Radiology Statistical Concepts Series
(November 2002 – March 2004)

Contents *(Page numbers in PDF file)*

Anthony V. Proto, MD, Editor

Radiology 2002—Statistical Concepts Series[1]

As of January 2001, we began review by statisticians of all manuscripts that have statistical content and that are to be published in *Radiology* (1). Although I believed, from its inception, this form of manuscript evaluation to be an essential component of the peer-review process, my belief has been further confirmed over the past many months. One of the most common comments by our statistical reviewers is that authors have selected inappropriate statistical tests for the analysis of their data. We urge authors to consult with statisticians regarding the analysis of their data. It is particularly important that a study be designed and data be collected in a manner that will allow the study hypothesis to be adequately evaluated. Statistical consultation in the study-planning stages can help ensure success in this regard.

Index terms:
Editorials
Radiology (journal)

Published online before print
10.1148/radiol.2252020996
Radiology 2002; 225:317
See also the article by Applegate and Crewson in this issue.

With the November 2002 issue of *Radiology,* we begin a special series of articles that will appear in the section entitled Statistical Concepts Series. As I announced earlier this year (2), we are indebted to Kimberly E. Applegate, MD, MS, and Philip E. Crewson, PhD, for coordinating this series. Dr Applegate, an RSNA Editorial Fellow in the year 2000, is currently associate professor of Radiology and Health Services Research at Indiana University, Indianapolis. Dr Crewson, formerly director of Clinical Studies, Research Development, at the American College of Radiology, is currently assistant director of Scientific Development, Health Services Research, and Development Service at the Office of Research and Development, Department of Veterans Affairs, Washington, DC. Both Dr Applegate and Dr Crewson have expended a substantial amount of time and effort in selecting topics for this series, identifying the authors for the various topics, and working with the authors to ensure an appropriate level of depth of coverage for each topic without undue overlap with other topics in the series. After review of their manuscripts by Drs Applegate and Crewson, authors submitted the manuscripts to the *Radiology* Editorial Office for peer review.

We hope that authors, reviewers, and readers will find this series of articles helpful—authors with regard to the design of their studies and the analysis of their data, reviewers with regard to their evaluation and critique of manuscripts during the peer-review process, and readers with regard to improved understanding and interpretation of articles published in *Radiology*. Since we have established the section title Statistical Concept Series for these articles, they will be accessible through *Radiology* Online *(radiology.rsnajnls.org)* by clicking first on Browse by Subspecialty and Category (Radiology Collections) and second on this section title. As noted by Drs Applegate and Crewson in their article "An Introduction to Biostatistics," the first in this series and published in the current issue of the Journal (3), "These articles are meant to increase understanding of how statistics can and should be applied in radiology research so that radiologists can appropriately interpret the results of a study."

References

1. Proto AV. Radiology 2001—the upcoming year. Radiology 2001; 218:1–2.
2. Proto AV. Radiology 2002—continued progress. Radiology 2002; 222:1–2.
3. Applegate KE, Crewson PE. An introduction to biostatistics. Radiology 2002; 225: 318–322.

Statistical Concepts Series

Kimberly E. Applegate, MD, MS
Philip E. Crewson, PhD

Index terms:
Radiology and radiologists, research
Statistical analysis

Published online before print
10.1148/radiol.2252010933
Radiology 2002; 225:318–322

An Introduction to Biostatistics[1]

This introduction to biostatistics and measurement is the first in a series of articles designed to provide *Radiology* readers with a basic understanding of statistical concepts. Although most readers of the radiology literature know that application of study results to their practice requires an understanding of statistical issues, many may not be fully conversant with how to interpret statistics. The goal of this series is to enhance the ability of radiologists to evaluate the literature competently and critically, not make them into statisticians.
© RSNA, 2002

There are three kinds of lies: lies, damned lies and statistics.
Benjamin Disraeli (1)

The use of statistics in both radiology journals and the broader medical literature has become a common feature of published clinical research (2). Although not often recognized by the casual consumer of research, errors in statistical analysis are common, and many believe that as many as 50% of the articles in the medical literature have statistical flaws (2). Most radiologists, however, are poorly equipped to properly interpret many of the statistics reported in the radiology literature. There are a number of reasons for this problem, but the reality for the radiology profession is that research methods have long been a low priority in the training of radiologists (3–5). Contributing to this deficit is a general indifference toward statistical teaching in medical school and physician training, insufficient numbers of statisticians, and limited collaboration and understanding between radiologists and statisticians (2,6,7).

If it has traditionally been such a low priority in the profession, why then do we need to improve our understanding of statistics? We are consumers of information. Statistics allow us to organize and summarize information and to make decisions by using only a sample of all available data. Nearly all readers of the radiology literature know that understanding a study's results and determining the applicability of the results to their practice requires an understanding of statistical issues (5). Even when learned, however, research skills can be quickly forgotten if not applied on a regular basis—something most radiologists are unlikely to do, given their increasing clinical demands.

This introduction to biostatistics and measurement is the first in a series of articles designed to provide *Radiology* readers with a basic understanding of statistical concepts. These articles are meant to increase understanding of how statistics can and should be applied in radiology research so that radiologists can appropriately interpret the results of a study. Each article will provide a short summary of a statistical topic. The series begins with basic measurement issues and progress from descriptive statistics to hypothesis testing, multivariate models, and selected technology-assessment topics. Key concepts presented in this series will be directly related to the practice of radiology and radiologic research. In some cases, formulas will be provided for those who wish to develop a deeper understanding; however, the goal of this series is to enhance the ability of radiologists to evaluate the literature competently and critically, not to make them statisticians.

The concepts presented in this introductory article are important for putting into perspective the substantive value of published research. Appendices A and B include two useful resources. One is a list of the common terms and definitions related to measurement. The other is a list of potentially useful Web resources. This list contains Web sites

[1] From the Department of Radiology, Riley Hospital for Children, 702 Barnhill Rd, Indianapolis, IN 46202 (K.E.A.); and the Department of Research, American College of Radiology, Reston, Va (P.E.C.). Received May 16, 2001; revision requested July 6; revision received September 5; accepted September 24. **Address correspondence to** K.E.A. (e-mail: *kiappleg@iupui.edu*).

See also the From the Editor article in this issue.

that are either primarily educational or have links to other resources such as statistical software. In addition, some suggested additional readings are listed in Appendix C.

ORIGINS OF STATISTICS

The term *statistic* simply means "numeric data." In contrast, the field of statistics is a human enterprise encompassing a wide range of methods that allow us to learn from experience (8,9). Tied to the emergence of the scientific method in the 16th and 17th centuries, statistical thinking involves deduction of explanations of reality and framing of these explanations into testable hypotheses. The results of these tests are used to reach conclusions (inferences) about future outcomes on the basis of past experiences.

Statistical thinking is probabilistic. For example, most radiologists are familiar with the notion that a *P* value of less than .05 represents a statistically significant result. Few understand that .05 is an arbitrary threshold. While performing agronomy research, Sir Ronald Fisher helped to establish the *P* value cutoff level of .05 (commonly referred to as the α level) in the early part of the 20th century. Fisher (10) was testing hypotheses about appropriate levels of fertilizer for potato plants and needed a basis for decision making. Today we often use this same basis for testing hypotheses about appropriate patient care.

The health profession gradually recognized that statistics were as applicable to people as they were to potato plants. By the 1960s, clinicians and health policy leaders were asking for statistical evidence that an intervention was effective (11). Over the past several decades, the use of statistics in medical journals has increased both in quantity and in sophistication (12,13). Advances in computers and statistical software have paralleled this increase. However, the benefits of easy access to the tools of statistical analysis can be overshadowed by the costs associated with misapplication of statistical methods. Statistical software makes it far too easy to conduct multiple tests of data without prior hypotheses (the so-called data-mining phenomenon) or to report overly precise results that portend a false sense of accuracy. There is also the potential for errors in statistical software and the ever-present risk that researchers will fail to take the time to carefully look at the raw data (14). These issues can result in poor science, erroneous or misleading results, and inappropriate patient care. The benefits of statistical software generally far outweigh the costs, but proper measurement, study design, and good judgment should prevail over the ease with which many analyses can be conducted. What follows is an introduction to the basics of measurement.

MEASUREMENTS: BUILDING BLOCKS OF STATISTICS

The interpretation and use of statistics require a basic understanding of the fundamentals of measurement. Although most readers of the radiology literature will recognize common terms such as variables, association, and causation, few are likely to understand how these terms interrelate with one another to frame the structure of a statistical analysis. What follows is a brief introduction to the principles and vocabulary of measurement.

Operationalization

Emmet (15) wrote, "We must beware always of thinking that because a word exists the 'thing' for which that word is supposed to stand necessarily exists too."

Measurement begins with the assignment of numbers to events or things to help us describe reality. Measurements range from the obvious (eg, diameter, length, time) to the more difficult (eg, patient satisfaction, quality of life, pain), but all are something we can quantify or count. This process is called *operationalization.* Operationalized concepts range from well-established measures such as lesion diameter in millimeters to less well-defined measures such as image quality, contrast agent toxicity, patient comfort, and imaging cost.

If this appears somewhat abstract, consider the following three points: First, researchers can operationalize anything that exists (16), but some measures will be more imprecise (quality of life) than others (diameter). Second, since there is likely to be more than one way to operationalize a concept, the choice of the best way may not be obvious. Third, the radiology profession and the research it generates are saturated with conceptualizations that have been operationalized, some more successfully than others.

Variables

Variables represent measurable indicators of a characteristic that can take on more than one value from one observation to the next. A characteristic may have a different value in different people, in different places, or at different times. Such variables are often referred to as random variables when the value of a particular outcome is determined by chance (ie, by means of random sampling) (17). Since many characteristics are measured imperfectly, we should not expect complete congruence between a measure and truth. Put simply, any measurement has an error component.

If a measure does not take on more than one value, it is referred to as a constant. As an example, patient sex is a variable: It can vary between male and female from one patient to the next. However, a study of breast imaging is likely to be limited to female patients. In this context, sex is no longer a variable in a statistical sense (we cannot analyze it because it does not vary). In contrast, holding the value of one variable constant in order to clarify variations in other variables is sometimes referred to as a statistical control. With mammography as an example, it may be useful to estimate the accuracy of an imaging technique separately for women with and for women without dense breast tissue. As noted previously, however, operationalizing what is and is not dense breast tissue may not be as simple as it first appears.

Measurement Scales

There are four levels of data, commonly referred to as nominal, ordinal, interval, and ratio data. Nominal data classifies objects according to type or characteristic but has no logical order. With imaging technology as an example, ultrasonography, magnetic resonance (MR) imaging, computed tomography, and conventional radiography are each exclusive technologic categories, generally without logical order. Other common examples would be sex, race, and a radiologist's primary subspecialty. Ordinal data also classify objects according to characteristic, but the categories can take on some meaningful order. The American College of Radiology Breast Imaging Reporting and Data System, or BI-RADS, classification system for final assessment is a good example of an ordinal scale. The categories are mutually exclusive (eg, a finding cannot be both "benign" and a "suspicious abnormality"), have some logical order (ranked from "negative" to "highly suggestive of malignancy"), and are scaled according to the amount of a particular characteristic they possess (suspicion of malignancy). Nominal and ordinal data are also referred to as qualitative variables, since their underlying meaning is nonnumeric.

Interval data classify objects according to type and logical order, but the differences between levels of a measure are equal (eg, temperature in degrees Celsius, T scores reported for bone mineral density). Ratio data are the same as interval data but have a true zero starting point. As noted in the examples above, the values of degrees Celsius and T score can take on both positive and negative numbers. Examples of ratio data would be heart rate, percentage vessel stenosis, and respirations per minute. Interval and ratio data are also referred to as quantitative variables, since they have a direct numeric interpretation. In most analyses, it does not matter whether the data are interval or ratio data.

Continuous and Discrete Variables

Variables such as weight and diameter are measured on a continuous scale, meaning they can take on any value within a given interval or set of intervals. As a general rule of thumb, if a subdivision between intervals makes sense, the data are continuous. As an example, a time interval of minutes can be further divided into seconds, milliseconds, and an infinite number of additional fractions. In contrast, discrete variables such as sex, the five-point BI-RADS final assessment scale, race, and number of children in a household have basic units of measurement that cannot be divided (one cannot have 1.5 children).

Reliability and Validity

Measurement accuracy is directly related to reliability and validity. Reliability is the extent to which the repeated use of a measure yields the same values when no change has occurred. Therefore, reliability can be evaluated empirically. Poor reliability negatively affects all studies. As an example, reliability can depend on who performs the measurement and when, where, how, and from whom the data are collected.

Validity is the extent to which a measure is an accurate representation of the concept it is intended to operationalize. Validity cannot be confirmed empirically—it will always be in question. Although there are several different conceptualizations of validity, the following provides a brief overview. Predictive validity refers to the ability of an indicator to correctly predict (or correlate with) an outcome (eg, imaged abnormal lesion and subsequent malignancy). Content validity is the extent to which the indicator reflects the full domain of interest (eg, tumor shrinkage may be indicated by tumor width, height, or both). Construct validity is the degree to which one measure correlates with other measures of the same concept (eg, does a positive MR study for multiple sclerosis correlate with physical examination findings, patient symptoms, or laboratory results?). Face validity evaluates whether the indicator appears to measure the concept. As an example, it is unlikely that an MR study of the lumbar spine will facilitate a diagnosis for lost memory and disorientation.

Association

The connection between variables is often referred to as *association*. Association, also known as *covariation,* is exhibited by measurable changes in one variable that occur concurrently with changes in another variable. A positive association is represented by changes in the same direction (eg, heart rate increases as physical activity increases). Negative association is represented by concurrent changes in opposite directions (hours per week spent exercising and percentage body fat). Spurious associations are associations between two variables that can be better explained by a third variable. As an example, if after taking medication for a common cold for 10 days the symptoms disappear, one could assume that the medication cured the illness. Most of us, however, would probably agree that the change is better explained in terms of the normal time course of a common cold rather than a pharmacologic effect.

Causation

There is a difference between the determination of association and that of causation. Causation cannot be proved with statistics. With this caveat in mind, statistical techniques are best used to explore (not prove) connections between independent and dependent variables. A dependent variable (sometimes called the response variable) is a variable that contains variations for which we seek an explanation. An independent variable is a variable that is thought to affect (cause) changes in the dependent variable. Causation is implied when statistically significant associations are found between an independent and a dependent variable, but causation can never be truly proved. Proof is always an exercise in logical deduction tempered with a degree of uncertainty (18,19), even in experimental designs (such as randomized controlled trials).

Statistical techniques provide evidence that a relationship exists between independent and dependent variables through the use of significance testing and measures of the strength of association. This evidence must be supported by the theoretical basis and logic of the research. The Table presents a condensed list of elements necessary for a claim of causation. The first attempt to provide an epidemiologic method for evaluating causation was performed by A. G. Hill and adapted for the well-known U.S. Surgeon General's report, *Smoking and Health* (1964) (18,19). The elements described in the Table serve to remind us that causation is neither a simple exercise nor a direct product of statistical significance. This is why many believe the optimal research technique to establish causation is to use a randomized controlled experiment.

Required Elements for Causation

Element	Explanation
Association	Do the variables covary empirically? Strong associations are more likely to be causal than are weak associations.
Precedence	Does the independent variable vary before the effect exhibited in the dependent variable?
Nonspuriousness	Can the empirical correlation between two variables be explained away by the influence of a third variable?
Plausibility	Is the expected outcome biologically plausible and consistent with theory, prior knowledge, and results of other studies?

MAINTAINING PERSPECTIVE

Rothman and Greenland (19) wrote, "The tentativeness of our knowledge does not prevent practical applications, but it should keep us skeptical and critical, not only of everyone else's work but of our own as well."

A basic understanding of measurement will enable radiologists to better understand and put into perspective the substantive importance of published research. Maintaining perspective not only requires an understanding that all analytic studies operate under a cloud of imperfect knowledge, but it also requires sufficient insight to recognize that statistical sophistication and significance testing are tools, not ends in themselves. Statistical techniques, however, are useful in providing summary measures of concepts and helping researchers decide, given certain assumptions, what is meaningful in a statistical sense (more about this in future articles). As new techniques are presented in this series, readers should remind themselves that statistical significance is meaningless without clinical significance.

WHAT COMES NEXT

This introduction to measurement will be followed by a series of articles on basic biostatistics. The series will cover topics on descriptive statistics, probability, statistical estimation and hypothesis testing, sample size, and power. There will also be more advanced topics introduced, such as correlation, regression modeling, statistical agreement, measures of risk and accuracy, technology assessment, receiver operating characteristic curves, and bias. Each article will be written by experienced researchers using radiologic examples to present a nontechnical explanation of a statistical topic.

APPENDIX A: KEY TERMS

Below is a list of the common terms and definitions related to measurement.

Abstract concept.—The starting point for measurement, an abstract concept is best understood as a general idea in linguistic form that helps us describe reality.

Association.—An association is a measurable change in one variable that occurs concurrently with changes in another variable. Positive association is represented by change in the same direction. Negative association is represented by concurrent changes in opposite directions.

Constant.—A constant is an attribute of a concept that does not vary.

Construct validity.—Construct validity is the degree to which one measure correlates with other measures of the same abstract concept.

Content validity.—Content validity is the extent to which the indicator reflects the full domain of interest.

Continuous variable.—This type of variable is a measure that can take on any value within a given interval or set of intervals: an infinite number of possible values.

Dependent variable.—The value of the dependent variable depends on variations in another variable.

Discrete variable.—This type of variable is a measure that is represented by a limited number of values.

Face validity.—Face validity evaluates whether the indicator appears to measure the abstract concept.

Independent variable.—The independent variable can be manipulated to affect variations or responses in another variable.

Interval data.—These variables classify objects according to type and logical order but also require that differences between levels of a category are equal.

Nominal data.—These are variables that classify objects according to type or characteristic.

Operationalize.—This is the process of creating a measure of an abstract concept.

Ordinal data.—These are variables that classify objects according to type or kind but also have some logical order.

Predictive validity.—This is the ability of an indicator to correctly predict (or correlate with) an outcome.

Random variable.—This type of variable is a measure where any particular value is based on chance by means of random sampling.

Ratio data.—These variables have a zero starting point and classify objects according to type and logical order but also require that differences between levels of a category be equal.

Reliability.—Reliability is the extent to which the repeated use of a measure yields the same value when no change has occurred.

Spurious association.—This is an association between two variables that can be better explained by or depends greatly on a third variable.

Statistical control.—This refers to holding the value of one variable constant in order to clarify associations among other variables.

Statistical inference.—This is the process whereby one reaches a conclusion about a population on the basis of information obtained from a sample drawn from that population. There are two such methods, statistical estimation and hypothesis testing.

Validity.—Validity is the extent to which a measure accurately represents the abstract concept it is intended to operationalize.

Variable.—A variable is a measure of a concept that can take on more than one value from one observation to the next.

APPENDIX B: WEB RESOURCES

The following is a list of links to general statistics resources available on the Web (accessed May 14, 2001).

www.StatPages.net

www.stats.gla.ac.uk

The following Web links are sources of statistics help (accessed May 14, 2001).

BMJ Statistics at Square One: *www.bmj.com/statsbk/*

The Little Handbook of Statistical Practice: *www.tufts.edu/~gdallal/LHSP.HTM*

Rice Virtual Lab in Statistics: *www.ruf.rice.edu/~lane/rvls.html*

Concepts and Applications of Inferential Statistics: *faculty.vassar.edu/~lowry/webtext.html*

StatSoft Electronic Textbook: *www.statsoft.com/textbook/stathome.html*

Hypertext Intro Stat Textbook: *www.stat.ucla.edu/textbook/*

Introductory Statistics: Concepts, Models, and Applications: *www.psychstat.smsu.edu/sbk00.htm*

Statnotes: An Online Textbook: *www2.chass.ncsu.edu/garson/pa765/statnote.htm*

Research Methods Knowledge Base: *trochim.human.cornell.edu/kb/*

The following is a list Web links to statistical software (accessed May 14, 2001).

Stata Software (links to software providers): *www.stata.com/links/stat_software.html*

EpiInfo, free software downloads available from the Centers for Disease Control and Prevention: *www.cdc.gov/epiinfo/*

APPENDIX C: SUGGESTED GENERAL READINGS

The following is a list of suggested readings.

Pagano M, Gauvreau K. Principles of biostatistics. Belmont, Calif: Duxbury, 1993.

Motulsky H. Intuitive biostatistics. New York, NY: Oxford University Press, 1995.

Rothman KJ, Greenland S, eds. Modern epidemiology. Philadelphia, Pa: Lippincott-Raven, 1998.

Gordis L, ed. Epidemiology. Philadelphia, Pa: Saunders, 1996.

Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA 1993; 270:2093–2095. [This is from an ongoing series through year 2000.]

References

1. Disraeli B. Quoted by: Twain M. An autobiography of Mark Twain. Neider C, ed. New York, NY: Columbia University Press, 1995; chapter 29.
2. Altman DG, Bland JM. Improving doctor's understanding of statistics. J R Stat Soc 1991; 154:223–267.
3. Hillman BJ, Putnam CE. Fostering research by radiologists: recommendations of the 1991 summit meeting. Radiology 1992; 182:315–318.
4. Doubilet PM. Statistical techniques for medical decision making: application to diagnostic radiology. AJR Am J Roentgenol 1988; 150:745–750.

