest means, then sequentially compares the next most extreme groups, and stops when a comparison is not significant.[33(p92)]

n-of-1 trial: randomized controlled trial that uses a single patient and an outcome measure agreed on by the patient and physician. The n-of-1 trial may be used by clinicians to assess which of 2 possible treatment options is superior for the individual patient.[32]

nominal variable: variable with named categories. If nominal data have more than 2 categories, the categories are not ordered (for example, gene alleles, race, or eye color). The nominal or discrete variable usually is assessed to determine its frequency within a population.[23(p332)] The variable can have either a `binomial` (equal chance for each category) or `Poisson` (the nominal event is extremely rare, eg, a genetic mutation) `distribution`.

nonconcurrent cohort study: `cohort study` in which an individual's group assignment is determined by information that exists at the time a study begins. The extreme of a nonconcurrent cohort study is one in which the outcome is determined retrospectively from existing records.[23(p332)]

nonnormal distribution: data that do not have a `normal` (bell-shaped curve) distribution; includes `binomial` or `Poisson distribution`.

⇒ Nonnormally distributed continuous data must be either transformed to a normal distribution to use parametric methods or, more commonly, analyzed by nonparametric methods.

nonparametric statistics: statistical procedures used for data that do not have a normal distribution. Nonparametric tests are most often used for ordinal or nominal data, or for nonnormally distributed continuous data converted to an ordinal scale[23(p332)] (for example, weight classified by tertile).

normal distribution: continuous data distributed in a symmetrical, bell-shaped curve with the mean value corresponding to the highest point of the curve. This distribution of data is assumed in many statistical procedures.[23(p330)] This is also called a Gaussian distribution.

⇒Descriptive statistics such as mean and SD can be used to accurately describe data only if the values are normally distributed or can be transformed into a normal distribution. Parametric statistics assume that data are normally distributed.

normal range: measure of the range of values on a particular test among those without the disease. Cut points for abnormal tests are arbitrary and are often defined as the central 95% of values, or the mean of values ±2 SDs.

null hypothesis: the assertion that no true association or difference in the study outcome or comparison of interest between comparison groups exists in the larger population from which the study samples are obtained.[23(p332)] The hypothesis is expressed as the null hypothesis to be proved (no significant difference) or disproved (a statistically significant difference) by statistical analysis.

number needed to treat (NNT): number of patients who must be treated with an intervention for a specific period of time to prevent 1 bad outcome or result in 1 good outcome.[23(pp332-333)] The NNT is the reciprocal of the absolute risk reduction, the difference between event rates in the intervention and placebo groups in a clinical trial.

⇒ The study patients from whom the NNT is calculated should be representative of the population to whom the numbers will be applied. The NNT does not take into account adverse effects of the intervention.

odds ratio (OR): ratio of 2 odds. Odds ratio may have different definitions depending on the study and therefore should be defined. For example, it may be the odds of having the disease if a particular risk factor is present to the odds of not having the disease if the risk factor is not present, or the odds of having a risk factor present if the person has the disease to the odds of the risk factor being absent if the person does not have the disease.

The odds ratio typically is used for a case-control or cohort study. For a study of incident cases with an infrequent disease (for example, <2% incidence), the odds ratio approximates the relative risk.[25(p118)]

⇒ The odds ratio is usually expressed by a point estimate and 95% confidence interval (CI). An odds ratio for which the CI includes 1 indicates no statistically significant effect on risk; if the point estimate and CI are both less than 1, there is a statistically significant reduction in risk; if the point estimate and CI are both greater than 1, there is a statistically significant increase in risk.

1-tailed test: test of statistical significance in which deviations from the null hypothesis in only 1 direction are considered.[23(p333)] Most commonly used for the *t* test.

⇒One-tailed tests are more likely to produce a statistically significant result than are 2-tailed tests. Since the use of a 1-tailed test implies that the intervention could have only 1 direction of effect, ie, beneficial or harmful, the justification for the use of a 1-tailed test must be provided.

ordinal data: type of data with a limited number of categories with an inherent ordering of the category from lowest to highest, but without fixed or equal spacing between increments.[23(p333)] Examples are Apgar scores, heart murmur rating, and cancer stage and grade. Discrete variables, such as family size, parity, or number of teeth, are special forms of ordinal data. Ordinal data can be summarized by means of the median and quantiles or range.

⇒ Since increments between the numbers for ordinal data generally are not fixed (eg, the difference between a grade 1 and a grade 2 heart murmur is not quantitatively the same as the difference between a grade 3 and a grade 4 heart murmur), ordinal data should be analyzed by nonparametric statistics.

ordinate: vertical or y-axis of a graph.

outcome: dependent variable or end point of an investigation. In retrospective studies such as case-control studies, the outcome occurs before the study; in prospective studies such as cohort studies and controlled trials, the outcome occurs during the study.[23(p333)]

outliers (outlying values): values at the extremes of a distribution. The median is preferred to describe data with outliers that influence the mean.