5. Black WC. How to evaluate the radiology literature. AJR Am J Roentgenol 1990; 154:17–22.
6. Moses LE, Lois TA. Statistical consultation in clinical research: a two-way street. In: Bailar JC III, Mosteller F, eds. Medical uses of statistics. 2nd ed. Boston, Mass: NEJM Books, 1992; 349–356.
7. Altman DG. Statistics and ethics in medical research. VIII. Improving the quality of statistics in medical journals. Br Med J 1981; 282:44–47.
8. Moses L. Statistical concepts fundamental to investigations. In: Bailar JC III, Mosteller F, eds. Medical uses of statistics. 2nd ed. Boston, Mass: NEJM Books, 1992; 5–44.
9. Bailar JC III, Mosteller F, eds. Medical uses of statistics. 2nd ed. Boston, Mass: NEJM Books, 1992.
10. Fisher RA. The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain 1926; 33:503–513.
11. Oxman AD, Guyatt GH. The science of reviewing research. Ann N Y Acad Sci 1993; 703:125–133; discussion 133–134.
12. Altman DG. Statistics: necessary and important. Br J Obstet Gynaecol 1986; 93:1–5.
13. Altman DG. Statistics in medical journals: some recent trends. Stat Med 2000; 19:3275–3289.
14. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall, 1991; 108–111.
15. Emmet ER. Handbook of logic. Lanham, Md: Littlefield Adams Quality Paperbacks, 1993.
16. Babbie E. The practice of social research. 6th ed. New York, NY: Wadsworth, 1995.
17. Fisher LD, Belle GV. Biostatistics: a methodology for the health sciences. New York, NY: Wiley, 1993.
18. Rothman KJ, Greenland S, eds. The emergence of modern epidemiology. In: Modern epidemiology. 2nd ed. Philadelphia, Pa: Lippincott-Raven, 1998; 3–6.
19. Rothman KJ, Greenland S, eds. Causation and causal inference. In: Modern epidemiology. 2nd ed. Philadelphia, Pa: Lippincott-Raven, 1998; 7–28.

Seema S. Sonnad, PhD

Index terms:
Data analysis
Statistical analysis

Published online before print
10.1148/radiol.2253012154
Radiology 2002; 225:622–628

[1] From the Department of Surgery, University of Michigan Medical Center, Ann Arbor. Received January 14, 2002; revision requested March 2; revision received May 20; accepted June 14. **Address correspondence to** the author, Department of Surgery, University of Pennsylvania Health System, 4 Silverstein, 3400 Spruce St, Philadelphia, PA 19104-4283 (e-mail: *seema.sonnad@uphs.upenn.edu*).

Describing Data: Statistical and Graphical Methods[1]

An important step in any analysis is to describe the data by using descriptive and graphic methods. The author provides an approach to the most commonly used numeric and graphic methods for describing data. Methods are presented for summarizing data numerically, including presentation of data in tables and calculation of statistics for central tendency, variability, and distribution. Methods are also presented for displaying data graphically, including line graphs, bar graphs, histograms, and frequency polygons. The description and graphing of study data result in better analysis and presentation of data.
© RSNA, 2002

A primary goal of statistics is to collapse data into easily understandable summaries. These summaries may then be used to compare sets of numbers from different sources or to evaluate relationships among sets of numbers. Later articles in this series will discuss methods for comparing data and evaluating relationships. The focus of this article is on methods for summarizing and describing data both numerically and graphically. Options for constructing measures that describe the data are presented first, followed by methods for graphically examining your data. While these techniques are not methodologically difficult, descriptive statistics are central to the process of organizing and summarizing anything that can be presented as numbers. Without an understanding of the key concepts surrounding calculation of descriptive statistics, it is difficult to understand how to use data to make comparisons or draw inferences, topics that will be discussed extensively in future articles in this series.

In this article, five properties of a set of numbers will be discussed. *(a)* Location or central tendency: What is the central or most typical value seen in the data? *(b)* Variability: To what degree are the observations spread or dispersed? *(c)* Distribution: Given the center and the amount of spread, are there specific gaps or concentrations in how the data cluster? Are the data distributed symmetrically or are they skewed? *(d)* Range: How extreme are the largest and smallest values of the observations? *(e)* Outliers: Are there any observations that do not fit into the overall pattern of the data or that change the interpretation of the location or variability of the overall data set?

The following tools are used to assess these properties: *(a)* summary statistics, including means, medians, modes, variances, ranges, quartiles, and tables; and *(b)* plotting of the data with histograms, box plots, and others. Use of these tools is an essential first step to understand the data and make decisions about succeeding analytic steps. More specific definitions of these terms can be found in the Appendix.

DESCRIPTIVE STATISTICS

Frequency Tables

One of the steps in organizing a set of numbers is counting how often each value occurs. An example would be to look at diagnosed prostate cancers and count how often in a 2-year period cancer is diagnosed as stage A, B, C, or D. For example, of 236 diagnosed cancers, 186 might be stage A, 42 stage B, six stage C, and two stage D. Because it is easier to understand these numbers if they are presented as percentages, we say 78.8% (186 of 236) are stage A, 17.8% (42 of 236) are stage B, 2.5% (six of 236) are stage C, and 0.9% (two of 236) are stage D. This type of calculation is performed often, and two definitions are important. The frequency of a value is the number of times that value occurs in a given data set. The relative frequency of a value is the proportion of all observations in the data set with that value. Cumulative frequency is obtained by adding relative frequency for one

TABLE 1
Frequencies, Relative Frequencies, and Cumulative Frequencies of Cancer Staging Distribution

Cancer Stage	Frequency	Cumulative Frequency	Percentage	Relative Frequency (proportion)
A	186	.788	78.8	.79
B	42	.967	17.9	.18
C	6	.992	2.5	.025
D	2	1.0	0.9	.009

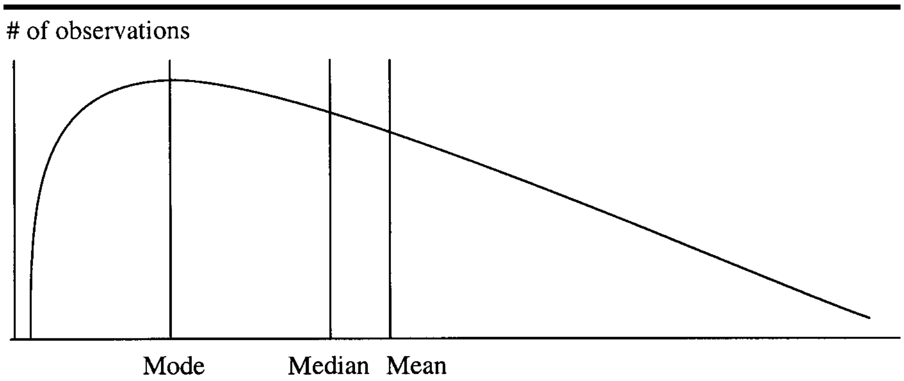


Figure 1. Line graph shows the mean, median, and mode of a skewed distribution. In a distribution that is not symmetric, such as this one, the mean (arithmetic average), the median (point at which half of the data lie above and half lie below), and the mode (most common value in the data) are not the same.

value at a time. The cumulative frequency of the first value would be the same as the relative frequency and that of the first two values would be the sum of their relative frequencies, and so on. Frequency tables appear regularly in *Radiology* articles and other scientific articles. An example of a frequency table, including both frequencies and percentages, is shown in Table 1. In the section of this article about graphic methods, histograms are presented. They are a graphic method that is analogous to frequency tables.

Measurement of the Center of the Data

The three most commonly used measures of the center of the data are the mean, median, and mode. The mean (often referred to as the average) is the most commonly used measure of center. It is most often represented in the literature as $\bar{x}$. The mean is the sum of the values of the observations divided by the number of observations. The median is the midpoint of the observations, when arranged in order. Half of the observations in a data set lie below the median and half lie above the median. The mode is the most frequent value. It is the value that occurs most commonly in the data set. In a data set like the one in Figure 1, the mean, median, and mode will be different numbers. In a perfectly normal distribution, they will all be the same number. A normal distribution is a commonly occurring symmetric distribution, which is defined by the familiar bell-shaped curve, that includes a set percentage of data between the center and each standard deviation (SD) unit.

When a median is used, the data are arranged in order and the middle observation is selected. If the number of observations is even, there will be a middle pair of values, and the median will be the point halfway between those two values. The numbers must be ordered when the median is selected, and each observation must be included. With a data set of 4, 4, 5, 5, 6, 6, 8, and 9, the median is 5.5. However, if values represented by the data were listed rather than listing the value of each observation, the median would be erroneously reported as 6. Finally, if there are 571 observations, there is a method to avoid counting in from the ends. Instead, the following formula is used: If there are n observations, calculate $(n + 1)/2$. Arrange the observations from smallest to largest, and count $(n + 1)/2$ observations up from the bottom. This gives the median. In real life, many statistical programs (including Excel; Microsoft, Redmond, Wash), will give the mean, median, and mode, as well as many other descriptive statistics for data sets.

To look for the mode, it is helpful to create a histogram or bar chart. The most common value is represented by the highest bar and is the mode. Some distributions may be bimodal and have two values that occur with equal frequency. When no value occurs more than once, they could all be considered modes. However, that does not give us any extra information. Therefore, we say that these data do not have a mode. The mode is not used often because it may be far from the center of the data, as in Figure 1, or there may be several modes, or none. The main advantage of the mode is that it is the only measure that makes sense for variables in nominal scales. It does not make sense to speak about the median or mean race or sex of radiologists, but it does make sense to speak about the most frequent (modal) race (white) or sex (male) of radiologists. The median is determined on the basis of order information but not on the basis of the actual values of observations. It does not matter how far above or below the middle a value is, but only that it is above or below.

The mean comprises actual numeric values, which may be why it is used so commonly. A few exceptionally large or small values can significantly affect the mean. For example, if one patient who received contrast material before computed tomography (CT) developed a severe life-threatening reaction and had to be admitted to the intensive care unit, the cost for care of that patient might be several hundred thousand dollars. This would make the mean cost of care associated with contrast material–enhanced CT much higher than the median cost because without such an episode costs might only be several hundred dollars. For this reason, it may be most appropriate to use the median rather than the mean for describing the center of a data set if the data contain some very large or very small outlier values or if the data are not centered (Fig 1).

“Skewness” is another important term. While there is a formula for calculating skew, which refers to the degree to which a distribution is asymmetric, it is not commonly used in data analysis. Data that are skewed right (as seen in Fig 1) are common in biologic studies because many measurements involve variables that have

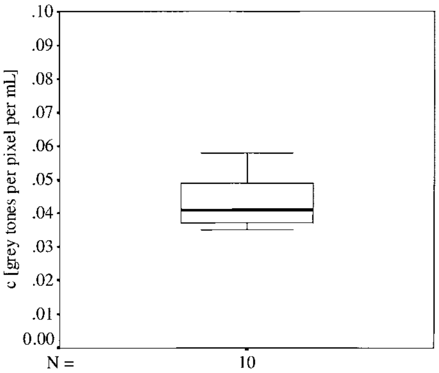


Figure 2. Box plot demonstrates low variation. A high-variation box plot would be much taller. The horizontal line is the median, the ends of the box are the upper and lower quartiles, and the vertical lines are the full range of values in the data. (Reprinted, with permission, from reference 1.)

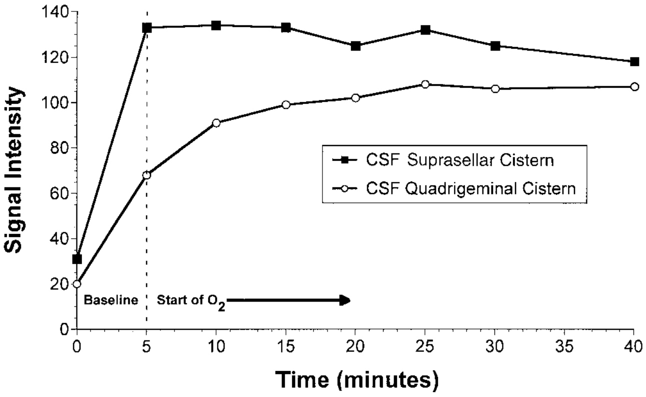


Figure 3. Line graph shows change in MR signal intensity in the cerebrospinal fluid *(CSF)* collected during oxygen inhalation over time. Signal intensity in the quadrigeminal plate cistern increases more gradually, and equilibration is reached at 15–20 minutes after the start of oxygen inhalation. (Reprinted, with permission, from reference 3.)

a natural lower boundary but no definitive upper boundary. For example, hospital length of stay can be no shorter than 1 day or 1 hour (depending on the units used by a given hospital), but it could be as long as several hundred days. The latter would result in a distribution with more values below some cutoff and then a few outliers that create a long "tail" on the right.

Measuring Variability in the Data

Measures of center are an excellent starting point in summarizing data, but they usually do not "tell the full story" and can be misleading if there is no information about the variability or spread of the data. An adequate summary of a set of data requires both a measure of center and a measure of variability. Just as with the center, there are several options for measuring variability. Each measure of variability is most often associated with one of the measures of center. When the median is used to describe the center, the variability and general shape of the data distribution are described by using percentiles. The *x*th percentile is the value at which *x* percent of the data lie below that percentile and the rest lie above it; therefore, the median is also the 50th percentile. As seen in the box plot (Fig 2), the 25th and 75th percentiles, also known as the lower and upper quartiles, respectively, are often used to describe data (1).

Although the percentiles and quartiles used for creating box plots are useful and simple, they are not the most common measures of spread. The most common measure is the SD. The SD (and the related variance) is used to describe spread around the center when the center is expressed as a mean. The formula for variance would be written as

$$\frac{\Sigma\ (\text{obs} - \text{mean})^2}{\text{no. of obs}},$$

where Σ represents a summation of all the values, and obs means observations. Squaring of the differences results in all positive values. Then the SD is

$$\sqrt{\frac{\Sigma\ (\text{obs} - \text{mean})^2}{\text{no. of obs}}},$$

the square root of the variance. It is helpful to think of the SD as the average distance of observations from the mean. Because it is a distance, it is always positive. Large outliers affect the SD drastically, just as they do the mean. Occasionally, the coefficient of variation—the SD or mean multiplied by 100 to get a percentage value—is used. This can be useful if the interest is in the percentage variation rather than the absolute value in numeric terms. Kaus and colleagues (2) presented an interesting table in their study. Table

TABLE 2
Intra- and Interobserver Variability in MR Reading: Automated versus Manual Method

Segmented Volume and Tumor Histologic Type	Manual Method		Automated Method	
	Intraobserver	Interobserver	Intraobserver	Interobserver
Brain				
Meningioma	0.42 ± 0.03	4.93 ± 1.75	0.36 ± 0.45	1.84 ± 0.65
Low-grade glioma	1.79 ± 1.53	6.31 ± 2.85	1.44 ± 1.33	2.71 ± 1.68
Tumor				
Meningioma	1.58 ± 0.98	7.08 ± 2.18	0.66 ± 0.72	2.66 ± 0.38
Low-grade glioma	2.08 ± 0.78	13.61 ± 2.21	2.06 ± 1.73	2.97 ± 1.58

Note.—Data are the mean coefficient of variation percentage plus or minus the SD. (Adapted and reprinted, with permission, from reference 2.)

TABLE 3
Standard Scores and Corresponding Percentiles

Standard Score	Percentile
−3.0	0.13
−2.5	0.62
−2.0	2.27
−1.5	6.68
−1.0	15.87
−0.5	30.85
0.0	50.00
0.5	69.15
1.0	84.13
1.5	93.32
2.0	97.73
2.5	99.38
3.0	99.87

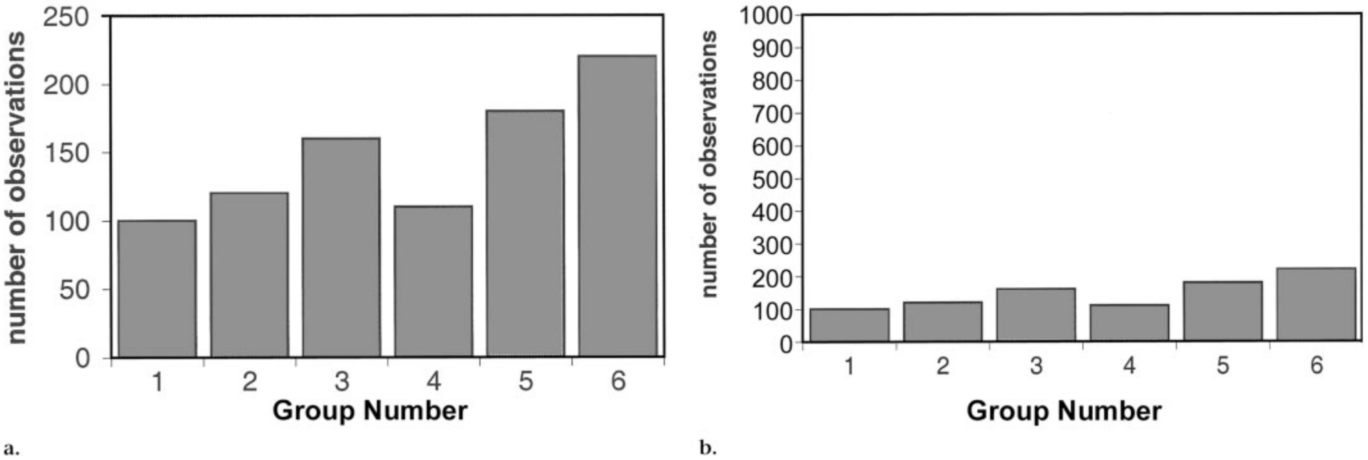


Figure 4. Bar graphs. (**a**) Use of a scale with a maximum that is only slightly higher than the highest value in the data shows the differences between the groups more clearly than does (**b**) use of a scale with a maximum that is much higher than the highest value.

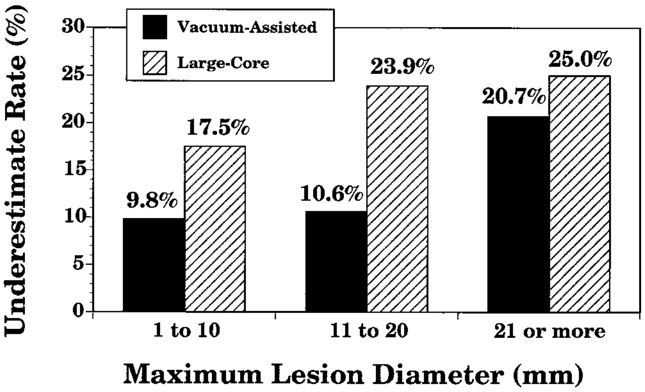


Figure 5. Bar graph shows the underestimation rate for lesion size with vacuum-assisted and large-core needle biopsy. Underestimation rates were lower with the vacuum-assisted device. It is helpful to label the bars with the value. (Reprinted, with permission, from reference 4.)

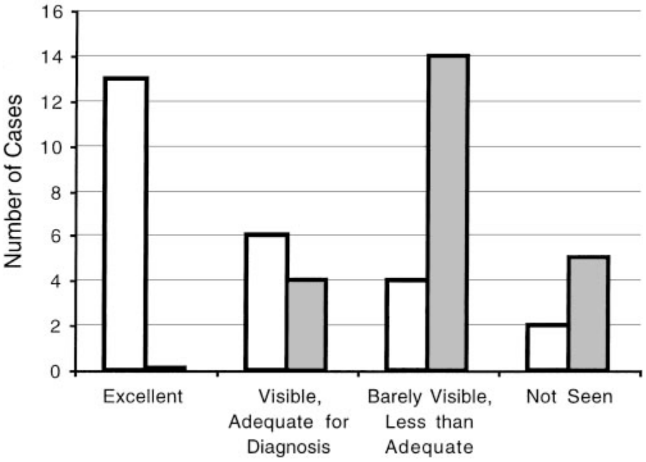


Figure 6. Bar graph shows the diagnostic adequacy of high-spatial-resolution MR angiography with a small field of view and that with a large field of view for depiction of the segmental renal arteries. (Reprinted, with permission, from reference 5.)

2 shows their data on the intra- and interobserver variability in the reading of magnetic resonance (MR) images with automated and manual methods.

Normal Distribution and Standard Scores

Once a mean and SD are calculated, they can be used to further simplify description of data or to create statistics that can be compared across data sets. One common measure is the standard score. With the standard score, data are assumed to come from a normal distribution, a symmetric bell-shaped distribution. In a normal distribution, 68% of all observations are within 1 SD of the mean (34% above the mean and 34% below). Another 27% are between 1 and 2 SDs; therefore, 95% of all observations are within 2 SDs of the mean. A total of 99.7% are within 3 SDs of the mean; therefore, any normal curve (or histogram that represents a large set of data drawn from a normal distribution) is about 6 SDs wide. Two SDs from the mean is often used as a cutoff for the assignment of values as outliers. This convention is related to the common use of 95% confidence intervals and the selection of confidence for testing of a hypothesis as 95% (these concepts will be defined in a future article). The use of 2 SDs from the mean in normally distributed data ensures that 95% of the data are included. As can be seen, the SD is the common measure of variability for data from normal distributions. These data can be expressed as standard scores, which are a measure of SD units from the mean. The standard score is calculated as $(\mathrm{OV} - M)/\mathrm{SD}$, where OV is the observation value and M is the mean. A standard score of 1 corresponds to the 84th percentile of data from a normal distribution. Standard scores are useful because if the data are drawn from a normal distribution, regardless of the original mean and SD, each standard score corresponds to a specific percentile. Table 3 shows percentiles that correspond to some standard scores.

GRAPHICAL METHODS FOR DATA SUMMARY

Line and Bar Graphs

It is often easier to understand and interpret data when they are presented graphically rather than descriptively or as a table. Line graphs are most often used to show the behavior of one variable over time. Time appears on the horizontal axis and the variable of interest on the vertical axis. Figure 3 is an example of a line graph. The variable is MR signal intensity in the cerebrospinal fluid collected during oxygen inhalation. It is apparent from this graph that signal intensity within the quadrigeminal plate cistern increases more gradually than that within the suprasellar cistern, which was the conclusion drawn by the authors (3).

When you look at graphs, it is important to examine both the horizontal and vertical axis scales. Selection of the scale to be used may influence how the reader interprets the graph. This can be seen in Figure 4, which depicts bar graphs.

Bar graphs are another common way to display data. They are used for comparing the value of multiple variables. In many cases, multiple bar graphs will appear in the same figure, such as in Figures 5 and 6. Figure 5 shows the diagnostic underestimation rate for three categories of lesion size with vacuum-assisted and large-core needle biopsies. This bar chart is presented with a percentage rate on the y axis and the categories on the x axis (4). Figure 6 is similar and shows the diagnostic adequacy of high-spatial-resolution MR angiography with a small field of view compared with that with a large field of view for depiction of the segmental renal arteries. In this case, the y axis represents the number of cases and the percentages appeared in the figure caption (5). Bars may be drawn either vertically, as in Figures 5 and 6, or horizontally. They may be drawn to be touching each other or to be separate. It is important that the width of the bars remains consistent so that one bar does not appear to represent more occurrences because it has a larger area.

TABLE 4
Data for Body Weight of 10 Patients Used to Construct Stem and Leaf Plot

Volunteer No./ Age (y)/Sex	Body Weight (kg)	Height (cm)	ΔQ_F (gray tones per pixel)*	*c* (gray tones per pixel per milliliter)
1/27/F	76	178	4.7	0.049
2/32/F	61	173	3.3	0.035
3/31/M	83	181	3.8	0.040
4/28/M	85	180	4.0	0.042
5/44/M	103	189	4.6	0.048
6/32/M	72	179	3.4	0.036
7/23/M	78	174	5.5	0.058
8/26/F	72	179	4.8	0.050
9/28/M	80	180	3.5	0.037
10/23/F	74	177	3.6	0.038

Note.—The mean values ± SEM were as follows: age, 29 years ± 2; body weight, 78 kg ± 3; height, 179 cm ± 1; ΔQ_F, 4.1 gray tones per pixel ± 0.2; *c* ($\Delta Q_F/V$), 0.043 gray tones per pixel per milliliter ± 0.003. (Adapted and reprinted, with permission, from reference 1).
* Ninety-five milliliters of saline solution was instilled.

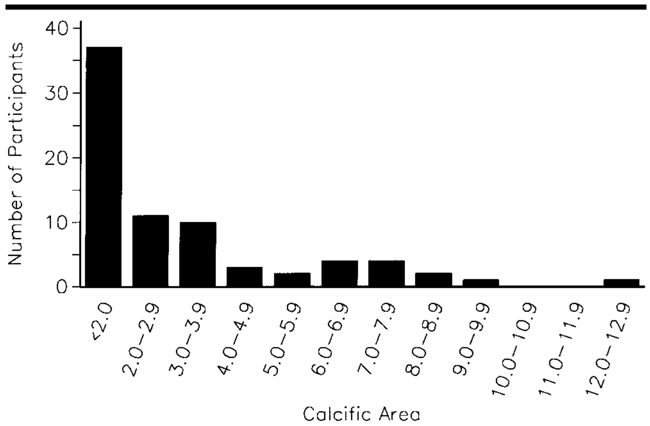


Figure 7. Histogram represents distribution of calcific area. Labels on the x axis indicate the calcific area in square millimeters for images with positive findings. (Reprinted, with permission, from reference 6.)

Histograms and Frequency Polygons

Histograms look somewhat like bar charts, but they serve a different purpose. Rather than displaying occurrence in some number of categories, a histogram is intended to represent a sampling distribution. A sampling distribution is a representation of how often a variable would have each of a given set of values if many samples were to be drawn from the total population of that variable. The bars in a histogram appear in a graph where the y axis is frequency and the x axis is marked in equal units. Remember that a bar chart does not have x-axis units. When a program automatically creates a histogram, it creates x-axis units of equal size. The histogram is analogous to the frequency table discussed earlier. Figure 7 shows a histogram, although it was called a bar chart in the original figure caption (6). This figure represents the distribution of calcific area among participants and has equal units on the x axis, which makes it a histogram.

Frequency polygons are a less used alternative to histograms. Figure 8a shows a histogram that might represent the distribution of the mean partition coefficients in a number of healthy individuals. The line in the figure is created by connecting the middle of each of the histogram bars. The figure represented by that line is called a frequency polygon. The frequency polygon is shown in Figure 8b. In Figure 8a, both the histogram and the frequency polygon are shown on the same graph. Most often, one or the other appears but not both.

Stem and Leaf Plots

Another graphic method for representing distribution is known as the stem and leaf plot, which is useful for relatively small data sets, as seen in many radiologic investigations. With this approach, more of the information from the original data is preserved than is preserved with the histogram or frequency polygon, but a graphic summary is still provided. To make a stem plot, the first digits of the data values represent the stems. Stems appear vertically with a vertical line to their right, and the digits are sorted into ascending order. Then the second digit of each value occurs as a leaf to the right of the proper stem. These leaves should also be sorted into ascend-

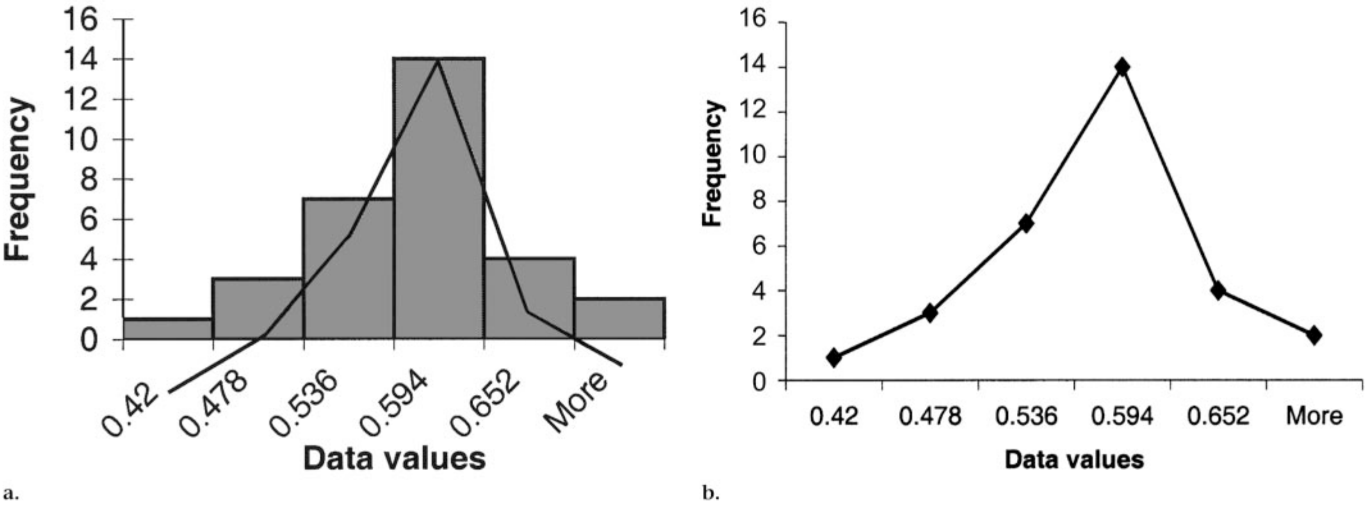


Figure 8. Representative histogram and frequency polygon constructed from hypothetical data. Alternate ways of showing the distribution of a set of data are shown **(a)** with both the histogram and the frequency polygon depicted or **(b)** with only the frequency polygon depicted.

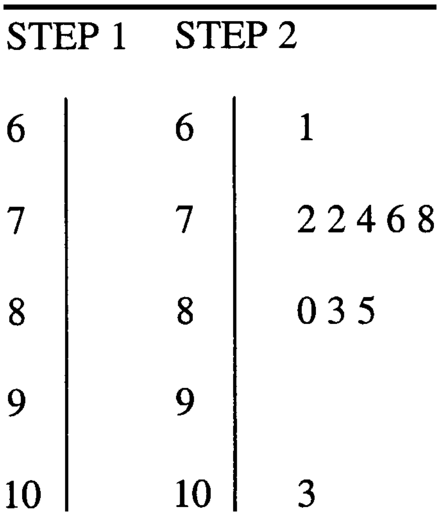


Figure 9. Stem and leaf plot constructed from data in Table 4. Step one shows construction of the stem, and step 2 shows construction of the leaves. These steps result in a sideways histogram display of the data distribution, which preserves the values of the data used to construct it. The number *9* appears on the stem as a place saver; the lack of digits to the right of this stem number indicates that no values began with this number in the original data.

ing order. In a simple example, Heverhagen and colleagues (1) used the body weight in kilograms of 10 patients. The original data appear in Table 4. Figure 9 shows the making of a stem plot from these data. The stem and leaf plot looks like a histogram with horizontal bars made of numbers. The plot's primary advantage is that all of the actual values of the observations are retained. Stem and leaf plots do not work well with very large data sets because there are too many leaves, which makes it difficult both to read and to fit on a page.

Box Plots

The final graphic method presented in this article is the box plot. The box plot shows the distribution of the data and is especially useful for comparing distributions graphically. It is created from a set of five numbers: the median, the 25th percentile or lower quartile, the 75th percentile or upper quartile, the minimum data value, and the maximum data value. The horizontal line in the middle of the box is the median of the measured values, the upper and lower sides of the box are the upper and lower quartiles, and the bars at the end of the vertical lines are the data minimum and maximum values.

Tombach and colleagues (7) used box plots in their study of renal tolerance of a gadolinium chelate to show changes over time in the distribution of values of serum creatinine concentration and creatinine clearance across different groups of patients. Some of these box plots are shown in Figure 10. They clearly demonstrate the changing distribution and the difference in change between patient groups.

In conclusion, a statistical analysis typically requires statistics in addition to a measure of location and a measure of variability. However, the plotting of data to see their general distribution and the computing of measures of location and spread are the first steps in being able to determine the interesting relationships that exist among data sets. In this article, methods have been provided for calculating the data center with the mean, median, or mode and for calculating data spread. In addition, several graphic methods were explained that are useful both when exploring and when presenting data. By starting with describing and graphing of study data, better analysis and clear presentation of data will result; therefore, descriptive and graphic methods will improve communication of important research findings.

APPENDIX

The following is a list of the common terms and definitions related to statistical and graphic methods of describing data.

Coefficient of variation.—SD divided by the mean and then multiplied by 100%.

Descriptive statistics.—Statistics used to summarize a body of data. Contrasted with inferential statistics.

Frequency distribution.—A table that shows a body of data grouped according to numeric values.

Frequency polygon.—A graphic method of presenting a frequency distribution.

Histogram.—A bar graph that represents a frequency distribution.

Inferential statistics.—Use of sample statistics to infer characteristics about the population.

Mean.—The arithmetic average for a group of data.

Median.—The middle item in a group of data when the data are ranked in order of magnitude.

Mode.—The most common value in any distribution.

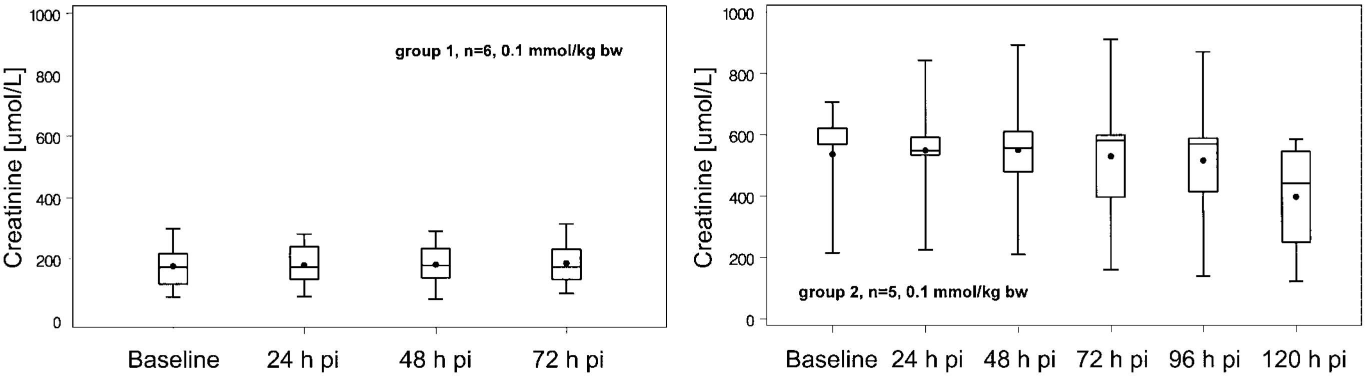


Figure 10. Box plots present follow-up data for serum creatinine concentration. Left: Within 72 hours after injection of a gadolinium chelate (group 1 baseline creatinine clearance, <80 mL/min [<0.50 mL/sec]). Right: Within 120 hours after injection (group 2 baseline creatinine clearance, <30 mL/min [<0.50 mL/sec]). (Reprinted, with permission, from reference 7.)

Nominal data.—Data with items that can only be classified into groups. The groups cannot be ranked.

Normal distribution.—A bell-shaped curve that describes the distribution of many phenomena. A symmetric curve with the highest value in the center and with set amounts of data on each side with the mathematical property that the logarithm of its probability density is a quadratic function of the standardized error.

Percentage distribution.—A frequency distribution that contains a column listing the percentage of items in each class.

Quartile.—Value below which 25% (lower quartile) or 75% (upper quartile) of data lie.

Sample.—A subset of the population that is usually selected randomly. Measures that summarize a sample are called sample statistics.

Sampling distribution.—The distribution actually seen (often represented with a histogram) when data are drawn from an underlying population.

Standard deviation.—A measure of dispersion, the square root of the average squared deviation from the mean.

Variance.—The average squared deviation from the mean, or the square of the SD.

References

1. Heverhagen JT, Muller D, Battmann A, et al. MR hydrometry to assess exocrine function of the pancreas: initial results of noninvasive quantification of secretion. Radiology 2001; 218:61–67.
2. Kaus MR, Warfield SK, Nabavi A, Black PM, Jolesz FA, Kikinis R. Automated segmentation of MR images of brain tumors. Radiology 2001; 218:586–591.
3. Deliganis AV, Fisher DJ, Lam AM, Maravilla KR. Cerebrospinal fluid signal intensity increase on FLAIR MR images in patients under general anesthesia: the role of supplemental O(2). Radiology 2001; 218: 152–156.
4. Jackman RJ, Burbank F, Parker SH, et al. Stereotatic breast biopsy of nonpalpable lesions: determinants of ductal carcinoma in situ underestimation rates. Radiology 2001; 218:497–502.
5. Fain SB, King BF, Breen JF, Kruger DG, Riederer SJ. High-spatial-resolution contrast-enhanced MR angiography of the renal arteries: a prospective comparison with digital subtraction angiography. Radiology 2001; 218:481–490.
6. Bielak LF, Sheedy PF II, Peyser PA. Automated segmentation of MR images of brain tumors. Radiology 2001; 218:224–229.
7. Tombach B, Bremer C, Reimer P, et al. Renal tolerance of a neutral gadolinium chelate (gadobutrol) in patients with chronic renal failure: results of a randomized study. Radiology 2001; 218:651–657.

Probability in Radiology[1]

Elkan F. Halpern, PhD
G. Scott Gazelle, MD, MPH, PhD

Index terms:
Statistical analysis

Published online before print
10.1148/radiol.2261011712
Radiology 2003; 226:12–15

Abbreviation:
PPV = positive predictive value

In this article, a summary of the basic rules of probability using examples of their application in radiology is presented. Those rules describe how probabilities may be combined to obtain the chance of "success" with either of two diagnostic or therapeutic procedures or with both. They define independence and relate it to the conditional probability. They describe the relationship (Bayes rule) between sensitivity, specificity, and prevalence on the one hand and the positive and negative predictive values on the other. Finally, the two distributions most commonly encountered in statistical models of radiologic data are presented: the binomial and normal distributions.

Radiologists routinely encounter probability in many forms. For instance, the sensitivity of a diagnostic test is really just a probability. It is the chance that disease (eg, a liver tumor) will be detected in a patient who actually has the disease. In the process of determining whether a sequence of successive diagnostic tests (say, both computed tomography [CT] and positron emission tomography) is a significant improvement over CT alone, radiologists must understand how those probabilities are combined to give the sensitivity of the combination.

Similarly, the prevalence of disease such as malignant liver cancer among patients with cirrhosis is a probability. It is the fraction of patients with a history of cirrhosis who have a malignant tumor. It can be determined simply from the number of patients with cirrhosis and the number of patients with both cirrhosis and malignant tumors of the liver or from the two prevalences.

The likelihood that a patient with a cirrhotic liver and positive CT findings has a malignant tumor (the positive predictive value [PPV] of CT in this setting) is another probability. It is determined by combining the prevalence of the disease among patients with cirrhosis with the sensitivity and specificity of CT.

In all its forms, probabilities obey a single set of rules that determine how they may be combined; this article presents an outline of these rules. The rules are also found and explained in greater detail in textbooks of biostatistics, such as that by Rosner (1).

THE EARLIEST DEFINITION OF PROBABILITY

Probability is a numeric expression of the concept of chance, in all its guises. As such, it obeys the rules of arithmetic and the logic of mathematics.

In its earliest manifestation, probability was a tool for gamblers. The earliest expressions of probability dealt with simple gambling games such as flipping a coin, rolling a die, or dealing a single card from the top of a deck. The essential property assumed of such games was that the possible outcomes were all equally likely. The probability of some event was proportional to the number of individual outcomes that comprised the event.

In more complicated scenarios, the fundamental outcomes might not be equally likely or there might be an infinite number of them. In these situations, the definition of a probability of an event became the fraction of times that the event would occur by chance. That is, it is the fraction of times in a sufficiently long sequence of trials where the chance of the event was the same in all trials and was not affected by the results of previous or subsequent trials.

With either definition, counting possible outcomes or measuring the frequency of occurrence, probability was susceptible to the laws and rules of simple arithmetic. The simplest rule of all was the rule of addition. Addition of the fraction of patients who have a single liver tumor that is malignant to the fraction of patients who have a single tumor that is benign must yield the fraction of patients who have a single tumor, either malignant or benign.

[1] From the DATA Group, Department of Radiology, Massachusetts General Hospital, Zero Emerson Pl, Suite 2H, Boston, MA 02114. Received October 19, 2001; revision requested December 26; revision received March 27, 2002; accepted April 9. **Address correspondence to** E.F.H. (e-mail: *elk@the-data-group.org*).

Algebraically, we express this "additive" rule for two events, A and B, as the following: If $\mathit{Prob}(A\ \textbf{AND}\ B) = 0$, then $\mathit{Prob}(A\ \textbf{OR}\ B) = \mathit{Prob}(A) + \mathit{Prob}(B)$, where we use $\mathit{Prob}(A)$ to denote the probability of A. The condition "$\mathit{Prob}(A\ \textbf{AND}\ B) = 0$" is a way of saying that the simultaneous occurrence of both A and B is impossible.

One consequence of this rule is that the chance that an event would not happen is immediately determined by the chance that it would happen. As "A" and "**NOT** A" (the event that A would not happen) cannot occur simultaneously, yet one or the other is certain, $\mathit{Prob}(A) + \mathit{Prob}(\textbf{NOT}\ A) = 1$ or $\mathit{Prob}(\textbf{NOT}\ A) = 1 - \mathit{Prob}(A)$. Thus, the fraction of patients who do not have a single tumor is just the complement to the fraction of patients who do.

The limitation "$\mathit{Prob}(A\ \textbf{AND}\ B) = 0$" is necessary, as the rule does not apply without modification when it is possible for both events to be true. For instance, addition of the fraction of patients who have at least one malignant tumor to the fraction of patients who have at least one benign tumor may cause overestimation of the fraction of patients with tumors, either malignant or benign. The extent of the overestimation is given by the fraction of patients who have both benign and malignant tumors. These patients were included in the counts for both tumor-type specific fractions but should be counted only once when patients who have at least one tumor are counted. This may be best seen in Figure 1, where the added shaded areas of A and B yield a sum greater than the total shaded area. The area representing the overlap of A and B must be accounted for.

The required modification to the additive rule of probabilities is: $\mathit{Prob}(A\ \textbf{OR}\ B \text{ or } [A\ \textbf{AND}\ B]) = \mathit{Prob}(A) + \mathit{Prob}(B) - \mathit{Prob}(A\ \textbf{AND}\ B)$.

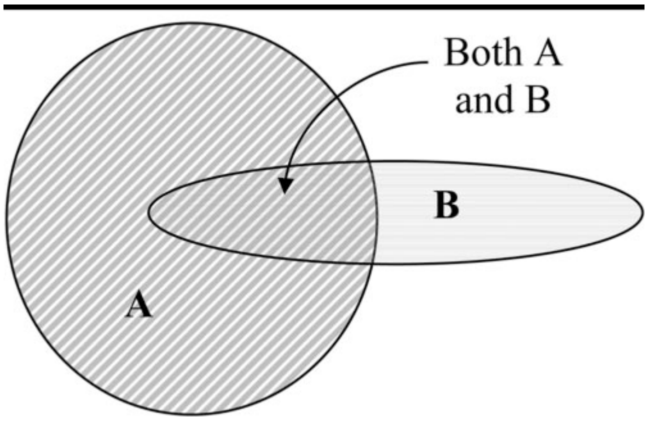


Figure 1. Venn diagram represents the probability of either of two events based on the probabilities of each and of both. The area of the shape made by combining A and B is the total of the areas of A and B less the area where they overlap.