⇒If outliers are excluded from an analysis, the exclusion should be explained in the text.

overmatching: obscuring by the matching process of a case-control study a true causal relationship between the independent and dependent variables because the variable used for matching is strongly related to the mechanism by which the independent variable exerts its effect.[25(pp119-120)] For example, matching cases and controls on residence within a certain area could obscure an environmental cause of a disease. Overmatching may also be used to refer to matching on variables that have no effect on the dependent variable, and therefore are unnecessary, or the use of so many variables for matching that no suitable controls can be found.[25(p120)]

paired samples: form of matching that can include self-pairing, when each subject serves as his or her own control, or artificial pairing, when 2 subjects are matched on prognostic variables.[27(p186)] Paired analyses provide greater power to detect a difference for a given sample size than do nonpaired analyses, since interindividual differences are minimized or eliminated. Pairing may also be used to match subjects in case-control or cohort studies. See Table 3.

paired *t* test: *t* test for paired data.

parameter: measurable characteristic of a population. One purpose of statistical analysis is to estimate population parameters from sample observations.[23(p333)] The

statistic is the numerical characteristic of the sample; the parameter is the numerical characteristic of the population. *Parameter* is also used to refer to aspects of a model (eg, a regression model).

parametric statistics: tests used for continuous data and that require the assumption that the data being tested are normally distributed, either as collected initially or after transformation to the ln or log of the value or other mathematical conversion.[25(p121)] The *t* test is a parametric statistic. See Table 3.

Pearson product moment correlation: test of correlation between 2 groups of normally distributed data. See diagnostic discrimination.

point estimate: single value calculated from sample observations that is used as the estimate of the population value, or parameter[23(p333)]; in most circumstances accompanied by an interval estimate (eg, 95% confidence interval).

Poisson distribution: distribution that occurs when a nominal event (often disease or death) occurs rarely.[25(p125)] The Poisson distribution is used instead of a binomial distribution when sample size is calculated for a study of events that occur rarely.

population: any finite or infinite collection of subjects from which a sample is drawn for a study to obtain estimates to approximate the values that would be obtained if the entire population were sampled.[27(p197)]

population attributable risk percentage: percentage of risk within a population that is associated with exposure to the risk factor. Population attributable risk takes into account the frequency with which a particular event occurs and the frequency with which a given risk factor occurs in the population. Population attributable risk does not necessarily imply a cause-and-effect relationship. It is also called *attributable fraction, attributable proportion*, and *etiologic fraction*.[23(p333)]

positive predictive value: proportion of those with a positive test result who have the condition or disease as measured by the criterion standard. This measure takes into account the prevalence of the condition or the disease. Clinically, it is the probability that an individual has the disease if the test is positive[23(p334)] (synonym: posttest probability). See Table 2 and diagnostic discrimination.

posterior probability: in Bayesian analysis, the probability obtained after the prior probability is combined with the probability from the study of interest.[25(p128)] If one assumes a uniform prior (no useful information for estimating probability exists before the study), the posterior probability is the same as the probability from the study of interest alone.

post hoc analysis: analysis performed after completion of a study and not based on a hypothesis considered before the study. Such analyses should be performed without prior knowledge of the relationship between the dependent and independent variables. A potential hazard of post hoc analysis is the type I error.

⇒While post hoc analyses may be used to explore intriguing results and generate new hypotheses for future testing, they should not be used to test hypotheses, since the comparison is not hypothesis-driven. See also data dredging.

posttest probability: the probability that an individual has the disease if the test result is positive (positive predictive value) or that the individual does not have the disease if the test result is negative (negative predictive value).[23(p158)]

power: ability to detect a significant difference with the use of a given sample size and variance; determined by frequency of the condition under study, magnitude of the effect, study design, and sample size.[25(p128)] Power should be calculated before a study is begun. If the sample is too small to have a reasonable chance (usually 80% or 90%) of rejecting the null hypothesis if a true difference exists, then a negative result may indicate a type II error rather than a true acceptance of the null hypothesis.

⇒ Power calculations are important to perform when designing a study; a statement providing the power of the study should be included in the methods section of all randomized controlled trials (see Table 1) and is appropriate for many other types of studies. A power statement is especially important if the study results are negative, to demonstrate that a type II error was not the reason for the negative result. Performing a post hoc power analysis is controversial, especially if it is based on the study results, but, if included, it should be placed in the discussion section and the fact that it was performed post hoc clearly stated.

> *Example:* We determined that a sample size of 800 patients would have 80% power to detect the clinically important difference of 10% at $\alpha = .05$.

precision: inverse of the variance in measurement (see measurement error)[25(p129)]; the degree of accuracy with which an instrument can make measurements.

pretest probability: see prevalence.

prevalence: proportion of persons with a particular disease at a given point in time. Prevalence can also be interpreted to mean the likelihood that a person selected at random from the population will have the disease (synonym: pretest probability).[23(p334)] See also incidence.

principal components analysis: procedure used to group related variables to help describe data. The variables are grouped so that the original set of correlated variables is transformed into a smaller set of uncorrelated variables called the *principal components*.[25(p131)] Variables are not grouped according to dependent and independent variables, unlike many forms of statistical analysis. Principal components analysis is similar to factor analysis.

prior probability: in Bayesian analysis, the probability of an event based on previous information before the study of interest is considered. The prior probability may be informative, based on previous studies or clinical information, or not, in which case the analysis uses a uniform prior (no information is known before the study of interest). A *reference prior* is one with minimal information, a *clinical prior* is based on expert opinion, and a *skeptical prior* is used when large treatment differences are not expected.[27(p201)] When Bayesian analysis is used to determine the posterior probability of a disease after a patient has undergone a diagnostic test, the prior probability is the prevalence of the disease in the population from which the patient is drawn (usually the clinic or hospital population).

probability: in clinical studies, the number of times an event occurs in a study group divided by the number of individuals being studied.[23(p334)]

product-limit method: see Kaplan-Meier method.

proportionate mortality ratio: number of individuals who die of a particular disease during a span of time divided by the number of individuals who die of all diseases during the same period.[23(p334)] This ratio may also be expressed as a rate if corrected to a standard unit of time (eg, cardiovascular deaths per total deaths per year).