SUBJECTIVE VERSUS OBJECTIVE PROBABILITIES

Before a coin is flipped, the probability that it will land heads up is 0.5. After it is flipped and is seen to have landed heads up, the probability that it landed heads up is 1 and that it landed tails up is 0. But what is the probability that it landed heads up before anyone saw which face was up? The face has been determined, and the probability is either 1 or 0 that it is heads, though which is not yet known. Yet, any gambler would be willing to bet at that point on which face had landed up in precisely the same way that he or she would have bet before the coin had been flipped. The gambler has his or her own "subjective" probability that the coin will be face up when observed, which reflects his or her beliefs concerning the probability before the coin was flipped. The subjective probabilities obey the same laws and rules as the "objective" probabilities in describing future events. Subjective probabilities may be revised as one learns more about what else may be true.

In radiology, before an examination, the probability that a patient will have an abnormality detected is a fraction between 0 and 1. After the procedure and after the image has been viewed, the probability is either 1 or 0, depending on whether an abnormality has been detected. In the interim, after the procedure but before the image has been viewed, the subjective probability is the same as the objective probability was before the procedure.

CONDITIONAL PROBABILITIES AND INDEPENDENCE VERSUS DEPENDENCE

While the probability that a randomly selected woman has an undetected malignant breast cancer at least 1 cm in diameter has real meaning, the probability is not the same for all women. It certainly is higher for women who have never undergone mammographic screening than for women who have—all other things being equal. The probability that a woman who has never been screened has such a tumor is called a "conditional" probability, because it is defined as the chance that the woman has a 1 cm or greater tumor, given that the woman has never been screened. The probability that a randomly chosen woman has such a tumor is the number of women with such tumors divided by the number of all women. The conditional probability that a randomly chosen woman who has never been screened has such a tumor is defined analogously. It is the number of women who have never been screened who have such tumors out of the number of women who have never been screened. By using $\mathit{Prob}(A|B)$ to denote the conditional probability of A (a woman with a 1 cm or greater breast tumor) given B (she underwent no prior screening), we have $\mathit{Prob}(A|B) = \mathit{Prob}(A\ \textbf{AND}\ B)/\mathit{Prob}(B)$.

This forms the basis of the definition of sensitivity and specificity. If we let A stand for a positive diagnostic examination result and B to represent the actual presence of the disease, then $\mathit{Prob}(A|B)$ is the sensitivity, the chance of a positive examination result among individuals with disease or of a true-positive result. The definition can also be used to calculate the chances derived from a succession of diagnostic tests. For instance, if confirmation of any positive test result, D_1+, is required by means of a second positive test result, D_2+, then the chance that we will obtain two positive test results is given by the "multiplicative" law of probabilities: $\mathit{Prob}(D_1+\ \textbf{AND}\ D_2+) = \mathit{Prob}(D_2+|D_1+) \times \mathit{Prob}(D_1+)$.

That is, the chance that both tests have positive results is a fraction of all second tests with positive results, once the first test had a positive result times the chance that the first test had a positive result. The rule can be similarly used to calculate

Example of Calculation of Positive and Negative Predictive Values Based on Expected Number of Cases Derived from Prevalence, Sensitivity, and Specificity

Disease	Positive Finding *B*	Negative Finding **NOT** *B*	Prevalence*
Present *A*	TP, 950	FN, 50	1,000 (10)
Absent **NOT** *A*	FP, 450	TN, 8,550	9,000 (90)
Total	1,400	8,600	10,000

Note.—Prevalence of disease is 10% out of every 10,000 screening examinations. *A* and *B* are events. FN = false-negative, FP = false-positive, TN = true-negative, TP = true-positive.

* Data in parentheses are percentages.

the chance of two successive negative findings or any other combination.

Occasionally, the chances for the second examination, D_2, are unaltered by the results of the initial examination, D_1. For instance, after the diagnostic image has been obtained, the interpretation by a blinded reader, D_2, should be unaffected by the prior interpretation by another blinded reader, D_1, of the same image. In these situations, the two interpretations are said to be independent (2).

If *A* and *B* are independent, then *Prob(A)* = *Prob*(*A*|*B*) = *Prob*(*A*|**NOT** *B*); that is, the (conditional) chance of *A* given that *B* occurs is the same as the (conditional) chance of *A* given that *B* does not occur and, as a result, also equals the (unconditional) chance of *A*.

A consequence is the multiplicative law of probabilities for independent events; namely, if *A* and *B* are independent, then *Prob*(*A* **AND** *B*) = *Prob(A)* × *Prob(B)*.

In practice, the diagnostic tests of radiology are rarely truly independent as CT, magnetic resonance imaging, and ultrasonography all rely similarly on the lesion size and differences in tissue characteristics. Yet, quite often, the multiplicative law is used as an approximation because the exact conditional probability has never been accurately determined in clinical trials of both diagnostic modalities.

The same rules may be used to calculate the risk of disease in the presence of multiple predictive characteristics (or cofactors). If the effects of the cofactors are synergistic, the more general multiplicative rule must be used. But if the effects of the cofactors are unrelated or independent, the multiplicative rule for independent events may be used. Similarly, the rules may be used to compute the overall chance of a successful treatment of disease with a succession of treatments based on the (conditional) chance of success of each treatment.

BAYES RULE AND POSITIVE AND NEGATIVE PREDICTIVE VALUES

While the sensitivity and specificity of a diagnostic test are important to the clinician when he or she determines which test to use, they do not entirely address the question of concern after the test has been performed. To the patient, the issue is not “How often does the test detect real disease?” but rather, “Now that the test results are known, what is the chance that I have the disease?” The patient wants to know a conditional probability that is the reverse of sensitivity. If we use D_x to denote a positive finding and *D* to denote the actual presence of disease, the patient is not as concerned with the sensitivity, $Prob(D_x|D)$, as with the PPV, $Prob(D|D_x)$, the chance that there is disease present given that the test result was positive.

Both sensitivity and specificity are considered to be inherent invariant test characteristics. In contrast, the PPV and the negative predictive value depend not only on the sensitivity and specificity but also on the prevalence of the disease, *Prob*(*D*). They may be combined by using Bayes rule (3), which relates the PPV, $Prob(D|D_x)$, to the sensitivity, $Prob(D_x|D)$, the specificity, $1 - Prob(D_x|\textbf{NOT}\ D)$, and the prevalence, *Prob(D)*.

$$Prob(D|D_x) = \frac{Prob(D_x|D) \times Prob(D)}{Prob(D_x|D) \times Prob(D) + Prob(D_x|\textbf{NOT}\ D) \times Prob(\textbf{NOT}\ D)}.$$

In this equation, the denominator is the total number of expected positive findings in the population, while the numerator is the number of positive findings that accompany the actual disease.

Suppose that we had a diagnostic test used for screening that had both 95% sensitivity (positive in 95% of all cases where the disease is present) and 95% specificity (negative in 95% of all cases where the disease is absent). When the prevalence of the disease is 10%, out of every 10,000 screening examinations, we can expect to see 1,400 cases that result in a positive finding with the diagnostic test (Table). Rather than go through the laborious exercise of constructing tables, Bayes rule gives us the PPV directly. For 10% prevalence, *Prob*(*D*) = 0.10, we have PPV = (0.95 × 0.10)/(0.95 × 0.10) + (0.05 × 0.90) = 0.095/(0.095 + 0.045) = 0.679.

P VALUES, POWER, AND BAYESIAN STATISTICS

Bayes rule applies for any two events, *A* and *B*, not just positive findings, D_X, and presence of disease, *D*. The distinction between the *Prob*(*A*|*B*) and the *Prob*(*B*|*A*) also forms the basis of the difference between conventional and Bayesian statistical analysis of a clinical trial. In conventional statistical analysis of the results of a study, *B* represents the null hypothesis. After the study has been performed and the results *A* have been observed, the conventional decision regarding the truth of *B* is based on the likelihood of *A**, any result as extreme or even more extreme than *A*, given that *B* is true, *Prob*(*A**|*B*). This probability is known as the *P* value. The less likely that any result as extreme as A, given B, the stronger the evidence that *B* is not true. Conventionally, some cutoff (known as the level of significance) is set in advance, and the study is deemed to have significant findings if the *P* value is smaller than the cutoff.

In a conventional analysis, the probability of significant study results, *S*, conditional on *B* being false, *Prob*(*S*|**NOT** *B*), is known as the power. It is not a factor in the conclusion drawn from the study. Rather, it is the major factor in the design before the study is conducted. It determines the number of patients in the study. The study sample size is chosen to provide the desired chance of successfully showing that *B* is not true.

Bayesian statistics differs from conventional statistics insofar as it depends on the probability that the hypothesis holds given the observed results of the study or studies. This probability is calculated by means of the Bayes rule. In order to be able to calculate it, the (subjective) probability reflecting the prior (before the study) chance or belief that the hypothesis was true is required. Much of the dispute regarding the use of Bayesian analysis centers around the possibility of conclusions that might be largely dictated by “opinion” before hard data are obtained.

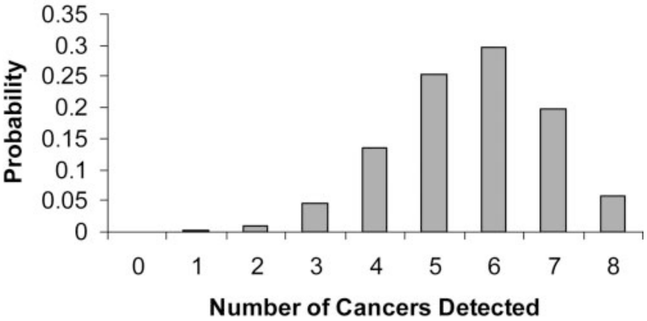


Figure 2. Binomial distribution, $B(i|n,p)$, for the number of cancers detected out of a total of eight, if the sensitivity of the detector is $P = .70$.

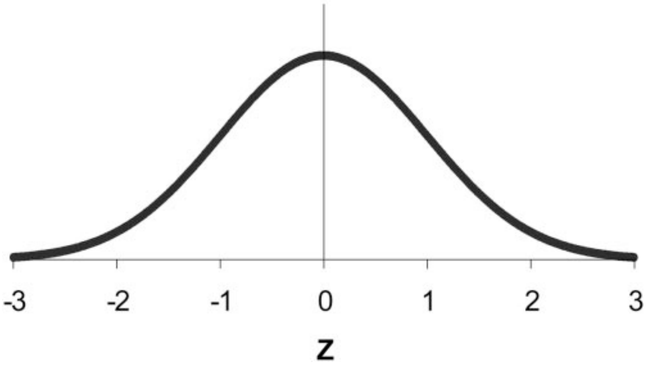


Figure 3. The normal, or Gaussian, distribution.

P values, power, and Bayesian analysis will be presented in later articles in this series. Software for Bayesian probabilities may be found at the University of Sheffield Web site at *www.shef.ac.uk/*~st1ao*/1b.html*.

PROBABILITY FOR CONTINUOUS OUTCOMES

The interpretation of the probability that a tumor consists of a specified number of cells differs in one essential regard from the interpretation of the probability that the tumor is a specified diameter or volume. The number of cells is "discrete" in the sense that it can only be an integer value. No fractional number of cells is possible. As such, each possible number of cells has its own probability, and if that number of cells is possible, the probability is greater than 0. But the diameter or volume of a tumor can take on all fractional values, as well as an integer. The probability that the tumor is between 2 and 3 cm in diameter could be treated in the same way as all of the probabilities that we have been discussing. However, the probability that the tumor is exactly π cm to the last decimal place (or any other exact value, even, say, 3 cm to the last decimal place) has to differ in meaning.

For continuous measures such as the diameter or volume of a tumor or time of an occurrence, the probability that it exactly equals a single value is analogous to the distance traveled or the radiation received in an instant. Instead, one can express the rate at each value and calculate the probability of any interval just as one calculates the distance traveled over any time interval from the instantaneous speed or the total radiation exposure from the instantaneous rate of exposure.

Happily, all of the rules discussed for discrete outcomes apply equally well to continuous ones, both for intervals and for the rates at specified values.

DISTRIBUTIONS

These rules may be used to provide formulas for calculating the distribution, the probabilities for all possible outcomes, under a variety of circumstances. Two of the distributions most commonly encountered by radiologists are the binomial and the normal distributions.

The binomial distribution, $B(i|n,p)$, (4) describes the probability of an event occurring i times out of n tries, where the chance, p, of the event is the same for all tries and the occurrence of the event in one try is unrelated (independent) to its occurrence in any other try. The distribution then gives the probabilities of each possible number of occurrences of the event out of n cases. The specific formula for the probability of i events out of n cases, $B(i|n,p)$, is $B(i|n,p) = [n!/i!(n - i)!]p^i(1 - p)^{n - \mathrm{i}}$, where ! indicates factorial, as in $n! = n \times (n - 1) \times (n - 2) \times \ldots \times 2 \times 1$.

In practice, this distribution is built into most statistical and spreadsheet software packages. For instance, by using Microsoft Excel software, $B(i|n,p)$ is calculated by the function BINOMDIST.

A binomial distribution that a radiologist might encounter is the probability of detecting i cancers with screening. In order for the binomial distribution to apply, the sensitivity would have to be identical for all of the cancers. Additionally, the detection (or nondetection) of any one cancer could not influence the chance that another was detected. If both conditions applied and there were n actual cases of cancer among those screened, $B(i|n,p)$ would be the chance that i cancers were detected if the sensitivity of the technique was given by p. Thus, if there were actually eight cancers and the sensitivity of screening was 70%, the chance that exactly six of those eight cancer were detected is $B(6|8,0.7) = [8!/6!(8 - 6)!]0.7^6(1 - 0.7)^{8 - 6} =$ 0.296. The full distribution is depicted in Figure 2.

The normal, or Gaussian (2,5,6), distribution describes the probabilities for a continuous outcome that is the result of averaging out a very large number of independent random contributions. The background component to number of x rays detected in a square millimeter of plain film is normally distributed. It is commonly described as a "bell-shaped" curve.

The distribution depends on two values or parameters: the mean, μ, and the SD, σ. (See the preceding article [7] in this series.) The mean determines the location of the high point of the curve. The SD gives the scale. The height of the curve at any point, x, is determined by the "z score," the difference of x and μ in units of σ. That is, the height depends only on $z = (x - \mu)/\sigma$.

Again, the height is found as a function in most spreadsheets and statistical software. With Excel software, it is given by NORMDIST. The distribution is depicted in Figure 3.

References

1. Rosner B. Fundamentals of biostatistics. 5th ed. Pacific Grove, Calif: Duxbury, 2000.
2. DeMoivre A. The doctrine of chance. London, England: W. Pearson, 1718.
3. Bayes T. An essay toward solving a problem in the doctrine of chances. Philos Trans R Soc London 1763; 53:370–418. (Reprinted in Biometrika 1958; 45:293–315.)
4. Bernoulli J. Ars conjectandi. Basel, Switzerland: 1713. (Reprinted in: Die werke von Jakob Bernoulli. Vol 3. Basel, Switzerland: Birkhäuser Verlag, 1975; 106-286.
5. Laplace PS. Théorie analytique des probabilités. Paris, France: Ve. Courcier, 1812.
6. Gauss CF. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Hamburg, Germany: F. Perthes et I. H. Besser, 1809.
7. Applegate KE, Crewson PE. An introduction to biostatistics. Radiology 2001; 225:318–322.

Measurement Variability and Confidence Intervals in Medicine: Why Should Radiologists Care?[1]

L. Santiago Medina, MD, MPH
David Zurakowski, PhD

Index terms:
Radiology and radiologists, research
Statistical analysis

Published online before print
10.1148/radiol.2262011537
Radiology 2003; 226:297–301

Abbreviations:
ACL = anterior cruciate ligament
OR = odds ratio
PSV = peak systolic velocity
SEM = standard error of the mean

In radiology, appropriate diagnoses are often based on quantitative data. However, these data contain inherent variability. Radiologists often see *P* values in the literature but are less familiar with other ways of reporting statistics. Statistics such as the SD and standard error of the mean (SEM) are commonly used in radiology, whereas the CI is not often used. Because the SEM is smaller than the SD, it is often inappropriately used in order to make the variability of the data look tighter. However, unlike the SD, which quantifies the variability of the actual data for a single sample, the SEM represents the precision for an estimated mean of a general population taken from many sample means. Since readers are usually interested in knowing about the variability of the single sample, the SD often is the preferred statistic. Statistical calculations combine sample size and variability (ie, the SD) to generate a CI for a population proportion or population mean. CIs enable researchers to estimate population values without having data from all members of the population. In most cases, CIs are based on a 95% confidence level. The advantage of CIs over significance tests (*P* values) is that the CIs shift the interpretation from a qualitative judgment about the role of chance to a quantitative estimation of the biologic measure of effect. Proper understanding and use of these fundamental statistics and their calculations will allow more reliable analysis, interpretation, and communication of clinical information among health care providers and between these providers and their patients.
© RSNA, 2003

Radiologists and physicians rely heavily on quantitative data to make specific diagnoses. Furthermore, the patients and the referring physicians place trust in the radiologist's assessment of these quantitative data for determination of appropriate treatment. For example, a radiologist can inform a patient that he or she has a significant stenosis of the carotid artery because the peak systolic velocity (PSV) determined by using ultrasonography (US) is 280 cm/sec. The radiologist's final diagnosis of a significant stenosis of the carotid artery may suggest that treatment with endarterectomy or stent placement is indicated for this patient. But how reliable is the PSV of 280 cm/sec as a true estimate of significant stenosis of the carotid artery? How does this value help in the differentiation between stenosis and nonstenosis? What is the variability of the PSV when stenosis is present? Which population was included in the study and what was the sample size for determining normal and abnormal arterial lumen size? These are very important questions because the well-being of the patients and the credibility of the radiologists depend, in part, on a clear understanding of the answers to these questions.

Precise knowledge of important statistical parameters, such as the SD, the standard error of the mean (SEM), and the CIs, will provide the radiologist with answers to the questions previously posed. Most of these parameters can be quickly and easily obtained with a small calculator. In addition, these parameters are useful while reading the literature. Appropriate understanding and use of these fundamental statistics, namely, the SD, the SEM, and the CI, will allow more reliable analysis, interpretation, and communication of clinical information among health care providers and between these providers and their patients.

[1] From the Department of Radiology and Health Outcomes Policy and Economics (H.O.P.E.) Center, Children's Hospital, 3100 SW 62nd Ave, Miami, FL 33155; and Departments of Orthopaedic Surgery and Biostatistics, Children's Hospital, Harvard Medical School, Boston, Mass. Received September 17, 2001; revision requested November 12; revision received December 17; accepted January 21, 2002. **Address correspondence to** L.S.M. (e-mail: *smedina@post.harvard.edu*).

WHAT ARE RANDOM SAMPLING AND THE CENTRAL LIMIT THEOREM?

Obtaining a sample that is representative of a larger population is key in any study design. A random sample is a sample chosen to minimize bias (1, pp 248–277). A simple random sample is a sample in which every subject of the population has an equal chance of being included in the sample (1, pp 248–277). The only way to be sure of a representative sample is to select the subjects at random, so that whether or not each subject in the population is chosen for the sample is purely a matter of chance and is not based on the subject's characteristics (2, pp 33–36).

Random sampling has other advantages. Because the sample is randomly selected, the methods of probability theory can be applied to the data obtained. This enables the clinician to estimate the likely size of the errors that may occur, for example, with the SD or CIs, and to present them as part of the results (2, pp 33–36).

In general, if one has any series of independent identically distributed random variables, then their sum tends to produce a normal distribution as the number of variables increases (2, pp 116–120) (Fig 1). This fundamental theorem in statistics is known as the central limit theorem (1, pp 248–277; 2, pp 116–120). Simply stated, as sample size increases, the means of samples from a population of any distribution will approach the normal (Gaussian) distribution. This is an important property because it allows clinicians to use the normal distribution to formulate inferences from the data about means of populations. In addition, the variability of means of samples obtained from a population decreases as the sample size increases. However, the sample size required to make use of the central limit theorem depends on the underlying distribution of the population, and skewed populations require larger samples.

For example, suppose a group of radiologists want to study PSV in the common carotid artery of children in a small Amazon Indian tribe by using Doppler US spectrum analysis. Figure 1 shows that as the number of children selected in each series of random samples increases, the sum of these numbers tends to produce a normal distribution, as shown by the bell-shaped Gaussian distribution of PSV in the pediatric population sampled. As the sample size increases, the mean and SD come closer to the population's mean and SD (Fig 2).

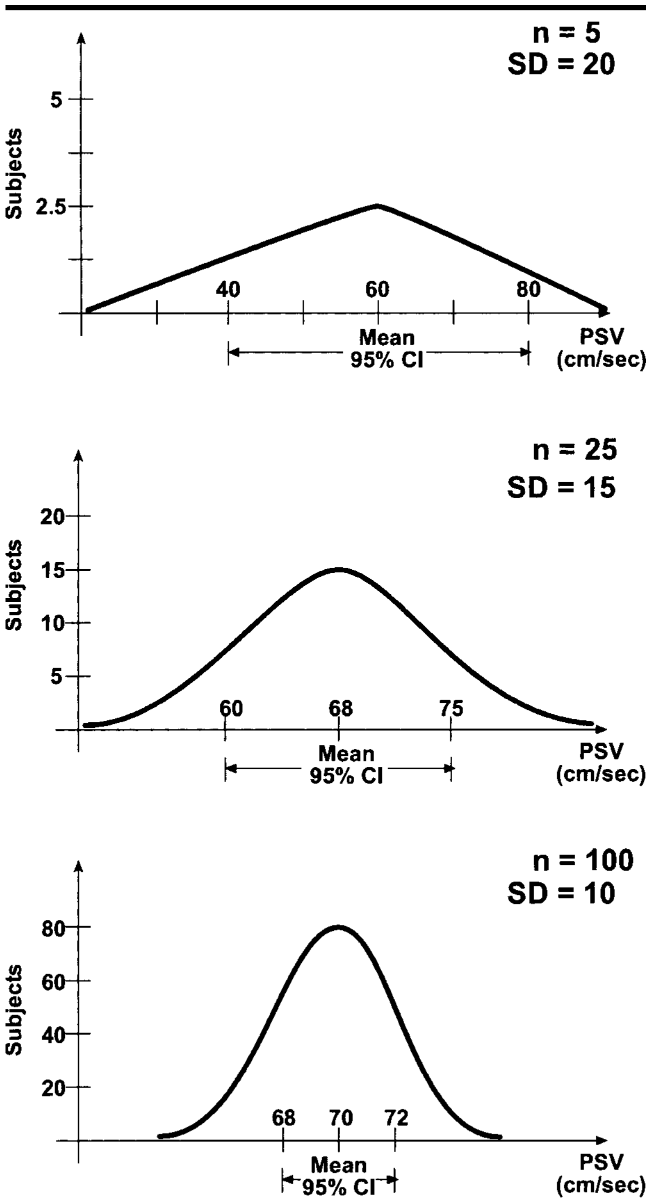


Figure 1. Graph shows PSV of the common carotid artery in an Amazon Indian population. Note that as the sample size increases from 5 to 100 subjects, the SD decreases and the 95% CI becomes narrower.

WHAT IS THE DIFFERENCE BETWEEN THE SD AND THE SEM?

The SD and the SEM measure two very different entities, but clinicians often confuse them. Some medical researchers summarize their data with the SEM because it is always smaller than the SD (3). Because the SEM is smaller, it often is inappropriately used to make the variability of the data look tighter. This kind of reporting of statistics should be discouraged.

The following example is given to illustrate the difference between the SD and the SEM and why one should summarize data by using the SD. Suppose that, in a study sample of patients with atherosclerotic disease, an investigator reported that the PSV in the carotid artery was 220 cm/sec and the SD was 10. Since the PSV in about 95% of all population members is within roughly 2 SDs of the mean, the results would tell one that, assuming the distribution is approximately normal, it would be unusual to observe a PSV less than 200 cm/sec or greater than 240 cm/sec in moderate atherosclerotic disease of the carotid artery. Therefore, a summary of the population and a range with which to compare spe-

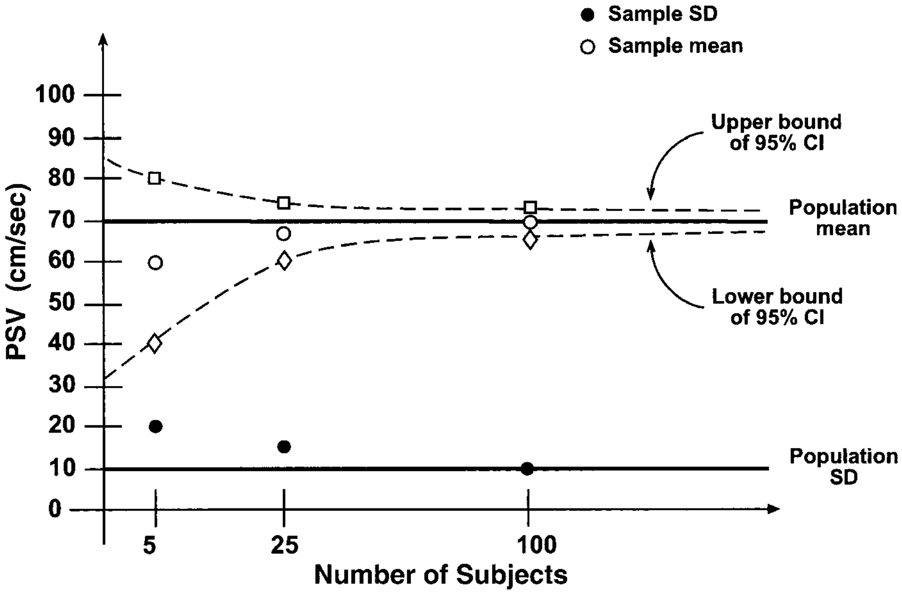


Figure 2. Graph summarizes the data from Figure 1. As the sample size increases, the sample mean and SD are a closer representation of the population's mean and SD. In addition, as the sample size increases, the 95% CI narrows.

cific patients who are examined by the clinician are described in the article.

Unfortunately, the investigator is quite likely to say that the PSV of the common carotid artery was 220 cm/sec ± 1.6 (SEM). If one confused the SEM with the SD, one would believe that the range of most of the population was narrow, between 216.8 and 223.2 cm/sec. These values describe the range that about 95% of the time includes the mean PSV of the entire population from which the sample of patients was chosen. The SEM is simply a measure of how far the sample mean is likely to be from the actual population mean. In practice, however, one generally wants to compare an individual patient's PSV with the spread of the population distribution as a whole and not with the population mean (3). This information is provided by the SD and not by the SEM.

WHAT ARE CIs?

Most biomedical research relies on the premise that what is true for a randomly selected sample from a population will be true, more or less, for the population from which the sample was chosen (1, pp 55–63). Therefore, measurements in the sample are used to estimate the characteristics of the population included in the study. The reliability of the results obtained from a sample is addressed by constructing CIs around statistics of the sample. The amount of variation associated with an estimate determined from a sample can be expressed by a CI.

A CI is the range of values that is believed to encompass the actual ("true") population value (1, pp 55–63). This true population value or parameter of interest usually is not known, but it does exist and can be estimated from an appropriately selected sample. CIs around population estimates provide information about how precise the estimate is. Wider CIs indicate lesser precision, while narrower ones indicate greater precision (Figs 1, 2). CIs provide bounds to estimates.

If one repeatedly obtained samples from the population and constructed CIs for each sample, then one could expect a certain percentage of the CIs to include the value of the true population and a certain percentage of them not to include that value. For example, with a 95% CI, the level of certainty is 95% of such CIs obtained in repeated sampling to include the true parameter value and only 5% of the CIs not to include the true parameter value.

HOW ARE CIs FOR A MEAN CALCULATED?

The mean of a set of measurements is calculated from the sample of patients. Therefore, the mean one calculates is unlikely to be exactly equal to the population mean. The size of the discrepancy depends on the size and variability of the sample (3, pp 163–190; 4). If the sample is small and variable, the sample mean may be far from the population mean. If the sample is large with little scatter, the sample mean will probably be very close to the population mean. Statistical calculations combine sample size and variability (ie, SD) to generate a CI for the population mean.

One can calculate an interval for any desired degree of confidence, although 95% CIs are by far the most commonly used. The following equation is the usual method for calculating a 95% CI that is based on a normally distributed sample with a known SD or one that is based on a sample from a population with an unknown SD but in which the population is known to be normally distributed and the sample itself is large (ie, $n > 100$):

$$95\%\ \mathrm{CI} = \mathrm{mean} \pm z \times [\mathrm{sample\ SD}/(\sqrt{n})], \quad (1)$$

where z (standardized score) is the value of the standard normal distribution with the specific level of confidence. For a 95% CI, $z = 1.96$ (approximately 2.0).

The scale of z scores is independent of the units of measurement. Therefore, for any measurement being investigated, one can calculate an individual's z score and compare it with that of other individuals. The z scores are calculated from the sample data as $(X - \mathrm{mean})/\mathrm{SD}$, where X is the actual individual's value. For example, if an individual's value is 1 SD above the mean for the group, that individual's z score is 1.0; a value 1 SD below the mean corresponds to a z score of -1.0. Approximately 68.0% of the area under the normal curve includes z scores between -1.0 and 1.0, approximately 95.0% of the area includes z scores between -2.0 and 2.0, and 99.7% of the area under the normal curve includes z scores between -3.0 and 3.0.

Equation (1) can be applied when the data conform to a normal (Gaussian) distribution and when the population SD is known. When the sample is small ($n < 100$) and information regarding the parametric SD is not known, one must rely on the sample SD, which requires setting CIs by using the t distribution. In this situation, the z value should be replaced with the appropriate critical value of the t distribution with $n - 1$ degrees of freedom, where n is the sample size.

The t distribution, or the Student t distribution, resembles the normal distribution, although its shape depends on the sample size. It is wider than the normal distribution to account for variability in estimating the mean and SD from the sample data (5). The *t* distribution differs from the normal distribution in that it assumes different shapes depending on the number of degrees of freedom. Therefore, when setting a CI around a mean, the appropriate critical value of the *t* distribution should be used in place of the *z* value in Equation (1). This *t* value can be found in a conventional *t* table included in most statistical textbooks. For example, in a study with a sample size of 25, the critical value for a *t* distribution that corresponds to a 95% CI, where $1 - \alpha$ is the confidence level and $n - 1$ indicates the degrees of freedom, is 2.064.

CIs can be constructed for any desired level of confidence. There is nothing magical about 95%, although it is traditionally used. If greater confidence is needed, then the CIs have to be wider. Consequently, 99% CIs are wider than 95% CIs, and 90% CIs are narrower than 95% CIs. Wider CIs are associated with greater confidence but less precision. This is the trade-off.

If one assumes that a sample was randomly selected from a certain population (that follows a normal distribution), one can be 95% sure that the CI includes the population mean. More precisely, if one generates many 95% CIs from many data sets, one can expect that the CI will include the true population mean in 95% of the cases and that the CI will not include the true mean value in the other 5%. Therefore, the 95% CI is related to statistical significance at the .05 level, which means that the CI itself can be used to determine if an estimated change is statistically significant at the .05 level (1, pp 55–63).

Whereas the *P* value is often interpreted as an indication of a statistically significant difference, the CI, by providing a range of values, allows the reader to interpret the implications of the results at either end of the range (1, pp 55–63; 6). For example, if one end of the range includes clinically important results but the other does not, the results can be regarded as inconclusive, not simply as an indication of a statistically significant difference or not. In addition, whereas *P* values are not presented in units, CIs are presented in the units of the variable of interest, and this latter presentation helps readers to interpret the results. CIs are generally preferred to *P* values because CIs shift the interpretation from a qualitative judgment about the role of chance to a quantitative estimation of the biologic measure of effect (1, pp 55–63; 6). More importantly, the CI quantifies the precision of the mean.

TABLE 1
Hypothetical Example: MR Imaging Depiction of ACL Tear

	Arthroscopic Findings		
MR Imaging Findings	Torn ACL	Intact ACL	Total
Torn ACL	394	32	426
Intact ACL	27	101	128
Total	421	133	554

Note.—Data are numbers of knees. Arthroscopic findings were the standard of reference. ACL = anterior cruciate ligament.

For example, findings in two hypothetical articles about US in the carotid artery in elderly patients indicate that a mean PSV of 200 cm/sec is associated with a 70% stenosis of the vascular diameter. Both articles reported the same SD of 50 cm/sec. However, one article was about a study that included 50 subjects, whereas the other one was about a study that included 500 subjects. At first glance, both articles appear to have the same information. This is delineated with the calculations here.

The calculations in the article with the smaller sample were as follows:

$$95\% \text{ CI} = 200 \pm 1.96\left(\frac{50}{\sqrt{50}}\right),$$

$$95\% \text{ CI} = 200 \pm 14.$$

The calculations in the article with the larger sample were as follows:

$$95\% \text{ CI} = 200 \pm 1.96\left(\frac{50}{\sqrt{500}}\right),$$

$$95\% \text{ CI} = 200 \pm 4.$$

However, in the article with the smaller sample, the 95% CI was 186 to 214 cm/sec, whereas in that with the larger sample, the 95% CI was 196 to 204 cm/sec. Therefore, the article with the larger sample has a narrower 95% CI.

WHY ARE CIs FOR SENSITIVITY AND SPECIFICITY OF A TEST IMPORTANT?

Most radiologists are familiar with the basic concepts of specificity and sensitivity and use them to evaluate the diagnostic accuracy of diagnostic tests in clinical practice. Since sensitivity and specificity are proportions, CIs can be calculated and should be reported in all research articles. CIs are needed to help one to be more certain about the clinical value of any screening or diagnostic test and to decide to what degree one can rely on the results. Omission of the precision of the sensitivity and specificity in a particular study can make a difference in the interpretation of the findings of that study (7).

The simplest diagnostic test is dichotomous, in which the results are used to classify patients into two groups according to the presence or absence of disease. Magnetic resonance (MR) imaging and arthroscopic findings from a hypothetical example are delineated in Table 1. In this hypothetical study, arthroscopy is considered the standard of reference. The question that arises in the clinical setting is, "How good is knee MR imaging at helping to distinguish torn and intact ACLs?" In other words, "To what degree can one rely on the interpretation of MR imaging in making judgments about the status of a patient's knee?"

One method of measuring the value of MR imaging in the detection of ACL tears is to calculate the proportion of torn ACLs and the proportion of intact ACLs that were correctly classified by using MR imaging. These proportions are known as the sensitivity and specificity of a test, respectively.

Sensitivity is calculated as the proportion of torn ACLs that were correctly classified by using MR imaging. In this example, of the 421 knees with ACL tears, 394 were correctly evaluated with MR imaging (Table 1). The sensitivity of MR imaging in the detection of ACL tears is, therefore, 94% (ie, sensitivity = 394/421 = 0.94). In other words, 94% of ACL tears were correctly classified as torn by using MR imaging. The 95% CI for a proportion can be determined by the equation shown here:

$$95\% \text{ CI} = p \pm z \times \sqrt{[p(1-p)]/n}. \tag{2}$$

By using Equation (2), the 95% CI for sensitivity is 0.94 ± 0.02, or 0.92 to 0.96. Therefore, one expects MR imaging to have a sensitivity between 92% and 96%.

TABLE 2
Data for Calculating OR and CI

Radiographs	Septic Arthritis	Transient Synovitis	Total
Effusion	63 (*a*)	33 (*b*)	96 (*a* + *b*)
No effusion	19 (*c*)	53 (*d*)	72 (*c* + *d*)
Total	82 (*a* + *c*)	86 (*b* + *d*)	168

Note.—Data are numbers of patients examined at radiography. The letters and/or symbols in parentheses are the variables that represent the given value in the equations used to calculate the OR and CI: OR = ad/bc = (63 × 53)/(33 × 19) = 3,339/627 = 5.3. CI = (OR) $\exp[\pm z\sqrt{(1/a + 1/b + 1/d)}]$. Adapted and reprinted, with permission, from Kocher et al (8).

Specificity is calculated as the proportion of intact ACLs that were correctly classified by using MR imaging. Of the 133 knees with an intact ACL, 101 were correctly classified. The specificity of MR imaging is, therefore, 76% (ie, specificity = 101/133 = 0.76). This means that 76% of intact ACLs were correctly classified as intact by using MR imaging. By using Equation (2), the 95% CI for specificity is 0.76 ± 0.07 or 0.69 to 0.83. Therefore, one expects MR imaging to have a specificity between 69% and 83%. It is also important to note that the CI was wider for specificity than it was for sensitivity because the sample groups were 133 (smaller) and 421 (larger), respectively.

CAN CIs FOR ODDS RATIOS BE CALCULATED?

CIs can also be calculated around risk measures, such as the relative risk or the odds ratio (OR). Consider an example in which radiographic effusion is examined to ascertain whether it is useful in the differentiation between septic arthritis and transient synovitis among children who have acute hip pain at presentation.

Data from a study by Kocher et al (8) are shown in Table 2 and can be summarized as follows: For those patients with radiographic effusion, the odds of having septic arthritis are 63/33 (effusion/no effusion) = 1.9. The odds of having septic arthritis for those with no radiographic effusion are 19/53 (effusion/no effusion) = 0.36. The OR is the ratio of these two odds: 1.9/0.36 = 5.3. This means that children with a radiographic effusion are approximately five times more likely to have septic arthritis than those without a radiographic effusion. The OR is sometimes referred to as the cross product ratio because it can be calculated by means of multiplication of the counts in the diagonal cells and division of data as follows (Table 2): OR = ad/bc = (63 × 53)/(33 × 19) = 3,339/627 = 5.3.

The OR is only a single number, that is, a "point estimate." The precision of this estimate can be described with a CI, which describes the statistical significance of the association between two variables within a specific range. The width of the CI reflects the amount of variability inherent in the OR. There is a trade-off between precision and confidence. Wider CIs provide greater certainty but are less precise. Narrower CIs are more precise but less certain that the truth is within the CI. In radiology and medicine, the most commonly reported CI corresponding to the OR is the 95% CI.

Several methods are commonly used to construct CIs around the OR. A simple method for constructing CIs (8) can be expressed as follows:

$$\mathrm{CI} = (\mathrm{OR})\exp\left[\pm z\sqrt{(1/a + 1/b + 1/c + 1/d)}\right], \quad (3)$$

where z is the value of the standard normal distribution with the specific level of confidence, and exp is the base of the natural logarithm (often symbolized as e).

By using Equation (3), the 95% CI in our example is calculated as follows:

$$95\%\ \mathrm{CI} = \log_e \mathrm{OR} \pm 1.96\sqrt{\frac{1}{63} + \frac{1}{33} + \frac{1}{19} + \frac{1}{53}}$$

$$95\%\ \mathrm{CI} = \log_e(5.3) \pm 1.96(0.348)$$

$$95\%\ \mathrm{CI} = \log_e(5.3) \pm 0.682$$

$$\mathrm{LL} = 1.67 - 0.682 = 0.988$$

$$\mathrm{UL} = 1.67 + 0.682 = 2.352$$

$$\mathrm{LL} = e^{0.988} = 2.69$$

$$\mathrm{UL} = e^{2.352} = 10.50,$$

where e is 2.718, LL represents the lower limit, and UL represents the upper limit.

Therefore, among children who have acute hip pain at presentation, those with a radiographic effusion are, on average, 5.3 times more likely to have septic arthritis compared with those with no radiographic effusion. The 95% CI lower limit of the OR is 2.7 and the upper limit is 10.5. When the 95% CI does not include 1.0 (as in this example), the results indicate a statistically significant difference at the .05 level (ie, $P < .05$).

CONCLUSION

The SD and SEM measure different parameters. The two are commonly confused in the medical literature. The SD can be thought of as a descriptive statistic that indicates the variation among measurements taken from a sample (1, pp 55–63). Investigators should not report summary statistics in terms of the SEM. The 95% CI is the preferred statistic for indicating the precision of an estimate of a population characteristic (1, pp 55–63).

CIs can be calculated for means as well as for proportions. Proportions commonly used in medicine include sensitivity, specificity, and the OR. Proportions should always be accompanied by 95% CIs. Proper understanding and use of fundamental statistics, such as the SD, the SEM, and the CI, and their calculations will allow more reliable analysis, interpretation, and communication of clinical data to patients and to referring physicians.

References

1. Lang TA, Secic M. How to report statistics in medicine. Philadelphia, Pa: American College of Physicians, 1997.
2. Bland M. An introduction to medical statistics. Oxford, England: Oxford Medical Publications, 1987.
3. Glantz SA. Primer of biostatistics. 2nd ed. New York, NY: McGraw-Hill, 1987.
4. Rosner B. Fundamentals of biostatistics, 4th ed. Belmont, Calif: Duxbury, 1995; 141–190.
5. Sokal RR, Rohlf FJ. Biometry. 3rd ed. New York, NY: Freeman, 1995; 143–150.
6. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ 1986; 292:746–750.
7. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. BMJ 1999; 318:1322–1323.
8. Kocher MS, Zurakowski D, Kasser JR. Differentiating between septic arthritis and transient synovitis of the hip in children: an evidence based prediction algorithm. J Bone Joint Surg Am 1999; 81:1662–1670.

Statistical Concepts Series

Kelly H. Zou, PhD
Julia R. Fielding, MD[2]
Stuart G. Silverman, MD
Clare M. C. Tempany, MD

Hypothesis Testing I: Proportions[1]

Index term:
Statistical analysis

Published online before print
10.1148/radiol.2263011500
Radiology 2003; 226:609–613

Abbreviations:
H_0 = null hypothesis
H_1 = alternative hypothesis
RDOG = Radiologic Diagnostic Oncology Group

[1] From the Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Mass (K.H.Z., J.R.F., S.G.S., C.M.C.T.); and Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115 (K.H.Z.). Received September 10, 2001; revision requested November 8; revision received December 12; accepted December 19. Supported in part by Public Health Service Grant NIH-U01 CA9398-03 awarded by the National Cancer Institute, Department of Health and Human Services. **Address correspondence to** K.H.Z. (e-mail: *zou@bwh.harvard.edu*).

Current address:
[2] Department of Radiology, University of North Carolina at Chapel Hill.

Statistical inference involves two analysis methods: estimation and hypothesis testing, the latter of which is the subject of this article. Specifically, *Z* tests of proportion are highlighted and illustrated with imaging data from two previously published clinical studies. First, to evaluate the relationship between nonenhanced computed tomographic (CT) findings and clinical outcome, the authors demonstrate the use of the one-sample *Z* test in a retrospective study performed with patients who had ureteral calculi. Second, the authors use the two-sample *Z* test to differentiate between primary and metastatic ovarian neoplasms in the diagnosis and staging of ovarian cancer. These data are based on a subset of cases from a multiinstitutional ovarian cancer trial conducted by the Radiologic Diagnostic Oncology Group, in which the roles of CT, magnetic resonance imaging, and ultrasonography (US) were evaluated. The statistical formulas used for these analyses are explained and demonstrated. These methods may enable systematic analysis of proportions and may be applied to many other radiologic investigations.

Statistics often involve a comparison of two values when one or both values are associated with some uncertainty. The purpose of statistical inference is to aid the clinician, researcher, or administrator in reaching a conclusion concerning a population by examining a sample from that population. Statistical inference consists of two components, estimation and hypothesis testing, and the latter component is the main focus of this article.