prospective study: study in which subjects with and without an exposure are identified and then followed up over time; the outcomes of interest have not occurred at the time the study commences.[27(p205)] *Prospective* is most commonly used in the context of a cohort study.

pseudorandomization: assigning of individuals to 1 of 2 groups in a nonrandom manner, eg, selecting every other individual for an intervention or assigning subjects by Social Security number or birth date.

publication bias: tendency of articles reporting positive and/or "new" results to be published, and studies with negative or confirmatory results to not be submitted or published; especially important in meta-analysis, but also in other systematic reviews. Substantial publication bias has been demonstrated from the "file-drawer" problem.[34]

purposive sample: set of observations obtained from a population in such a way that the sample distribution of independent variable values is determined by the researcher and is not necessarily representative of distribution of the values in the population.[23(p334)]

***P* value:** probability of obtaining the observed data (or data that are more extreme) if the null hypothesis were exactly true.[27(p206)]

⇒While hypothesis testing often results in the *P* value, *P* values themselves can only provide information about whether the null hypothesis is accepted or rejected. Confidence intervals (CIs) are much more informative since they provide a plausible range of values for an unknown parameter, as well as some indication of the power of the study as indicated by the width of the CI.[21(pp186-187)] (For example, an odds ratio of 0.5 with a 95% CI of 0.05-4.5 indicates to the reader the [im]precision of the estimate, whereas $P = .63$ does not provide such information.) Confidence intervals are preferred whenever possible. Including both the CI and the *P* value provides more information than either alone.[21(p187)] This is especially true if the CI is used to provide an interval estimate and the *P* value to provide the results of hypothesis testing.

⇒ When any *P* value is expressed, it should be clear to the reader what parameters and groups were compared, what statistical test was performed, and the degrees of freedom (*df*) and whether the test was 1-tailed or 2-tailed (if these distinctions are relevant for the statistical test).

⇒ For expressing *P* values in manuscripts and articles, the actual value for *P* should be expressed to 2 digits for $P \geq .01$, whether or not *P* is significant. (When rounding a *P* value expressed to 3 digits would make the *P* value nonsignificant, such as $P = .049$ rounded to .05, the *P* value can be left as 3 digits.) If $P < .01$,

P should be expressed to 3 digits. The actual *P* value should be expressed ($P = .04$), rather than expressing a statement of inequality ($P < .05$), unless $P < .001$. Expressing *P* to more than 3 significant digits does not add useful information to $P < .001$, since precise *P* values with extreme results are sensitive to biases or departures from the statistical model.[21(p198)]

P values should not be listed simply as not significant (NS), since for meta-analysis the actual values are important and not providing exact *P* values is a form of incomplete reporting.[21(p195)] Because the *P* value represents the result of a statistical test and not the strength of the association or the clinical importance of the result, *P* values should be referred to simply as statistically significant or not significant; terms such as highly significant or very highly significant should be avoided.

⇒The AMA style does not use a zero to the left of the decimal point, since statistically it is not possible to prove or disprove the null hypothesis completely when only a sample of the population is tested (*P* cannot equal 0 or 1, except by rounding). If $P < .00001$, *P* should be expressed as $P < .001$ as discussed. If $P > .999$, *P* should be expressed as $P > .99$.

qualitative data: data that fit into discrete categories according to their attributes, such as nominal or ordinal data, as opposed to quantitative data.[25(p136)]

qualitative study: form of study based on observation and interview with individuals that uses inductive reasoning and a theoretical sampling model, with emphasis on validity rather than reliability of results. Qualitative research is used traditionally in sociology, psychology, and group theory, but also occasionally in clinical medicine to explore beliefs and motivations of patients and physicians.[35]

quality-adjusted life-year (QALY): method used in economic analyses to reflect the existence of chronic conditions that cause impairment, disability, and loss of independence. Numerical weights representing severity of residual disability are based on assessments of disability by study subjects, parents, physicians, or other researchers made as part of utility analysis.[25(p136)]

quantile: method used for grouping and describing dispersion of data. Commonly used quantiles are the tertile (3 equal divisions of data into lower, middle, and upper ranges), quartile (4 equal divisions of data), quintile (5 divisions), and decile (10 divisions). Quantiles are also referred to as *percentiles*.[22(p165)]

⇒ Data may be expressed as median (quantile range), eg, length of stay was 7.5 days (interquartile range, 4.3-9.7 days). See also interquartile range.

quantitative data: data in numerical quantities such as continuous data or counts[25(p137)] (as opposed to qualitative data).

quasi-experiment: experimental design in which variables are specified and subjects assigned to groups, but interventions cannot be controlled by the experimenter. One type of quasi-experiment is the natural experiment.[25(p137)]

***r*:** correlation coefficient for bivariate analysis.

***R*:** correlation coefficient for multivariate analysis.

r^2: coefficient of determination for bivariate analysis. See also correlation coefficient.

R^2: coefficient of determination for multivariate analysis. See also correlation coefficient.

random-effects model: model used in meta-analysis that assumes that there is a universe of conditions and that the effects seen in the studies are only a sample, ideally a random sample, of the possible effects.[29(p349)] Antonym is fixed-effects model.

randomization: method of assignment in which individuals have a random chance of being assigned to a particular study or control. Individuals may be randomly assigned at a 2:1 or 3:1 frequency, in addition to the usual 1:1 frequency. Subjects may or may not be representative of a larger population.[23(p334)] Simple methods of randomization include coin flip or use of a random numbers table. See also block randomization.

randomized controlled trial: see 17.2.1, Randomized Controlled Trial.

random sample: method of obtaining a sample that ensures that every individual in the population has a known (but not necessarily equal, eg, in weighted sampling techniques) chance of being selected for the sample.[23(p335)]

range: the highest and lowest values of a variable measured in a sample.