Estimation can be carried out on the basis of sample values from a larger population (1). Point estimation involves the use of summary statistics, including the sample mean and SD. These values can be used to estimate intervals, such as the 95% confidence level. For example, by using summary statistics, one can determine the sensitivity or specificity of the size and location of a ureteral stone for prediction of the clinical management required. In a study performed by Fielding et al (2), it was concluded that stones larger than 5 mm in the upper one-third of the ureter were very unlikely to pass spontaneously.

In contrast, hypothesis testing enables one to quantify the degree of uncertainty in sampling variation, which may account for the results that deviate from the hypothesized values in a particular study (3,4). For example, hypothesis testing would be necessary to determine if ovarian cancer is more prevalent in nulliparous women than in multiparous women.

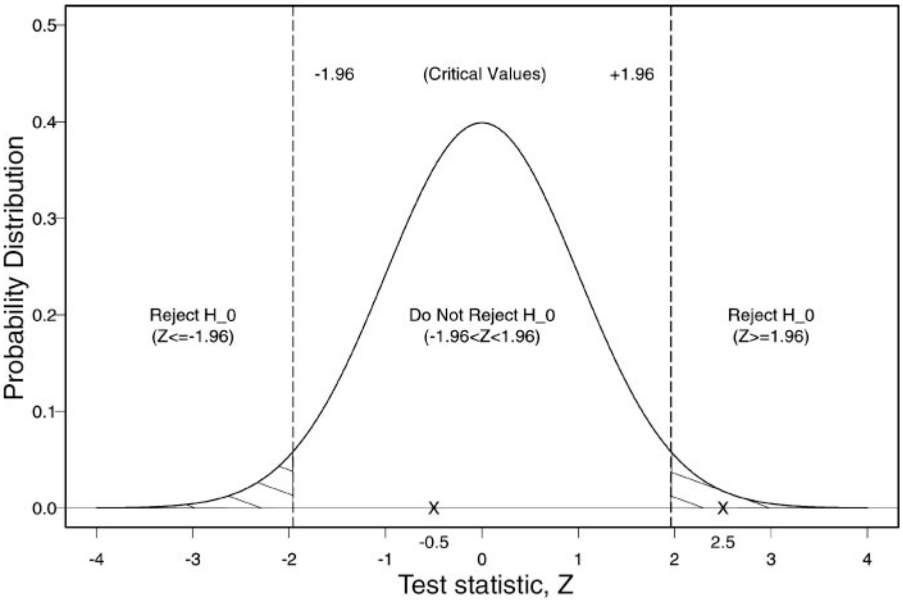
It is important to distinguish between a research hypothesis and a statistical hypothesis. The research hypothesis is a general idea about the nature of the clinical question in the population of interest. The primary purpose of the statistical hypothesis is to establish the basis for tests of significance. Consequently, there is also a difference between a clinical conclusion based on a clinical hypothesis and a statistical conclusion of significance based on a statistical hypothesis. In this article, we will focus on statistical hypothesis testing only.

In this article we review and demonstrate the hypothesis tests for both a single proportion and a comparison of two independent proportions. The topics covered may provide a basic understanding of the quantitative approaches for analyzing radiologic data. Detailed information on these concepts may be found in both introductory (5,6) and advanced textbooks (7–9). Related links on the World Wide Web are listed in Appendix A.

STATISTICAL HYPOTHESIS TESTING BASICS

A general procedure is that of calculating the probability of observing the difference between two values if they really are not different. This probability is called the *P* value, and this condition is called the null hypothesis (H_0). On the basis of the *P* value and whether it is low enough, one can conclude that H_0 is not true and that there really is a difference. This act of conclusion is in some ways a “leap of faith,” which is why it is known as statistical significance. In the following text, we elaborate on these key concepts and the definitions needed to understand the process of hypothesis testing.

There are five steps necessary for conducting a statistical hypothesis test: *(a)* formulate the null (H_0) and alternative (H_1) hypotheses, *(b)* compute the test statistic for the given conditions, *(c)* calculate the resulting *P* value, *(d)* either reject or do not reject H_0 (reject H_0 if the *P* value is less than or equal to a prespecified significance level [typically .05]; do not reject H_0 if the *P* value is greater than this significance level), and *(e)* interpret the results according to the clinical hypothesis relevant to H_0 and H_1. Each of these steps are discussed in the following text.



Graph illustrates the normal distribution of the test statistic *Z* in a two-sided hypothesis test. Under H_0, *Z* has a standard normal distribution, with a mean of 0 and a variance of 1. The critical values are fixed at ±1.96, which corresponds to a 5% significance level (ie, type I error) under H_0. The rejection regions are the areas marked with oblique lines under the two tails of the curve, and they correspond to any test statistic lying either below −1.96 or above +1.96. Two hypothetical test statistic values, −0.5 and 2.5, result in not rejecting or rejecting H_0, respectively.

Null and Alternative Hypotheses

In general, H_0 assumes that there is no association between the predictor and outcome variables in the study population. In such a case, a predictor (ie, explanatory or independent) variable is manipulated, and this may have an effect on another outcome or dependent variable. For example, to determine the effect of smoking on blood pressure, one could compare the blood pressure levels in nonsmokers, light smokers, and heavy smokers.

It is mathematically easier to frame hypotheses in null and alternative forms, with H_0 being the basis for any statistical significance test. Given the H_0 of no association between a predictor variable and an outcome variable, a statistical hypothesis test can be performed to estimate the probability of an association due to chance that is derived from the available data. Thus, one never accepts H_0, but rather one rejects it with a certain level of significance.

In contrast, H_1 makes a claim that there is an association between the predictor and outcome variables. One does not directly test H_1, which is by default accepted when H_0 is rejected on the basis of the statistical significance test results.

One- and Two-sided Tests

The investigator must also decide whether a one- or two-sided test is most suitable for the clinical question (4). A one-sided H_1 test establishes the direction of the association between the predictor and the outcome—for example, that the prevalence of ovarian cancer is higher in nulliparous women than in parous women. In this example, the predictor is parity and the outcome is ovarian cancer. However, a two-sided H_1 test establishes only that an association exists without specifying the direction—for example, the prevalence of ovarian cancer in nulliparous women is different (ie, either higher or lower) from that in parous women. In general, most hypothesis tests involve two-sided analyses.

Test Statistic

The test statistic is a function of summary statistics computed from the data. A general formula for many such test statistics is as follows: test statistic = (relevant statistic − hypothesized parameter value)/(standard error of the relevant statistic), where the relevant statistics and standard error are calculated on the basis of the sample data. The standard error is the indicator of variability, and much of the complexity of the hypothesis test involves estimating the standard error correctly. H_0 is rejected if the test statistic exceeds a certain level (ie, critical value).

For example, for continuous data, the Student *t* test is most often used to determine the statistical significance of an observed difference between mean values with unknown variances. On the basis of large samples with underlying normal distributions and known variances (5), the *Z* test of two population means is often conducted. Similar to the *t* test, the *Z* test involves the use of a numerator to compare the difference between the sample means of the two study groups with the difference that would be expected with H_0, that is, zero difference. The denominator includes the sample size, as well as the variances, of each study group (5).

Once the *Z* value is calculated, it can be converted into a probability statistic by means of locating the *P* value in a standard reference table. The Figure illustrates a standard normal distribution (mean of 0, variance of 1) of a test statistic, *Z,* with two rejection regions that are either below −1.96 or above 1.96. Two hypothetical test statistic values, −0.5 and 2.5, which lie outside and inside the rejection regions, respectively, are also included. Consequently, one does not reject H_0 when *Z* equals −0.5, but one does reject H_0 when *Z* equals 2.5.

P Value

When we conclude that there is statistical significance, the *P* value tells us

Cross Tabulation Showing Relationship between the Two Error Types

Test Result, Underlying Truth	H_0 True	H_0 False
Do not reject H_0	Correct action ($1 - \alpha$)	Type II error (β)
Reject H_0	Type I error (α)	Correct action ($1 - \beta$)

Note.—A statistical power of $1 - \beta$ is analogous to the sensitivity of a diagnostic test. The probability $1 - \alpha$ is analogous to the specificity of a diagnostic test.

what the probability is that our conclusion is wrong when in fact H_0 is correct. The lower the *P* value, the less likely that our rejection of H_0 is erroneous. By convention, most analysts will not claim that they have found statistical significance if there is more than a 5% chance of being wrong ($P = .05$).

Type I and II Errors

Two types of errors can occur in hypothesis testing: A type I error (significance level α) represents the probability that H_0 was erroneously rejected when in fact it is true in the underlying population. Note that the *P* value is not the same as the α value, which represents the significance level in a type I error. The significance level α is prespecified (5% conventionally), whereas the *P* value is computed on the basis of the data and thus reflects the strength of the rejection of H_0 on the test statistic. A type II error (significance level β) represents the probability that H_0 was erroneously retained when in fact H_1 is true in the underlying population. There is always a trade-off between these two types of errors, and such a relationship is similar to that between sensitivity and specificity in the diagnostic literature (Table) (10). The probability $1 - \beta$ is the statistical power and is analogous to the sensitivity of a diagnostic test, whereas the probability $1 - \alpha$ is analogous to the specificity of a diagnostic test.

STATISTICAL TESTS OF PROPORTIONS: THE *Z* TEST

We now focus on hypothesis testing for either a proportion or a comparison of two independent proportions. First, we study a one-sample problem. In a set of independent trials, one counts the number of times that a certain interesting event (eg, a successful outcome) occurs. The underlying probability of success (a proportion) is compared against a hypothesized value. This proportion can be the diagnostic accuracy (eg, sensitivity or specificity) or the proportion of patients whose cancers are in remission. We also study a two-sample problem in which trials are conducted independently in two study groups. For example, one may compare the sensitivities or specificities of two imaging modalities. Similarly, patients in one group receive a new treatment, whereas independently patients in the control group receive a conventional treatment, and the proportions of remission in the two patient populations are compared.

When sample sizes are large, the approximate normality assumptions hold for both the sample proportion and the test statistic. In the test of a single proportion (π) based on a sample of *n* independent trials at a hypothesized success probability of π_0 (the hypothesized proportion), both $n\pi_0$ and $n(1 - \pi_0)$ need to be at least 5 (Appendix B). In the comparison of two proportions, π_1 and π_2, based on two independent sample sizes of n_1 and n_2 independent trials, respectively, both n_1 and n_2 need to be at least 30 (Appendix C) (5). The test statistic is labeled *Z*, and, hence, the analysis is referred to as the *Z* test of a proportion. Other exact hypothesis-testing methods are available if these minimum numbers are not met.

Furthermore, the *Z* and Student *t* tests both are parametric hypothesis tests—that is, they are based on data with an underlying normal distribution. There are many situations in radiology research in which the assumptions needed to use a parametric test do not hold. Therefore, nonparametric tests must be considered (9). These statistical tests will be discussed in a future article.

TWO RADIOLOGIC EXAMPLES

One-Sample *Z* Test of a Single Proportion

Fielding et al (2) evaluated the unenhanced helical CT features of 100 ureteral calculi, 71 of which passed spontaneously and 29 of which required intervention. According to data in the available literature (11–13), approximately 80% of the stones smaller than 6 mm in diameter should have passed spontaneously. Analysis of the data in the Fielding et al study revealed that of 66 stones smaller than 6 mm, 57 (86%) passed spontaneously. To test if the current finding agrees with that in the literature, we conduct a statistical hypothesis test with five steps:

1. H_0 is as follows: 80% of the ureteral stones smaller than 6 mm will pass spontaneously ($\pi = 0.80$). H_1 is as follows: The proportion of the stones smaller than 6 mm that pass spontaneously does not equal 80%—that is, it is either less than or greater than 80% ($\pi \neq 0.80$). This is therefore a two-sided hypothesis test.
2. The test statistic *Z* is calculated to be 1.29 on the basis of the results of the *Z* test of a single proportion (5).
3. The *P* value, .20, is the sum of the two tail probabilities of a standard normal distribution for which the *Z* values are beyond ±1.29 (Figure).
4. Because the *P* value, .20, is greater than the significance level α of 5%, H_0 is not rejected.
5. Therefore, our data support the belief that 80% of the stones smaller than 6 mm in diameter will pass spontaneously, as reported in the literature. Thus, H_0 is not rejected, given the data at hand. Consequently, it is possible that a type II error will occur if the true proportion in the population does not equal 80%.

Two-Sample *Z* Test to Compare Two Independent Proportions

Brown et al (14) hypothesized that the imaging appearances (eg, multilocularity) of primary ovarian tumors and metastatic tumors to the ovary might be different. Data were obtained from 280 patients who had an ovarian mass and underwent US in the Radiologic Diagnostic Oncology Group (RDOG) ovarian cancer staging trial (15,16). The study results showed that 30 (37%) of 81 primary ovarian cancers, as compared with three (13%) of 24 metastatic neoplasms, were multilocular at US. To test if the respective underlying proportions are different, we conduct a statistical hypothesis test with five steps:

1. H_0 is as follows: There is no difference between the proportions of multilocular metastatic tumors (π_1) and multilocular primary ovarian tumors (π_2) among the primary and secondary ovarian cancers—that is, $\pi_1 - \pi_2 = 0$. H_1 is as follows: There is a difference in these proportions: One is either less than or greater than the other—that is, $\pi_1 - \pi_2 \neq 0$. Thus, a two-sided hypothesis test is conducted.

2. The test statistic Z is calculated to be 2.27 on the basis of the results of the Z test to compare two independent proportions (5).

3. The P value, .02, is the sum of the two tail probabilities of a standard normal distribution for which the Z values are beyond ±2.27 (Figure).

4. Because the P value, .02, is less than the significance level α of 5%, H_0 is rejected.

5. Therefore, there is a statistically significant difference between the proportion of multilocular masses in patients with primary tumors and that in patients with metastatic tumors.

SUMMARY AND REMARKS

In this article, we reviewed the hypothesis tests of a single proportion and for comparison of two independent proportions and illustrated the two test methods by using data from two prospective clinical trials. Formulas and program codes are provided in the Appendices. With large samples, the normality of a sample proportion and test statistic can be conveniently assumed when conducting Z tests (5). These methods are the basis for much of the scientific research conducted today; they allow us to make conclusions about the strength of research evidence, as expressed in the form of a probability.

Alternative exact hypothesis-testing methods are available if the sample sizes are not sufficiently large. In the case of a single proportion, the exact binomial test can be conducted. In the case of two independent proportions, the proposed large-sample Z test is equivalent to a test based on contingency table (ie, χ^2) analysis. When large samples are not available, however, the Fisher exact test based on contingency table analysis can be adopted (8,17–19). For instance, in the clinical example involving data from the RDOG study, the sample of 24 metastatic neoplasms is slightly smaller than the required sample of 30 neoplasms, and, thus, use of the exact Fisher test may be preferred.

The basic concepts and methods reviewed in this article may be applied to similar inferential and clinical trial design problems related to counts and proportions. More complicated statistical methods and study designs may be considered, but these are beyond the scope of this tutorial article (20–24). A list of available software packages can be found by accessing the Web links given in Appendix A.

TABLE B1
Testing a Single Proportion by Using a One-Sample Z Test

Procedure	H_1: $\pi > \pi_0$	H_1: $\pi < \pi_0$
Do not reject H_0 if p is inconsistent with H_1—that is, if	$p < \pi_0$	$p > \pi_0$
Perform the Z test if	$p \geq \pi_0$	$p \leq \pi_0$
Compute the Z test statistic	$z = \dfrac{p - \pi_0}{\sqrt{[\pi_0(1-\pi_0)]/n}}$	Same
Compute the P value	Probability ($Z > z$)	Probability ($Z < z$)
With $\alpha = .05$, reject H_0 if	P value $\leq \alpha$	Same

Note.—Under H_0, Z has a standard normal distribution, with a mean of 0 and a variance of 1. The P value from a two-sided test is twice that from a one-sided test, as shown above. Large sample assumption requires that both $n\pi_0$ and $n(1 - \pi_0)$ are greater than or equal to 5, where $n\pi_0$ and $n(1 - \pi_0)$ represent numbers of trials; the sum of these two numbers represents the total number of trials (n) in the sample.

TABLE C1
Comparing Two Independent Proportions by Using a Two-Sample Z Test

Procedure	H_1: $\pi_1 - \pi_2 > 0$	H_1: $\pi_1 - \pi_2 < 0$
Do not reject H_0 if $p_1 - p_2$ is inconsistent with H_1—that is, if	$p_1 - p_2 < 0$	$p_1 - p_2 > 0$
Perform the test if	$p_1 - p_2 \geq 0$	$p_1 - p_2 \leq 0$
Compute the test statistic	$z = \dfrac{p_1 - p_2}{\sqrt{[p_c(1 - p_c)]/(1/n_1 + 1/n_2)}}$	Same
Compute the P value	Probability ($Z > z$)	Probability ($Z < z$)
With $\alpha = .05$, reject H_0 if	P value $\leq \alpha$	Same

Note.—Under H_0, Z has a standard normal distribution, with a mean of 0 and a variance of 1. The P value from a two-sided test is twice that from a one-sided test, as shown above. Large sample assumption requires that both numbers of trials, n_1 and n_2, are greater than or equal to 30.

APPENDIX A

Statistical Resources Available on the World Wide Web

The following are links to electronic textbooks on statistics: *www.davidmlane.com/hyperstat/index.html, www.statsoft.com/textbook/stathome.html, www.ruf.rice.edu/~lane/rvls.html, www.bmj.com:/collections/statsbk/index.shtml, and espse.ed.psu.edu/statistics/investigating.htm.* In addition, statistical software packages are available at the following address: *www.amstat.org/chapters/alaska/resources.htm.*

APPENDIX B

Testing a Single Proportion by Using a One-Sample Z Test

Let π be a population proportion to be tested (Table B1). The procedure for deciding whether or not to reject H_0 is as follows: $\pi = \pi_0$; this is based on the results of a one-sided, one-sample Z test at the significance level of α with n independent trials (Table B1). The observed number of successes is x, and, thus, the sample proportion of successes is $p = x/n$. In our first clinical example, that in which the unenhanced helical CT features of 100 ureteral calculi were evaluated (2), $\pi = 0.80$, $n = 66$, $x = 57$, and $p = 66/57$ (0.86).

APPENDIX C

Comparing Two Independent Proportions by Using a Two-Sample Z Test

Let π_1 and π_2 be the two independent population proportions to be compared (Table C1). The procedure for deciding whether or not to reject H_0 is as follows: $\pi_1 - \pi_2 = 0$; this is based on the results of a one-sided, two-sample Z test at the significance level of α with two independent trials: sample sizes of n_1 and n_2, respectively (Table C1). The observed numbers of successes in these two samples are $p_1 = x_1/n_1$ and $p_2 = x_2/n_2$, respectively. To denote the pooled proportion of successes over the two samples, use the following equation: $p_c = (x_1 + x_2)/(n_1 + n_2)$. In our second clinical example, that involving 280 patients with ovarian masses in the RDOG ovarian cancer staging trial (15,16), $n_1 = 81$, $x_1 = 30$, $n_2 = 24$, $x_2 = 3$, $p_1 = x_1/n_1$ (30/81 [0.37]), $p_2 = x_2/n_2$ (3/24 [0.13]), and $p_c = (x_1 + x_2)/(n_1 + n_2)$, or 33/105 (0.31).

Acknowledgments: We thank Kimberly E. Applegate, MD, MS, and Philip E. Crewson, PhD, for their constructive comments on earlier versions of this article.

References

1. Altman, DG. Statistics in medical journals: some recent trends. Statist Med 2000; 1g:3275-3289.
2. Fielding JR, Silverman SG, Samuel S, Zou KH, Loughlin KR. Unenhanced helical CT of ureteral stones: a replacement for excretory urography in planning treatment. AJR Am J Roentgenol 1998; 171:1051–1053.
3. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986; 292:746–750.
4. Bland JM, Altman DF. One and two sided tests of significance. BMJ 309:248, 1994.
5. Goldman RN, Weinberg JS. Statistics: an introduction. Englewood Cliffs, NJ: Prentice Hall, 1985; 334–353.
6. Hulley SB, Cummings SR. Designing clinical research: an epidemiologic approach. Baltimore, Md: Williams & Wilkins, 1988; 128–138, 216–217.
7. Freund JE. Mathematical statistics. 5th ed. Englewood Cliffs, NJ: Prentice Hall, 1992; 425–430.
8. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY: Wiley, 1981; 1–49.
9. Gibbons JD. Sign tests. In: Kotz S, Johnson NL, eds. Encyclopedia of statistical sciences. New York, NY: Wiley, 1982; 471–475.
10. Browner WS, Newman TB. Are all significant p values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987; 257:2459–2463.
11. Drach GW. Urinary lithiasis: etiology, diagnosis and medical management. In: Walsh PC, Staney TA, Vaugham ED, eds. Campbell's urology. 6th ed. Philadelphia, Pa: Saunders, 1992; 2085–2156.
12. Segura JW, Preminger GM, Assimos DG, et al. Ureteral stones: clinical guidelines panel summary report on the management of ureteral calculi. J Urol 1997; 158: 1915–1921.
13. Motola JA, Smith AD. Therapeutic options for the management of upper tract calculi. Urol Clin North Am 1990; 17:191–206.
14. Brown DL, Zou KH, Tempany CMC, et al. Primary versus secondary ovarian malignancy: imaging findings of adnexal masses in the Radiology Diagnostic Oncology study. Radiology 2001; 219:213–218.
15. Kurtz AB, Tsimikas JV, Tempany CMD, et al. Diagnosis and staging of ovarian cancer: comparative values of Doppler and conventional US, CT, and MR imaging correlated with surgery and histopathologic analysis—report of the Radiology Diagnostic Oncology Group. Radiology 1999; 212:19–27.
16. Tempany CM, Zou KH, Silverman SG, Brown DL, Kurtz AB, McNeil BJ. Stating of ovarian cancer: comparison of imaging modalities—report from the Radiology Diagnostic Oncology Group. Radiology 2000; 215:761–767.
17. Agresti A. Categorical data analysis. New York, NY: Wiley, 1990; 8–35.
18. Joe H. Extreme probabilities for contingency tables under row and column independence with application to Fisher's exact test. Comm Stat A Theory Methods 1988; 17:3677–3685.
19. MathSoft/Insightful. S-Plus 4 guide to statistics. Seattle, Wash: MathSoft, 1997; 89-96. Available at: http://www.insightful.com/products.splus.
20. Lehmann EL, Casella G. Theory of point estimation. New York, NY: Springer Verlag, 1998.
21. Lehmann EL. Testing statistical hypotheses. 2nd ed. New York, NY: Springer Verlag, 1986.
22. Hettmansperger TP. Statistical inference based on ranks. Malabar, Fla: Krieger, 1991.
23. Joseph L, Du Berger R, Belisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. Stat Med 1997; 16:769–781.
24. Zou KH, Norman SL. On determination of sample size in hierarchical binomial models. Stat Med 2001; 20:2163–2182.