Example: The mean age of the participants was 45.6 years (range, 20-64 years).

rank sum test: see Mann-Whitney test or Wilcoxon rank sum test.

rate: measure of the occurrence of a disease or outcome per unit of time, usually expressed as a decimal if the denominator is 100 (eg, the surgical mortality rate was 0.02). See also 16.7.3, Numbers and Percentages, Proportions and Rates.

ratio: fraction in which the numerator is not necessarily a subset of the denominator, unlike a proportion[23(p335)] (eg, the assignment ratio was 1:2:1 for each drug dose [twice as many individuals were assigned to the second group as to the first and third groups]).

recall bias: systematic error resulting from individuals in one group being more likely than individuals in the other group to remember past events.[25(p141)]

⇒Recall bias is especially common in case-control studies that assess risk factors for serious illness in which individuals are asked about past exposures or behaviors, such as environmental exposure in an individual who has cancer.[23(p335)]

receiver operating characteristic curve (ROC curve): graphic means of assessing the extent to which a screening test can be used to discriminate between persons with and without disease,[25(p142)] and to select an appropriate cut point for defining normal vs abnormal results. The ROC curve is created by plotting sensitivity vs (1 - specificity). The area under the curve provides some

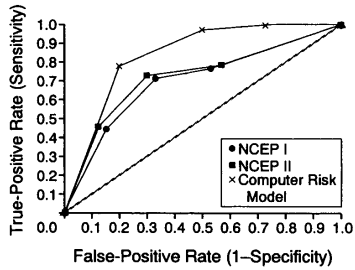


FIGURE 4 Receiver operating characteristic curve. The 45° line represents the point at which the test is no better than chance. The area under the curve measures the performance of the test; the larger the area under the curve, the better the test performance. Adapted from Grover SA, Coupal L, Hu X-P. Identifying adults at increased risk of coronary disease: how well do the current cholesterol guidelines work? *JAMA.* 1995;274:801-806.

measure of how well the test performs; the larger the area, the better the test. See Figure 4.

⇒ The appropriate cut point is a function of the test. A screening test would require high `sensitivity`, whereas a diagnostic test would require high `specificity`. See Table 2 and `diagnostic discrimination`.

reference group: group of presumably disease-free individuals from which a sample of individuals is drawn and tested to establish a range of normal values for a test.[23(p335)]

regression analysis: statistical techniques used to describe a `dependent variable` as a function of 1 or more `independent variables`; often used to control for `confounding variables`.[23(p335)] See also `linear regression`, `logistic regression`.

regression line: diagrammatic presentation of a `linear regression` equation, with the `independent variable` plotted on the x-axis and the `dependent variable` plotted on the y-axis. As many as 3 variables may be depicted on one graph.[25(p145)]

regression to the mean: principle that unusual events are unlikely to recur. A common example is blood pressure measurement; on repeated measurements individuals who are initially hypertensive often will have a blood pressure closer to the population mean than the initial measurement was.[23(p335)]

relative risk (RR): probability of developing an outcome within a specified period if a risk factor is present, divided by the probability of developing the outcome in that same period if the risk factor is absent. The relative risk is applicable to randomized clinical trials and `cohort studies`[23(p335)]; for `case-control studies` the `odds ratio` can be used to approximate the relative risk if the outcome is infrequent.

⇒ The relative risk should be accompanied by confidence intervals.

Example: The individuals with untreated mild hypertension had a relative risk of 2.4 (95% confidence interval, 1.9-3.0) for stroke or transient ischemic attack.

[In this example, individuals with untreated mild hypertension were 2.4 times more likely than the rest of the cohort to have a stroke or transient ischemic attack.]

relative risk reduction: proportion of the control group experiencing a given outcome minus the proportion of the treatment group experiencing the outcome, divided by the proportion of the control group experiencing the outcome.

reliability: ability of a test to replicate a result given the same measurement conditions, as distinguished from validity, which is the ability of a test to measure what it is intended to measure.[25(p145)]

repeated measures: analysis designed to take into account the lack of independence of events when measures are repeated in each subject over time (eg, blood pressure, weight, or test scores). This type of analysis emphasizes the change measured for a subject over time, rather than the differences between subjects over time.

repeated-measures ANOVA: see analysis of variance.

reporting bias: a bias in assessment that can occur when individuals in one group are more likely than individuals in another group to report past events. Reporting bias is especially likely to occur when different groups have different reasons to report or not report information.[23(pp335-336)] For example, when examining behaviors, adolescent girls may be less likely than adolescent boys to report being sexually active. See also recall bias.

reproducibility: ability of a test to produce consistent results when repeated under the same conditions and interpreted without knowledge of the first test results[23(p336)]; same as reliability.

residual: measure of the discrepancy between observed and predicted values. The residual SD is a measure of the goodness of fit of the regression line to the data and gives the uncertainty of estimating a point y from a point x.[22(p176)]

response rate: number of individuals who respond to a survey divided by the number of individuals who are contacted for the survey, usually expressed as a percentage.