Statistical Concepts Series

Hypothesis Testing II: Means[1]

Richard Tello, MD, MSME, MPH
Philip E. Crewson, PhD

Index term:
Statistical analysis

Published online before print
10.1148/radio1.2271020085
Radiology 2003; 227:1–4

Abbreviation:
ANOVA = analysis of variance

Whenever means are reported in the literature, they are likely accompanied by tests to determine statistical significance. The *t* test is a common method for statistical evaluation of the difference between two sample means. It provides information on whether the means from two samples are likely to be different in the two populations from which the data originated. Similarly, paired *t* tests are common when comparing means from the same set of patients before and after an intervention. Analysis of variance techniques are used when a comparison involves more than two means. Each method serves a particular purpose, has its own computational formula, and uses a different sampling distribution to determine statistical significance. In this article, the authors discuss the basis behind analysis of continuous data with use of paired and unpaired *t* tests, the Bonferroni correction, and multivariate analysis of variance for readers of the radiology literature.
© RSNA, 2003

To establish if there is a statistically significant difference in two groups that are measured with a continuous variable, such as patient height versus sex, a test of the hypothesis that there is no difference must be performed. In this article, we discuss the application of three commonly used methods for testing the difference of means. The three methods are the independent samples *t* test, the paired samples *t* test, and one-way analysis of variance (ANOVA). Each method is used to compare means obtained from continuous sample data, but each is designed to serve a particular purpose. All three approaches require normally distributed variables and produce an estimate of whether there is a significant difference between means. Independent samples *t* tests provide information on whether the means from two samples are likely to be different in the two populations from which the data originated. Paired samples *t* tests compare means from the same set of observations (patients) before and after an intervention is performed. ANOVA is used to test the differences between three or more sample means. The *t* test is a commonly used statistical test that is easy to calculate but can be misapplied (1,2). If a radiologist wishes to compare two proportions, such as the sensitivity of two tests, the appropriate test is the χ^2 test, which is addressed in other articles in this series. What follows in this article is a brief introduction to three techniques for testing the difference of means.

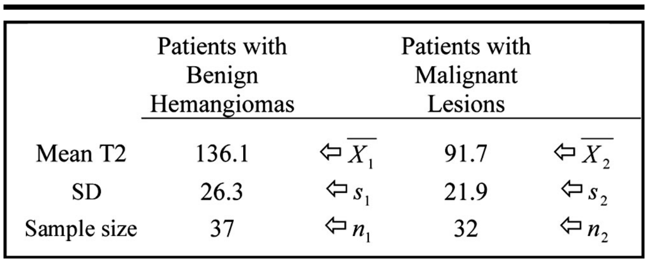
HYPOTHESIS TESTING FOR TWO SAMPLE MEANS

The *t* test can be used to evaluate the difference between two means from two independent samples or between two samples for which the observations in the second sample are not independent of those in the first sample. The latter is commonly referred to as a paired *t* test. Both paired and unpaired *t* tests use the *t* sampling distribution to determine the *P* value. As reported in a previous article (3), test statistics are compared with sampling distributions to determine the probability of error. The *t* distribution is similar to the standard normal distribution, except that it compensates for small sample sizes (especially fewer than 30 observations). As total sample size of the two groups increases beyond 120 observations, the two sampling distributions are virtually identical. This attribute allows the *t* distribution to be used for all sample sizes, large or small.

Independent Samples *t* Tests

If a researcher wants to compare means collected from two patient populations, a *t* test for independent samples will often be used. The term *independent samples* indicates that none of the patients in one sample are included in the second sample. The following research question will serve as an example: Are T2s in magnetic resonance (MR) imaging of malignant hepatic masses different from those for benign hepatic masses? Table 1

[1] From the Department of Radiology, Boston University School of Medicine, 88 E Newton St, Atrium 2, Boston, MA 02118 (R.T.); and Health Services Research and Development Service, Department of Veterans Affairs, Washington, DC (P.E.C.). Received February 11, 2002; revision requested March 18; revision received April 1; accepted May 1. Address correspondence to R.T. (e-mail: *tello@alum.mit.edu*).



	Patients with Benign Hemangiomas		Patients with Malignant Lesions	
Mean T2	136.1	⇦ $\overline{X_1}$	91.7	⇦ $\overline{X_2}$
SD	26.3	⇦ s_1	21.9	⇦ s_2
Sample size	37	⇦ n_1	32	⇦ n_2

Calculations for the independent samples *t* test statistic reported in Table 1.

TABLE 1
Example of the Independent Samples (unpaired) *t* Test of T2s for Benign Hemangiomas and Malignant Lesions

Parameter	Benign Hemangiomas	Malignant Lesions
Mean T2 (msec) ± SD	136.1 ± 26.3	91.7 ± 21.9
No. of patients	37	32

Note.—*t* Test results indicate a significant difference between mean T2 values in the two groups ($t = 7.44$, $P < .001$) (4).

presents summary statistics from two patient samples to answer this question (Figure). One sample includes only patients with malignant tumors. The second sample includes only patients with benign tumors (hemangiomas). The average T2 for malignant lesions is about 92 msec, while hemangiomas had an average T2 of 136 msec. The *t* test is used to test the null hypothesis that the T2 for malignant tumors is not different from that for benign tumors: To conduct this test, the difference between the two means is used in conjunction with the variation found in both samples (SD) and sample sizes to compute a *t* test statistic. The *t* test formula is

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_{\overline{x_1} - \overline{x_2}}},$$

where $\overline{X_1}$ is the mean for sample 1, $\overline{X_2}$ is the mean for sample 2, and S is the variance. The pooled standard error of the difference between two sample means is

$$S_{\overline{x_1} - \overline{x_2}} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

where n_1 is the size of sample 1, n_2 is the size of sample 2, s_1 is the variance of sample 1, and s_2 is the variance of sample 2.

As shown in Table 1, the calculations result in a *t* test statistic of 7.44, which, when compared with the *t* distribution, produces a *P* value of less than .001 (4) by performing the following equations:

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_{\overline{X_1} - \overline{X_2}}} = \frac{136.1 - 91.7}{S_{\overline{X_1} - \overline{X_2}}};$$

$$S_{\overline{X_1} - \overline{X_2}}$$

$$= \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

$$= \sqrt{\frac{37(26.3)^2 + 32(21.9)^2}{37 + 32 - 2}} \sqrt{\frac{37 + 32}{37(32)}}$$

$$= 5.96;$$

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_{\overline{X_1} - \overline{X_2}}} = \frac{136.1 - 91.7}{5.96} = 7.44.$$

Assuming a .05 cutoff *P* value, the null hypothesis is rejected in favor of the conclusion that on average, there is a statistically significant difference in the T2 for MR imaging of malignant tumors compared with that for benign tumors.

Homogeneity of Variance

The *t* test assumes that the variances in each group are equal (termed *homogeneous).* Alternative methods can be used when this assumption is not valid. Statistical software programs often automatically compute *t* tests and report results for both equal and unequal variances. Alternate approximations of the *t* test when the variances are unequal can be found in a publication by Rosner (5). It is in this setting that the need to determine if variances are equal requires another statistical test.

Determination of whether the assumption of equal variances is valid requires the use of an F test. The F test involves conducting a variance ratio test (6). This calculation tests the null hypothesis that the variances are equal. If the results of the F test are statistically significant ($P <$ .05), this suggests that the variance of the two groups is not equal. If this occurs, two recommended solutions are to either modify the *t* test to compensate for unequal variances or use a nonparametric test, called the Mann-Whitney test (7).

Paired *t* Tests

Two samples are paired when each data point of the first sample is matched and related to a data point in the second sample. This is common in studies in which measurements are collected from the same patients before and after intervention. Table 2 presents data on the size of a cancerous tumor in the same patient before and after receiving a new treatment. The research question represented in Table 2 is whether the new therapy affects tumor size. There are two groups represented by the same seven patients. One group is represented by the patient sample before therapy. The second group is represented by the same sample of patients after therapy. Before therapy, mean tumor size was 4.86 cm. After therapy, mean tumor size was 4.50 cm, representing a mean decrease of .36 cm. The *t* test is used to test the null hypothesis that the mean difference in tumor size between the groups before and after therapy does not differ significantly from zero, with the assumption that this difference is distributed normally. The test is used to compare the observed difference obtained from the sample of seven patients with the hypothesized value of no difference in the population. The paired *t* test formula is

$$t = \frac{\bar{d} - 0}{S_{\bar{d}}},$$

where $\bar{d}$ is the observed mean difference, and $S_{\bar{d}}$ is the standard error of the observed mean difference. On the basis of a *t* test statistic of 5.21, calculated as follows:

$$t = \frac{\bar{d} - 0}{S_{\bar{d}}} = \frac{0.36 - 0}{0.18/\sqrt{7}} = \frac{0.36}{0.068} = 5.21,$$

the probability of falsely rejecting the null hypothesis of no change in size (ie, the observed difference is due to random chance) is .002. Hence, the null hypothesis is rejected in favor of the conclusion that tumors shrank in patients who underwent therapy.

ANOVA

In the previous section, we compared the means of two normally distributed variables with the two-sample *t* test. When the means of more than two distributions

TABLE 2
Example of the Paired *t* Test to Evaluate Tumor Size before and after Therapy in Seven Patients

	Tumor Size (cm)		
Patient No.	Before Therapy	After Therapy	Change (cm)
1	5.3	5.0	0.3
2	4.4	4.0	0.4
3	4.9	4.5	0.4
4	6.4	6.0	0.4
5	3.4	3.0	0.4
6	5.0	5.0	0.0
7	4.6	4.0	0.6
Mean size ± SD	4.86 ± 0.91	4.50 ± 0.96	0.36 ± 0.18

Note.—The calculated statistic suggests a significant difference between the means before and after treatment ($t = 5.21$, $P < .002$).

TABLE 3
Example of ANOVA to Compare Differences in Vertebral Density in Three Age Groups of Women

	Age Groups		
Parameter	46–55 Years	56–65 Years	66–75 Years
Mean density ± SD	1.12 ± 0.18	1.13 ± 0.20	1.03 ± 0.19
Sample size	46	63	50

Note.—F = 4.499, $P < .013$ (8).

must be compared, one-way ANOVA is used. With ANOVA, the means of two or more independent groups (each of which follow a normal distribution and have similar SDs) can be evaluated to ascertain the relative variability between the groups compared with the variability within the groups.

ANOVA calculations are best performed with statistical software (software easily capable of calculating the *t* test and variances include Excel version 5.0 [Microsoft, Bothell, Wash]; more sophisticated analyses for performance of ANOVA include Stata version 5.0 [College Park, Tex], SPSS [Chicago, Ill], or SAS [Cary, NC]), but the basic approach is to compare the means and variances of independent groups to determine if the groups are significantly different from one another. The null hypothesis proposes that the samples come from populations with the same mean and variance. The alternative hypothesis is that at least two of the means are not the same. If ANOVA is used with two groups, it would produce results comparable to those obtained with the two independent samples *t* test.

Table 3 presents data on vertebral bone density for three groups of women (8). The research question represented in Table 3 is whether the mean vertebral bone density varies among the three groups; hence, ANOVA is used to determine if there is a significant difference. In this example, age groups of 46–55 years, 56–65 years, and 66–75 years are used to represent the three populations for patient screening. As reported in Table 3, the mean vertebral bone density is 1.12 for women 46–55 years of age, 1.13 for women 56–65 years of age, and 1.03 for women 66–75 years of age. Calculation of the test statistic involves estimation of a ratio of the variance between the groups to the variance within the groups. The ratio, called the F statistic, is compared with the F sampling distribution instead of the *t* distribution discussed earlier (5). An F statistic of 1.0 occurs when the variance between the groups is the same as the variance within the groups.

The F statistic is the mean squares between groups, divided by the mean squares within groups:

$$\mathrm{F} = [\Sigma n_k(\overline{X}_i - \overline{X}_g)^2/K - 1]/ \{[\Sigma(X_i - \overline{X}_1)^2 + \Sigma(X_i - \overline{X}_2)^2 + \Sigma(X_i - \overline{X}_3)^2]/N - K\},$$

where $\overline{X}_i$ is each group mean, $\overline{X}_g$ is the grand mean for all the groups [(sum of all scores)/N], n_k is the number of patients in each group, K is the number of groups, and N is the total number of scores.

The sum of squares between groups is $\Sigma n_k(\overline{X}_i - \overline{X}_g)^2$. The sum of squares within groups is $\Sigma(X_i - \overline{X}_1)^2 + \Sigma(X_i - \overline{X}_2)^2 + \Sigma(X_i - \overline{X}_3)^2$. The example presented in Table 3 can be calculated by using the F statistic formula; however, the descriptive data in Table 3 would require manipulation. The sum of squares within groups is the variance multiplied by the number of scores in a group.

The F statistic in Table 3 is 4.499. This indicates that there is greater variation between the group means than within each group. In this case, there is a statistically significant difference ($P < .013$) between at least two of the age groups.

THE MULTIPLE COMPARISON PROBLEM

The vertebral bone density example could also be analyzed by using three *t* tests (46–55-year-old group vs 56–65-year-old group; 46–55-year-old group vs 66–75-year-old group; and 56–65-year-old group vs 66–75-year-old group), which is commonly performed (although often incorrectly) for simplicity of communication. Similarly, it is not uncommon for investigators who evaluate many outcomes to report statistical significance with *P* values at the .04 and .02 levels (9,10). This approach, however, leads to a multiple comparisons problem (11,12). In this situation, one may falsely conclude a significant effect where there is none. In particular, use of a .05 cut-off value for significance theoretically guarantees that if there were 20 pairwise comparisons, there will by chance alone appear to be one with significance at the .05 level (20 × .05 = 1).

There are corrections for this problem (11,14). Some are useful for unordered groups, such as patient height versus sex, while others are applied to ordered groups (to evaluate a trend), such as patient height versus sex when stratified by age. It is also worth noting that there is a debate about which method to use, and some hold the view that this correction is overused (13). We focus our attention solely on unordered groups and the most commonly used correction, the Bonferroni method.

The Bonferroni correction is critical in adjusting the threshold for significance (14), which is equal to the desired *P* value (eg, .05, .01) divided by the number of outcome variables being examined. Consequently, when multiple statistical tests are conducted between the same variables, which would occur if multiple *t* tests were conducted for a comparison of age and bone density, the significance cut-off value is often adjusted to represent a more conservative estimate of statistical significance. One limitation of the Bonferroni correction is that by reducing the level of significance associated with each test, we have reduced the power of the test, thereby increasing the chance of incorrectly keeping the null hypothesis.

TABLE 4
Adjustment for Multiple Comparisons of Vertebral Density according to Age Group

	P Values		
Age Group Comparisons	Mean Difference	*t* Test	*t* Test with Bonferroni Adjustment
46–55 Years and 56–65 years	.014	.824	.99
46–55 Years and 66–75 years	.090	.021	.072

Note.—Values provided by statistical software, indicating a significant difference in density for the three groups.

Table 4 presents the results of *t* tests by using the same data presented in Table 3. Use of the common threshold of .05 would result in the conclusion that there is a significant difference in vertebral bone density between those 46–55 years of age and those 66–75 years of age (P = .021) and also between those 56–65 years of age and those 66–75 years of age (P = .006). However, compensating for multiple comparisons would reduce the threshold from .05 to .017 (.05/3). This results in only the comparison between those 56–65 years of age and those 66–75 years of age, which reaches statistical significance.

The far right column of Table 4 shows the exact probability of error with use of the Bonferroni adjustment provided by SPSS statistical software. The Bonferroni adjustment is a common option in statistical software when using ANOVA to determine if the group means are different from each other. The P value for each comparison is adjusted so it can be compared directly with the $P < .05$ cutoff. Again, by using $P < .05$ as the standard for significance, the results of the Bonferroni adjustment listed in Table 4 indicate that the only statistically significant difference is between those 56–65 years of age and those 66–75 years of age (P = .015).

In summary, it is often necessary to test hypotheses that relate a continuous outcome to an intervention—for example, tumor size versus treatment options. Depending on the number of groups (more than two) being analyzed, the ANOVA technique may be used to test for an effect, or a *t* test may be used if there are only two groups. A paired *t* test is used to examine two groups if the control population is linked on an individual basis, such as when a pre- and posttreatment comparison is made in the same patients. For radiologists, the comparisons may involve a new imaging technique, the use of contrast material, or a new MR imaging sequence.

Fundamental limitations in using these tests include the understanding that they generate an estimate of the probability that the differences observed would be due to random chance alone. This estimate is based not only on differences between means but also on sample variability and sample size. In addition, the assumption that the underlying population is distributed normally is not always appropriate—in which case, special nonparametric techniques are available. In a more common misapplication, the *t* test is used inappropriately to compare two groups of categoric or binary data. Finally, use of the tests presented in this article is limited to comparisons between two variables (such as patient age and bone density), which may often oversimplify much more complex relationships. More complex relationships are best analyzed with other techniques, such as multiple regression or ANOVA.

Acknowledgment: The authors thank Miriam Bowdre for secretarial support in preparation of the manuscript.

References

1. Mullner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. Ann Intern Med 2002; 136:122–126.
2. Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994; 309:248.
3. Zou KH, Fielding JR, Silverman SG, Tempany CM. Hypothesis testing I: proportions. Radiology 2003; 226:609–613.
4. Fenlon HM, Tello R, deCarvalho VLS, Yucel EK. Signal characteristics of focal liver lesions on double echo T2-weighted conventional spin echo MRI: observer performance versus quantitative measurements of T2 relaxation times. J Comput Assist Tomogr 2000; 24:204–211.
5. Rosner B. Fundamentals of biostatistics. 4th ed. Boston, Mass: Duxbury, 1995; 270–273.
6. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall/CRC, 1997.
7. Rosner B. Fundamentals of biostatistics. 4th ed. Boston, Mass: Duxbury, 1995; 570–575.
8. Bachman DM, Crewson PE. Comparison of central DXA with heel ultrasound and finger DXA for detection of osteoporosis. J Clin Densitom 2002; 5:131–141.
9. Sheafor DH, Keogan MT, Delong DM, Nelson RC. Dynamic helical CT of the abdomen: prospective comparison of pre- and postprandial contrast enhancement. Radiology 1998; 206:359–363.
10. Tello R, Seltzer SE. Hepatic contrast-enhanced CT: statistical design for prospective analysis. Radiology 1998; 209:879–881.
11. Gonen M, Panageas KS, Larson SM. Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient. Radiology 2001; 221:763–767.
12. Ware JH, Mosteller F, Delgado F, Donnelly C, Ingelfinger JA. P values. In: Bailar JC III, Mosteller F, eds. Medical uses of statistics. 2nd ed. Boston, Mass: NEJM Books, 1992; 181.
13. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998; 316:1236–1238.
14. Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford, England: Blackwell Scientific, 1994; 331.

Sample Size Estimation: How Many Individuals Should Be Studied?[1]

John Eng, MD

Index terms:
Radiology and radiologists, research
Statistical analysis

Published online
10.1148/radiol.2272012051
Radiology 2003; 227:309–313

The number of individuals to include in a research study, the sample size of the study, is an important consideration in the design of many clinical studies. This article reviews the basic factors that determine an appropriate sample size and provides methods for its calculation in some simple, yet common, cases. Sample size is closely tied to statistical power, which is the ability of a study to enable detection of a statistically significant difference when there truly is one. A trade-off exists between a feasible sample size and adequate statistical power. Strategies for reducing the necessary sample size while maintaining a reasonable power will also be discussed.

How many individuals will I need to study? This question is commonly asked by the clinical investigator and exposes one of many issues that are best settled before actually carrying out a study. Consultation with a statistician is worthwhile in addressing many issues of study design, but a statistician is not always readily available. Fortunately, many studies in radiology have simple designs for which determination of an appropriate *sample size*—the number of individuals that should be included for study—is relatively straightforward.

Superficial discussions of sample size determination are included in typical introductory biostatistics texts (1–3). The goal of this article is to augment these introductory discussions with additional practical material. First, the need for considering sample size will be reviewed. Second, the study design parameters affecting sample size will be identified. Third, formulae for calculating appropriate sample sizes for some common study designs will be defined. Finally, some advice will be offered on what to do if the calculated sample size is impracticably large. To assist the reader in performing the calculations described in this article and to encourage experimentation with them, a World Wide Web page has been developed that closely parallels the equations presented in this article. This page can be found at *www.rad.jhmi.edu/jeng/javarad/samplesize/*.