⇒In general, response rates of less than 60% may not reflect the population surveyed and results may be unreliable.

retrospective study: study performed after the outcomes of interest have already occurred[25(p147)]; most commonly a case-control study, but also may be a retrospective cohort study or case series.

right-censored data: see censored data.

risk: probability that an event will occur during a specified period. Risk is equal to the number of individuals who develop the disease during the period divided by the number of disease-free persons at the beginning of the period.[23(p336)]

risk factor: characteristic or factor that is associated with an increased probability of developing a condition or disease. Also called a risk marker, a risk factor does not necessarily imply a causal relationship. A modifiable risk factor is one that can be modified through an intervention[25(p148)] (eg, stopping smoking or treating elevated cholesterol level, as opposed to a genetically linked characteristic for which there is no effective treatment).

robust: term used to indicate that a statistical procedure's assumptions (most commonly, normal distribution of data) can be violated without a substantial effect on its conclusions.[25(p149)]

root-mean-square: see `standard deviation`.

rule of three: method used to estimate the number of observations required to have a 95% chance of observing at least 1 episode of a serious adverse effect. For example, to observe at least 1 case of penicillin anaphylaxis that occurs in about 1 in 10 000 cases treated, 30 000 treated cases must be observed. If an adverse event occurs 1 in 15 000 times, 45 000 cases need to be treated and observed.[23(p114)]

sample: subset of a larger population, selected for investigation to draw conclusions or make estimates about the larger population.[34(p336)]

sampling error: error introduced by chance differences between the estimate obtained from the sample and the true value in the population from which the sample was drawn. Sampling error is inherent in the use of sampling methods and is measured by the `standard error`.[23(p336)]

Scheffé test: see `multiple comparisons procedures`.

SD: see `standard deviation`.

SE: see `standard error`.

SEE: see `standard error of the estimate`.

selection bias: bias in assignment that occurs when the way the study and control groups are chosen causes them to differ from each other by at least 1 factor that affects the outcome of the study.[23(p336)]

⇒A common type of selection bias occurs when individuals from the study group are drawn from one population (eg, patients seen in an emergency department or admitted to a hospital) and the control subjects are drawn from another (eg, clinic patients). Regardless of the disease under study, the clinic patients will be healthier overall than the patients seen in the emergency department or hospital and will not be comparable controls.

SEM: see `standard error of the mean`.

sensitivity: proportion of those with the disease or condition as measured by the `criterion standard` who have a positive test result (true positives divided by all positives).[23(p336)] See Table 2 and `diagnostic discrimination`.

sensitivity analysis: method to determine the robustness of an assessment by examining the extent to which results are changed by differences in methods,

values of variables, or assumptions[25(p154)]; applied in decision analysis to test the robustness of the conclusion to changes in the assumptions.

signed rank test: see Wilcoxon signed rank test.

significance: statistically, the testing of an hypothesis that an effect is not present. A significant result rejects the null hypothesis. Statistical significance is highly dependent on sample size and provides no information about the clinical significance of the result. Clinical significance, on the other hand, involves a judgment as to whether the risk factor or intervention studied would affect a patient's outcome enough to make the intervention worthwhile. The level of clinical significance considered important is sometimes defined prospectively (often by consensus of a group of physicians) as the minimal clinically important difference, but the cutoff is arbitrary.

sign test: a nonparametric test of significance that depends on the signs (positive or negative) of variables and not on their magnitude; used when combining the results of several studies, as in meta-analysis.[25(p156)] See also Cox-Stuart trend test.

skewed distribution: asymmetric frequency distribution. Data for a given variable with a longer tail on the right of the distribution curve are referred to as positively skewed; data with a longer left tail are negatively skewed.[27(pp238-239)]

Spearman rank correlation (p): statistical test used to determine the covariance between 2 nominal or ordinal variables.[27(p243)] The nonparametric equivalent to the Pearson product moment correlation, it can also be used to calculate the coefficient of determination.

specificity: proportion of those without the disease or condition as measured by the criterion standard who have negative results by the test being studied[23(p326)] (true negatives divided by all negatives). See Table 2 and diagnostic discrimination.

standard deviation (SD; does not need to be expanded at first mention): commonly used descriptive measure of the spread or dispersion of data; the positive square root of the variance.[23(p336)] The mean ±2 SDs represents the middle 95% of values obtained.

⇒Describing data by means of SD implies that the data are normally distributed; if not, then the interquartile range or a similar measure is more appropriate to describe the data. If the interquantile range cannot be provided, and particularly if the mean ±2 SDs would be impossible (eg, length of stay 9 ± 15 days or age at evaluation 4 ± 5.3 days), it is preferable to use the format mean (SD) rather than the ± construction.[36]

standard error (SE; does not need to be expanded at first mention): positive square root of the variance of the sampling distribution of the statistic.[22(p195)] There are several types of SE; the type intended should be clear.

⇒The SE is not interchangeable with SD. The SD is a descriptive statistic; SE is an inferential statistic. In text and tables that provide descriptive statistics, SD is usually appropriate; in figures where the SE is frequently depicted for error bars, the 95% confidence interval is preferred (see 2.14, Manuscript Preparation, Figures).[37]

standard error of the difference: measure of the dispersion of the differences between samples of 2 populations, usually the differences between the means of 2 samples; used in the *t* test.

standard error of the estimate (SEE): SD of the observed values about the regression line.[22(p195)]

standard error of the mean (SEM): quantification of the certainty with which the mean computed from a random sample estimates the true mean of the population from which the sample was drawn.[33(p21)] If multiple samples of a population were taken, then 95% of the sample means (equal to the 95% confidence interval) would fall within the sample mean ±2 SEMs.

standard error of the proportion: SD of the population of all possible values of the proportion computed from samples of a given size.[33(p109)]

standardization (of a rate): adjustment of a rate to account for factors such as age or sex[23(pp336-337)]; also referred to as *age-adjusted rate*.

standardized mortality ratio: ratio in which the numerator contains the observed number of deaths and the denominator contains the number of deaths that would be expected in a comparison population. This ratio implies that confounding factors have been controlled for by means of indirect standardization. It is distinguished from proportionate mortality ratio, which is the mortality rate for a specific disease.[23(p337)]

standard normal distribution: normal distribution in which the mean has a *z* score of 0 and the SD has a *z* score of 1.[27(p245)] By definition, 68% of the curve is contained within 1 SD and 95% within 2 SDs. The mean, median, and mode are equal.

standard score: *z* score.[22(p196)]

statistic: value calculated from sample data that is used to estimate a value or parameter in the larger population from which the sample was obtained,[23(p337)] as distinguished from data, which refers to the actual values obtained via measurement, chart review, patient interview, and the like.

stochastic: type of measure that implies the presence of a random variable.[22(p197)]

stopping rule: rule, based on a test statistic or other function, specified as part of the design of the trial and established before patient enrollment, that specifies a limit for the observed treatment difference for the primary outcome measure, which, if exceeded, will lead to the termination of the trial or one of the study arms.[3(p258)] The stopping rules are designed to ensure that a study does not continue to enroll patients after a significant treatment difference has been demonstrated that would still exist regardless of the treatment results of subsequently enrolled patients.

stratification: division into groups. Stratification may be used to compare groups separated according to similar confounding characteristics. Stratified sampling may be used to increase the number of individuals sampled in rare categories of

independent variables, or to obtain an adequate sample size to examine differences among individuals with certain characteristics of interest.[17(p337)] One example is stratified sampling of blacks and Hispanics in epidemiologic studies to ensure that adequate numbers of individuals are included for comparisons of study characteristics by race.

Student-Newman-Keuls test: see Newman-Keuls test.

Student *t* test: see *t* test. Student is not the name of the originator of the test; W. S. Gossett wrote under the name Student since his employment precluded individual publication.[25(p166)]

study group: in a controlled clinical trial, the group of individuals who undergo an intervention; in a cohort study, the group of individuals with the exposure or characteristic of interest; and in a case-control study, the group of cases.[23(p337)]

sufficient cause: characteristic that will bring about or cause the disease.[23(p337)]

supportive criteria: substantiation of the existence of a contributory cause. Potential supportive criteria include the strength and consistency of the relationship, the presence of a dose-response relationship, and biological plausibility.[23(p337)]

survey: method of study that depends on self-report, conducted with the use of a form mailed or otherwise distributed to individuals, or completed by an interviewer in person or on the telephone. A survey is used to collect information regarding an individual's demographic characteristics, medical history, attitudes, knowledge, and behaviors.[25(p163)] Methods that describe a survey should include how the survey was developed and performed, how the sample was selected, and the response rate.

⇒The most important factors to consider when a survey is assessed include the validity of the survey instrument, how the study sample was obtained, and the response rate. Measures of the validity of the survey instrument should be included in the methods section. The method of sampling (patients attending a clinic, individuals with a telephone, people responding to a mailing) affects the generalizability of the results. A low response rate (eg, <60%) may also affect the generalizability of results, since those surveyed may not be representative of the population regardless of the representativeness of the initial sample.

⇒If more than one attempt to reach individuals was made, each response rate should be reported individually. If the overall response rate is low, a comparison can be made between the different waves of respondents. If significant differences exist between groups of respondents, the individuals who responded may not be representative of the overall group that was surveyed.

survival analysis: statistical procedures for estimating the survival function and for making inferences about how it is affected by treatment and prognostic factors.[25(p163)] See life table.

target population: group of individuals to whom one wishes to apply or extrapolate the results of an investigation, not necessarily the population studied.[23(p337)] If

the target population is different from the population studied, whether the study results can be extrapolated to the target population should be discussed.

τ (tau): see Kendall τ rank correlation.

trend, test for: see χ^2 test.

trial: controlled experiment with an uncertain outcome[22(p208)]; used most commonly to refer to a randomized study.

true negative: negative test result in an individual who does not have the disease or condition as determined by the criterion standard.[23(p338)] See also Table 2.

true-negative rate: number of individuals who have a negative test result and do not have the disease by the criterion standard, divided by the total number of individuals who do not have the disease as determined by the criterion standard; usually expressed as a decimal (eg, the true-negative rate was 0.85). See also Table 2.

true positive: positive test result in an individual who has the disease or condition as determined by the criterion standard.[23(p338)] See also Table 2.

true-positive rate: number of individuals who have a positive test result and have the disease as determined by the criterion standard, divided by the total number of individuals who have the disease as measured by the criterion standard; usually expressed as a decimal (eg, the true-positive rate was 0.92). See also Table 2.

***t* test:** statistical test used when the dependent variable is continuous and the parameter of interest is the location of the variable. Use of the *t* test assumes that the variable has a normal distribution; if not, nonparametric statistics must be used.[23(p266)]

⇒ Usually the *t* test is unpaired, unless the data have been measured in the same individual over time. A paired *t* test is appropriate to assess the change of the parameter in the individual from baseline to final measurement; in this case the dependent variable is the change from one measurement to the next.

⇒ Presentation of the *t* statistic should include the degrees of freedom (*df*), whether the *t* test was paired or unpaired, and whether a 1- or 2-tailed test was used. Since a 1-tailed test assumes that the study effect has only 1 possible direction (ie, either beneficial or harmful), justification for use of the 1-tailed test must be provided. (The 1-tailed test is similar to testing at $\alpha = .10$ for a 2-tailed test and therefore is more likely to give a significant result.)

> *Example:* The difference was significant by a 2-tailed test for paired samples ($t_{15} = 2.78$, $P = .05$).

⇒ The *t* test can also be used to compare different coefficients of variation.

Tukey test: a type of multiple comparisons procedure.

2-tailed test: test of statistical significance in which deviations from the null hypothesis in either direction are considered.[23(p338)] For most outcomes, the 2-

tailed test is appropriate unless there is a plausible reason why only 1 direction of effect is considered and a 1-tailed test is appropriate. Commonly used for the *t* test.

2-way analysis of variance: see analysis of variance.

type I error: data demonstrating a statistically significant result, although no true association or difference exists in the population.[23(p338)] The α level is the size of a type I error that will be permitted, usually .05.

⇒A frequent cause of a type I error is performing multiple comparisons, which increase the likelihood that a significant result will be found by chance. To avoid a type I error, one of several multiple comparisons procedures can be used.

type II error: failure of the data to demonstrate a statistically significant result although a true association or difference exists in the population.[23(p338)]

⇒A frequent cause of a type II error is insufficient sample size. Therefore, a power calculation should be performed when a study is planned to determine the sample size needed to avoid a type II error.

uncensored data: continuous data reported as collected, without adjustment, as opposed to censored data.

uniform prior: assumption that no useful information regarding the outcome of interest is available prior to the study. See Bayesian analysis.

unity: number 1; a relative risk of 1 is a relative risk of unity, and a regression line with a slope of 1 is said to have a slope of unity.

univariable analysis: another name for univariate analysis.

univariate analysis: statistics involving 1 dependent variable and no independent variables; uses measures of central tendency (mean or median) and location or dispersion. The purpose of the analysis is to describe the sample, determine how the sample compares with the population, and determine whether chance has skewed 1 or more of the variables in the study. If the characteristics of the sample do not reflect those of the population from which the sample was drawn, the results may not be generalizable to the population.[23(pp245-246)]

unpaired analysis: method that compares 2 treatment groups when the 2 treatments are not given to the same individual. Most case-control studies also use unpaired analysis..

unpaired *t* test: see *t* test.

***U* test:** see Wilcoxon rank sum test.

utility: in decision theory and clinical decision analysis, a scale used to judge the importance of achieving a particular outcome (used in studies to quantify the importance of an outcome vs the discomfort of the intervention to a patient) or the discomfort experienced by the patient with a disease.[25(p170)] Commonly used

methods are the *time trade-off* and the *standard gamble*. The result is expressed as a single number along a continuum from death (0) to full health or absence of disease (1.0). This quality number can then be multiplied by the number of years a patient is in the health state produced by a particular treatment to obtain the quality-adjusted life-year. See also 17.2.8, Cost-effectiveness Analysis, Cost-benefit Analysis.

validity (of a measurement): degree to which a measurement is appropriate for the question being addressed or measures what it is intended to measure. For example, a test may be highly consistent and reproducible over time, but unless it is compared with a criterion standard or other validation method, the test cannot be considered valid (see also diagnostic discrimination). *Construct validity* refers to the extent to which the measurement corresponds to theoretical concepts (eg, a measure thought to change over time does change). *Content validity* is the extent to which the measurement incorporates the domain under study (eg, a measurement to assess delirium evaluates cognition). *Criterion validity* is the extent to which the measurement is correlated with an external criterion of the phenomenon under study. Validity can be *concurrent* (assessed simultaneously) or *predictive* (eg, ability of a standardized test to predict school performance).[25(p171)]

⇒Validity of a test is sometimes mistakenly used as a synonym of reliability; the two are distinct statistical concepts and should not be used interchangeably.

validity (of a study): *internal validity* means that the observed differences between the control and comparison groups may, apart from sampling error, be attributed to the effect under study; *external validity* or generalizability means that a study can produce unbiased inferences regarding the target population, beyond the subjects in the study.[25(p171)]

Van der Waerden test: nonparametric test that is sensitive to differences in location for 2 samples from otherwise identical populations.[22(p216)]

variable: characteristic measured as part of a study. Variables may be dependent (usually the outcome of interest) or independent (characteristics of individuals that may affect the dependent variable).

variance: variation measured in a set of data for 1 variable, defined as the sum of squares of the deviation of each data point from the mean for the data, divided by the *df* (sample observation − 1).[27(p266)]

variance components analysis: process of isolating the sources of variability in the outcome variable for the purpose of analysis.

variance ratio distribution: synonym for F distribution.[25(p61)]

visual analog scale: scale used to quantify subjective factors such as pain or satisfaction. Subjects are asked to indicate where their current feelings fall by marking a straight line with 1 extreme, such as "worst pain ever experienced," at 1 end of the scale and the other extreme, such as "pain-free," at the other end. The feeling (eg, degree of pain) is quantified by measuring the distance from the mark on the scale to the end of the scale.[25(p268)]

TABLE 3. STATISTICAL METHODS FREQUENTLY USED TO TEST HYPOTHESES*

Scale of Measurement	2 Treatment Groups	3 or More Treatment Groups	Before and After 1 Treatment in the Same Individual	Multiple Treatments in the Same Individual	Association Between 2 Variables
Interval (assumes normally distributed data)†	Unpaired *t* test	Analysis of variance	Paired *t* test	Repeated-measures analysis of variance	Linear regression and Pearson product moment correlation
Nominal‡	χ^2 analysis-of-contingency table	χ^2 analysis-of-contingency table	McNemar test	Cochran *Q*	Contingency coefficients
	Fisher exact test if ≤ 6 in any cell	Fisher exact if ≤ 6 in any cell	Wilcoxon signed rank test	Friedman statistic	Spearman rank correlation
Ordinal	Mann-Whitney rank sum test	Kruskal-Wallis statistic			

* *Adapted with permission from Glantz SA.* Primer of Biostatistics. *2nd ed. New York, NY: McGraw-Hill Book Co Inc; 1981.*[33]

† *If data are not normally distributed, then rank the observations and use the methods for data measured on an ordinal scale.*

‡ *For a nominal dependent variable that is time dependent (such as mortality over time), use life-table analysis for nominal independent variables and Cox regression for continuous and/or nominal independent variables.*

Wilcoxon rank sum test: a nonparametric test that ranks and sums observations from combined samples and compares the result with the sum of ranks from 1 sample.[22(p20)] U is the statistic that results from the test. Alternative name for the Mann-Whitney test.

Wilcoxon signed rank test: nonparametric test in which 2 treatments that have been evaluated by means of matched samples are compared. Each observation is ranked according to size and given the sign of the treatment difference (ie, positive if the treatment effect was positive and vice versa) and the ranks are summed.[22(p220)]

Wilks λ (lambda): a test used in multivariate analysis of variance (MANOVA).

x-axis: horizontal axis of a graph. By convention, the dependent variable is plotted on the x-axis.

Yates correction: continuity correction used to bring a distribution based on discontinuous frequencies closer to the continuous χ^2 distribution from which χ^2 tables are derived.[25(p176)]

y-axis: vertical axis of a graph. By convention, the independent variable is plotted on the y-axis.

z-axis: third axis of a 3-dimensional graph, generally placed so that it appears to project out toward the reader. The z- and y-axes are both used to plot independent variables and are often used to demonstrate that the 2 independent variables each contribute independently to the dependent variable. See x-axis and y-axis.

***z* score:** score used to analyze continuous variables that represents the deviation of a value from the mean value, expressed as the number of SDs from the mean. This score is frequently used to compare children's height and weight measurements and for behavioral scores.[25(p176)]

■ **STATISTICAL SYMBOLS AND ABBREVIATIONS.**—The following may be used without expansion except where noted by an asterisk. For a term expanded at first mention, the abbreviation may be placed in parentheses after the expanded term and the abbreviation used thereafter (see also 11.11, Abbreviations, Clinical and Technical Terms). Most terms other than mathematical symbols can also be found in 17.4, Glossary of Statistical Terms.

Symbol or Abbreviation	*Description*
$\|x\|$	absolute value
Σ	sum
$>$	greater than
$\geq$	greater than or equal to
$<$	less than
$\leq$	less than or equal to

Symbol or Abbreviation	*Description*
^	hat, used to denote an estimate
ANOVA	analysis of variance*
ANCOVA	analysis of covariance*
α	alpha, probability of type I error
$1 - \alpha$	confidence coefficient
β	beta, probability of type II error; or population regression coefficient
$1 - \beta$	power of a statistical test
b	sample regression coefficient
CI	confidence interval*
CV	coefficient of variation $(s/\bar{x}) \times 100$*
D	difference
df	degrees of freedom (v is the international symbol[38] and also may be used if familiar to readers)
D^2	Mahalanobis distance, distance between the means of 2 groups
Δ	delta, change
δ	delta, true sampling error
ε	epsilon, true experimental error
e	exponential
$E(x)$	expected value of the variable x
f	frequency; or a function of, usually followed by an expression in parentheses, eg, $f(x)$
$F_{v_1, v_2}(1 - \alpha)$	F test, ratio of 2 variances, with $df = v_1$, v_2 for numerator and denominator, respectively, and $(1 - \alpha)$ = confidence coefficient
$G^2(df)$	likelihood ratio χ^2
H_0	null hypothesis
H_1	alternate hypothesis; specify whether 1- or 2-sided
κ	kappa statistic
λ_i	lambda, hazard function for interval i; eigenvalue; or estimate of parameter for log-linear models
Λ	Wilks lambda
ln	natural logarithm
log	logarithm to base 10
MANOVA	multivariate analysis of variance*
μ	population mean
n	size of a subsample

Symbol or Abbreviation	*Description*
N	total sample size
$n!$	(n) factorial
OR	odds ratio*
P	statistical probability
χ^2	Yates corrected χ^2 (1 df)
χ^2	χ^2 test
r	bivariate correlation coefficient
R	multivariate correlation coefficient
r^2	bivariate coefficient of determination
R^2	multivariate coefficient of determination
RR	relative risk*
ρ	rho, population correlation coefficient
S_D	standard deviation of a difference D
s^2	sample variance
σ^2	sigma squared, population variance
σ	sigma, population SD
SD	standard deviation of a sample
SE	standard error
SEM	standard error of the mean
t	Student t; specify α level, df, 1-tailed vs 2-tailed
τ	Kendall tau
T^2	Hotelling T^2 statistic
U	Mann-Whitney U (Wilcoxon) statistic
$\|x\|$	arithmetic mean
z	z score

U.S. Medical & Scientific Affairs

MEDICAL SERVICES

U.S. Human Health

MSP 00-0011-ST 2/2000