Even if a statistician is readily available, the investigator may find that a working knowledge of the factors affecting sample size will result in more fruitful communication with the statistician and in better research design. A working knowledge of these factors is also required to use one of the numerous Web pages (4–6) and computer programs (7–9) that have been developed for calculating appropriate sample sizes. It should be noted that Web pages for calculating sample size are typically limited for use in situations involving the well-known *parametric statistics*, which are those involving the calculation of summary means, proportions, or other parameters of an assumed underlying statistical distribution such as the normal, Student *t*, or binomial distributions. The calculation of sample size for nonparametric statistics such as the Wilcoxon rank sum test is performed by some computer programs (7,9).

IMPORTANCE OF SAMPLE SIZE

In a comparative research study, the means or proportions of some characteristic in two or more comparison groups are measured. A statistical test is then applied to determine whether or not there is a significant difference between the means or proportions observed in the comparison groups. We will first consider the comparative type of study.

[1] From the Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University, 600 N Wolfe St, Central Radiology Viewing Area, Rm 117, Baltimore, MD 21287.
Received December 17, 2001; revision requested January 29, 2002; revision received March 7; accepted March 13.
Address correspondence to the author (e-mail: *jeng@jhmi.edu*).

Sample size is important primarily because of its effect on statistical *power*. Statistical power is the probability that a statistical test will indicate a significant difference when there truly is one. Statistical power is analogous to the sensitivity of a diagnostic test (10), and one could mentally substitute the word "sensitivity" for the word "power" during statistical discussions.

In a study comparing two groups of individuals, the power (sensitivity) of a statistical test must be sufficient to enable detection of a statistically significant difference between the two groups if a difference is truly present. This issue becomes important if the study results were to demonstrate no statistically significant difference. If such a negative result were to occur, there would be two possible interpretations. The first interpretation is that the results of the statistical test are correct and that there truly is no statistically significant difference (a true-negative result). The second interpretation is that the results of the statistical test are erroneous and that there is actually an underlying difference, but the study was not powerful enough (sensitive enough) to find the difference, yielding a false-negative result. In statistical terminology, a false-negative result is known as a *type II error*. An adequate sample size gives a statistical test enough power (sensitivity) so that the first interpretation (that the results are true-negative) is much more plausible than the second interpretation (that a type II error occurred) in the event no statistically significant difference is found in the study.

It is well known that many published clinical research studies possess low statistical power owing to inadequate sample size or other design issues (11,12). One could argue that it is as wasteful and inappropriate to conduct a study with inadequate power as it is to obtain a diagnostic test of insufficient sensitivity to rule out a disease.

PARAMETERS THAT DETERMINE APPROPRIATE SAMPLE SIZE

An appropriate sample size generally depends on five study design parameters: minimum expected difference (also known as the effect size), estimated measurement variability, desired statistical power, significance criterion, and whether a one- or two-tailed statistical analysis is planned.

Minimum Expected Difference

This parameter is the smallest measured difference between comparison groups that the investigator would like the study to detect. As the minimum expected difference is made smaller, the sample size needed to detect statistical significance increases. The setting of this parameter is subjective and is based on clinical judgment and experience with the problem being investigated. For example, suppose a study is designed to compare a standard diagnostic procedure of 80% accuracy with a new procedure of unknown but potentially higher accuracy. It would probably be clinically unimportant if the new procedure were only 81% accurate, but suppose the investigator believes that it would be a clinically important improvement if the new procedure were 90% accurate. Therefore, the investigator would choose a minimum expected difference of 10% (0.10). The results of pilot studies or a literature review can also guide the selection of a reasonable minimum difference.

Estimated Measurement Variability

This parameter is represented by the expected SD in the measurements made within each comparison group. As statistical variability increases, the sample size needed to detect the minimum difference increases. Ideally, the estimated measurement variability should be determined on the basis of preliminary data collected from a similar study population. A review of the literature can also provide estimates of this parameter. If preliminary data are not available, this parameter may have to be estimated on the basis of subjective experience, or a range of values may be assumed. A separate estimate of measurement variability is not required when the measurement being compared is a proportion (in contrast to a mean), because the SD is mathematically derived from the proportion.

Statistical Power

This parameter is the power that is desired from the study. As power is increased, sample size increases. While high power is always desirable, there is an obvious trade-off with the number of individuals that can feasibly be studied, given the usually fixed amount of time and resources available to conduct a study. In randomized controlled trials, the statistical power is customarily set to a number greater than or equal to 0.80, with many clinical trial experts now advocating a power of 0.90.

Significance Criterion

This parameter is the maximum *P* value for which a difference is to be considered statistically significant. As the significance criterion is decreased (made more strict), the sample size needed to detect the minimum difference increases. The significance criterion is customarily set to .05.

One- or Two-tailed Statistical Analysis

In a few cases, it may be known before the study that any difference between comparison groups is possible in only one direction. In such cases, use of a one-tailed statistical analysis, which would require a smaller sample size for detection of the minimum difference than would a two-tailed analysis, may be considered. The sample size of a one-tailed design with a given significance criterion—for example, α—is equal to the sample size of a two-tailed design with a significance criterion of 2α, all other parameters being equal. Because of this simple relationship and because truly appropriate one-tailed analyses are rare, a two-tailed analysis is assumed in the remainder of this article.

SAMPLE SIZES FOR COMPARATIVE RESEARCH STUDIES

With knowledge of the design parameters detailed in the previous section, the calculation of an appropriate sample size simply involves selecting an appropriate equation. For a study comparing two means, the equation for sample size (13) is

$$N = \frac{4\sigma^2(z_{\mathrm{crit}} + z_{\mathrm{pwr}})^2}{D^2}, \qquad (1)$$

where *N* is the total sample size (the sum of the sizes of both comparison groups), σ is the assumed SD of each group (assumed to be equal for both groups), the z_{crit} value is that given in Table 1 for the desired significance criterion, the z_{pwr} value is that given in Table 2 for the desired statistical power, and *D* is the minimum expected difference between the two means. Both z_{crit} and z_{pwr} are cutoff points along the x axis of a standard normal probability distribution that demarcate probabilities matching the specified significance criterion and statistical power, respectively. The two groups that make up

TABLE 1
Standard Normal Deviate (z_{crit}) Corresponding to Selected Significance Criteria and CIs

Significance Criterion*	z_{crit} Value†
.01 (99)	2.576
.02 (98)	2.326
.05 (95)	1.960
.10 (90)	1.645

* Numbers in parentheses are the probabilities (expressed as a percentage) associated with the corresponding CIs. Confidence probability is the probability associated with the corresponding CI.

† A stricter (smaller) significance criterion is associated with a larger z_{crit} value. Values not shown in this table may be calculated in Excel version 97 (Microsoft, Redmond, Wash) by using the formula z_{crit} = NORMSINV(1−(*P*/2)), where *P* is the significance criterion.

TABLE 2
Standard Normal Deviate (z_{pwr}) Corresponding to Selected Statistical Powers

Statistical Power	z_{pwr} Value*
.80	0.842
.85	1.036
.90	1.282
.95	1.645

* A higher power is associated with a larger value for z_{pwr}. Values not shown in this table may be calculated in Excel version 97 (Microsoft, Redmond, Wash) by using the formula z_{pwr} = NORMSINV(*power*). For calculating power, the inverse formula is *power* = NORMSDIST(z_{pwr}), where z_{pwr} is calculated from Equation (1) or Equation (2) by solving for z_{pwr}.

N are assumed to be equal in number, and it is assumed that two-tailed statistical analysis will be used. Note that *N* depends only on the difference between the two means; it does not depend on the magnitude of either one.

As an example, suppose a study is proposed to compare a renovascular procedure versus medical therapy in lowering the systolic blood pressure of patients with hypertension secondary to renal artery stenosis. On the basis of results of preliminary studies, the investigators estimate that the vascular procedure may help lower blood pressure by 20 mm Hg, while medical therapy may help lower blood pressure by only 10 mm Hg. On the basis of their clinical judgment, the investigators might also argue that the vascular procedure would have to be twice as effective as medical therapy to justify the higher cost and discomfort of the vascular procedure. On the basis of results of preliminary studies, the SD for blood pressure lowering is estimated to be 15 mm Hg. According to the normal distribution, this SD indicates an expectation that 95% of the patients in either group will experience a blood pressure lowering within 30 mm Hg (2 SDs) of the mean. A significance criterion of .05 and power of 0.80 are chosen. With these assumptions, $D = 20 - 10 = 10$ mm Hg, $\sigma = 15$ mm Hg, $z_{crit} = 1.960$ (from Table 1), and $z_{pwr} = 0.842$ (from Table 2). Equation (1) yields a sample size of $N = 70.6$. Therefore, a total of 70 patients (rounding *N* to the nearest even number) should be enrolled in the study: 35 to undergo the vascular procedure and 35 to receive medical therapy.

For a study in which two proportions are compared with a χ^2 test or a *z* test, which is based on the normal approximation to the binomial distribution, the equation for sample size (14) is

$$N = 2 \cdot [z_{crit}\sqrt{2\bar{p}(1-\bar{p})} + z_{pwr}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\,]^2/D^2 \,, \qquad (2)$$

where p_1 and p_2 are pre-study estimates of the two proportions to be compared, $D = |p_1 - p_2|$ (ie, the minimum expected difference), $\bar{p} = (p_1 + p_2)/2$, and *N*, z_{crit}, and z_{pwr} are defined as they are for Equation (1). The two groups comprising *N* are assumed to be equal in number, and it is assumed that two-tailed statistical analysis will be used. Note that in this case, *N* depends not only on the difference between the two proportions but also on the magnitude of the proportions themselves. Therefore, Equation (2) requires the investigator to estimate p_1 and $p_{2,}$ as well as their difference, before performing the study. However, Equation (2) does not require an independent estimate of SD because it is calculated from p_1 and p_2 within the equation.

As an example, suppose a standard diagnostic procedure has an accuracy of 80% for the diagnosis of a certain disease. A study is proposed to evaluate a new diagnostic procedure that may have greater accuracy. On the basis of their experience, the investigators decide that the new procedure would have to be at least 90% accurate to be considered significantly better than the standard procedure. A significance criterion of .05 and a power of 0.90 are chosen. With these assumptions, $p_1 = 0.80$, $p_2 = 0.90$, $D = 0.10$, $\bar{p} = 0.85$, $z_{crit} = 1.960$, and $z_{pwr} = 0.842$. Equation (2) yields a sample size of $N = 398$. Therefore, a total of 398 patients should be enrolled: 199 to undergo the standard diagnostic procedure and 199 to undergo the new one.

SAMPLE SIZES FOR DESCRIPTIVE STUDIES

Not all research studies involve the comparison of two groups. The purpose of many studies is simply to describe, with means or proportions, one or more characteristics in one particular group. In these types of studies, known as descriptive studies, sample size is important because it affects how precise the observed means or proportions are expected to be. In the case of a descriptive study, the minimum expected difference reflects the difference between the upper and lower limit of an expected *confidence interval*, which is described with a percentage. For example, a 95% CI indicates the range in which 95% of results would fall if a study were to be repeated an infinite number of times, with each repetition including the number of individuals specified by the sample size.

In studies designed to estimate a mean, the equation for sample size (2,15) is

$$N = \frac{4\sigma^2(z_{crit})^2}{D^2} \,, \qquad (3)$$

where *N* is the sample size of the single study group, σ is the assumed SD for the group, the z_{crit} value is that given in Table 1, and *D* is the total width of the expected CI. Note that Equation (3) does not depend on statistical power because this concept only applies to statistical comparisons.

As an example, suppose a fetal sonographer wants to determine the mean fetal crown-rump length in a group of pregnancies. The sonographer would like the limits of the 95% confidence interval to be no more than 1 mm above or 1 mm below the mean crown-rump length of the group. From previous studies, it is known that the SD for the measurement is 3 mm. Based on these assumptions, $D = 2$ mm, $\sigma = 3$ mm, and $z_{crit} = 1.960$ (from Table 1). Equation (3) yields a sample size of $N = 35$. Therefore, 35 fetuses should be examined in the study.

In studies designed to measure a characteristic in terms of a proportion, the equation for sample size (2,15) is

$$N = \frac{4(z_{crit})^2 p(1-p)}{D^2} \,, \qquad (4)$$

where *p* is a pre-study estimate of the proportion to be measured, and *N*, z_{crit}, and *D* are defined as they are for Equation (3). Like Equation (2), Equation (4) depends not only on the width of the expected CI but also on the magnitude of the proportion itself. Also like Equation (2), Equation (4) does not require an independent estimate of SD because it is calculated from p within the equation.

As an example, suppose an investigator would like to determine the accuracy of a diagnostic test with a 95% CI of ±10%. Suppose that, on the basis of results of preliminary studies, the estimated accuracy is 80%. With these assumptions, $D = 0.20$, $p = 0.80$, and $z_{\text{crit}} = 1.960$. Equation (4) yields a sample size of $N = 61$. Therefore, 61 patients should be examined in the study.

MINIMIZING THE SAMPLE SIZE

Now that we understand how to calculate sample size, what if the sample size we calculate is too large to be feasibly studied? Browner et al (16) list a number of strategies for minimizing the sample size. These strategies are briefly discussed in the following paragraphs.

Use Continuous Measurements Instead of Categories

Because a radiologic diagnosis is often expressed in terms of a binary result, such as the presence or absence of a disease, it is natural to convert continuous measurements into categories. For example, the size of a lesion might be encoded as "small" or "large." For a sample of fixed size, the use of the actual measurement rather than the proportion in each category yields more power. This is because statistical tests that incorporate the use of continuous values are mathematically more powerful than those used for proportions, given the same sample size.

Use More Precise Measurements

For studies in which Equation (1) or Equation (2) applies, any way to increase the precision (decrease the variability) of the measurement process should be sought. For some types of research, precision can be increased by simply repeating the measurement. More complex equations are necessary for studies involving repeated measurements in the same individuals (17), but the basic principles are similar.

Use Paired Measurements

Statistical tests like the paired t test are mathematically more powerful for a given sample size than are unpaired tests because in paired tests, each measurement is matched with its own control. For example, instead of comparing the average lesion size in a group of treated patients with that in a control group, measuring the change in lesion size in each patient after treatment allows each patient to serve as his or her own control and yields more statistical power. Equation (1) can still be used in this case. D represents the expected change in the measurement, and σ is the expected SD of this change. The additional power and reduction in sample size are due to the SD being smaller for changes within individuals than for overall differences between groups of individuals.

Use Unequal Group Sizes

Equations (1) and (2) involve the assumption that the comparison groups are equal in size. Although it is statistically most efficient if the two groups are equal in size, benefit is still gained by studying more individuals, even if the additional individuals all belong to one of the groups. For example, it may be feasible to recruit additional individuals into the control group even if it is difficult to recruit more individuals into the noncontrol group. More complex equations are necessary for calculating sample sizes when comparing means (13) and proportions (18) of unequal group sizes.

Expand the Minimum Expected Difference

Perhaps the minimum expected difference that has been specified is unnecessarily small, and a larger expected difference could be justified, especially if the planned study is a preliminary one. The results of a preliminary study could be used to justify a more ambitious follow-up study of a larger number of individuals and a smaller minimum difference.

DISCUSSION

The formulation of Equations (1–4) involves two statistical assumptions which should be kept in mind when these equations are applied to a particular study. First, it is assumed that the selection of individuals is random and unbiased. The decision to include an individual in the study cannot depend on whether or not that individual has the characteristic or outcome being studied. Second, in studies in which a mean is calculated from measurements of individuals, the measurements are assumed to be normally distributed. Both of these assumptions are required not only by the sample size calculation method, but also by the statistical tests themselves (such as the t test). The situations in which Equations (1–4) are appropriate all involve parametric statistics. Different methods for determining sample size are required for nonparametric statistics such as the Wilcoxon rank sum test.

Equations for calculating sample size, such as Equations (1) and (2), also provide a method for determining statistical power corresponding to a given sample size. To calculate power, solve for z_{pwr} in the equation corresponding to the design of the study. The power can be then determined by referring to Table 2. In this way, an "observed power" can be calculated after a study has been completed, where the observed difference is used in place of the minimum expected difference. This calculation is known as retrospective power analysis and is sometimes used to aid in the interpretation of the statistical results of a study. However, retrospective power analysis is controversial because it can be shown that observed power is completely determined by the P value and therefore cannot add any additional information to its interpretation (19). Power calculations are most appropriate when they incorporate a minimum difference that is stated prospectively.

The accuracy of sample size calculations obviously depends on the accuracy of the estimates of the parameters used in the calculations. Therefore, these calculations should always be considered estimates of an absolute minimum. It is usually prudent for the investigator to plan to include more than the minimum number of individuals in a study to compensate for loss during follow-up or other causes of attrition.

Sample size is best considered early in the planning of a study, when modifications in study design can still be made. Attention to sample size will hopefully result in a more meaningful study whose results will eventually receive a high priority for publication.

References

1. Pagano M, Gauvreau K. Principles of biostatistics. 2nd ed. Pacific Grove, Calif: Duxbury, 2000; 246–249, 330–331.
2. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7th ed. New York, NY: Wiley, 1999; 180–185, 268–270.
3. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall, 1991.
4. Bond J. Power calculator. Available at: *http://calculators.stat.ucla.edu/powercalc/*. Accessed March 11, 2003.

5. Uitenbroek DG. Sample size: SISA—simple interactive statistical analysis. Available at: *http://home.clara.net/sisa/samsize.htm*. Accessed March 3, 2003.
6. Lenth R. Java applets for power and sample size. Available at: *www.stat.uiowa.edu/~rlenth/Power/index.html*. Accessed March 3, 2003.
7. NCSS Statistical Software. PASS 2002. Available at: *www.ncss.com/pass.html*. Accessed March 3, 2003.
8. SPSS. SamplePower. Available at: *www.spss.com/SPSSBI/SamplePower/*. Accessed March 3, 2003.
9. Statistical Solutions. nQuery Advisor. Available at: *www.statsolusa.com/nquery/nquery.htm*. Accessed March 3, 2003.
10. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987; 257:2459–2463.
11. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994; 272:122–124.
12. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. N Engl J Med 1978; 299:690–694.
13. Rosner B. Fundamentals of biostatistics. 5th ed. Pacific Grove, Calif: Duxbury, 2000; 308.
14. Feinstein AR. Principles of medical statistics. Boca Raton, Fla: CRC, 2002; 503.
15. Snedecor GW, Cochran WG. Statistical methods. 8th ed. Ames, Iowa: Iowa State University Press, 1989; 52, 439.
16. Browner WS, Newman TB, Cummings SR, Hulley SB. Estimating sample size and power. In: Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Designing clinical research: an epidemiologic approach. 2nd ed. Philadelphia, Pa: Lippincott Williams & Wilkins, 2001; 65–84.
17. Frison L, Pocock S. Repeated measurements in clinical trials: analysis using mean summary statistics and its implications for design. Stat Med 1992; 11:1685–1704.
18. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY: Wiley, 1981; 45.
19. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 2001; 55:19–24.

Correlation and Simple Linear Regression[1]

Kelly H. Zou, PhD
Kemal Tuncali, MD
Stuart G. Silverman, MD

Index terms:
Data analysis
Statistical analysis

Published online
10.1148/radiol.2273011499
Radiology 2003; 227:617–628

In this tutorial article, the concepts of correlation and regression are reviewed and demonstrated. The authors review and compare two correlation coefficients, the Pearson correlation coefficient and the Spearman ρ, for measuring linear and nonlinear relationships between two continuous variables. In the case of measuring the linear relationship between a predictor and an outcome variable, simple linear regression analysis is conducted. These statistical concepts are illustrated by using a data set from published literature to assess a computed tomography–guided interventional technique. These statistical methods are important for exploring the relationships between variables and can be applied to many radiologic studies.
© RSNA, 2003

Results of clinical studies frequently yield data that are dependent of each other (eg, total procedure time versus the dose in computed tomographic [CT] fluoroscopy, signal-to-noise ratio versus number of signals acquired during magnetic resonance imaging, and cigarette smoking versus lung cancer). The statistical concepts correlation and regression, which are used to evaluate the relationship between two continuous variables, are reviewed and demonstrated in this article.

Analyses between two variables may focus on *(a)* any association between the variables, *(b)* the value of one variable in predicting the other, and *(c)* the amount of agreement. Agreement will be discussed in a future article. Regression analysis focuses on the form of the relationship between variables, while the objective of correlation analysis is to gain insight into the strength of the relationship (1,2). Note that these two techniques are used to investigate relationships between continuous variables, whereas the χ^2 test is an example of a test for association between categorical variables. Continuous variables, such as procedure time, patient age, and number of lesions, have no gaps on the measurement scale. In contrast, categorical variables, such as patient sex and tissue classification based on segmentation, have gaps in their possible values. These two types of variables and the assumptions about their measurement scales were reviewed and distinguished in an article by Applegate and Crewson (3) published earlier in this Statistical Concepts Series in *Radiology*.

Specifically, the topics covered herein include two commonly used correlation coefficients, the Pearson correlation coefficient (4,5) and the Spearman ρ (6–10) for measuring linear and nonlinear relationship, respectively, between two continuous variables. Correlation analysis is often conducted in a retrospective or observational study. In a clinical trial, on the other hand, the investigator may also wish to manipulate the values of one variable and assess the changes in values of another variable. To evaluate the relative impact of the predictor variable on the particular outcome, simple regression analysis is preferred. We illustrate these statistical concepts with existing data to assess patient skin dose based on total procedure time by using a quick-check method in CT fluoroscopy–guided abdominal interventions (11).

These statistical methods are useful tools for assessing the relationships between continuous variables collected from a clinical study. However, it is also important to distinguish these statistical methods. While they are similar mathematically, their purposes are different. Correlation analysis is generally overused. It is often interpreted incorrectly (to establish "causation") and should be reserved for generating hypotheses rather than for testing them. On the other hand, regression modeling is a more useful statistical technique that allows us to assess the strength of the relationships in the data and the uncertainty in the model by using confidence intervals (12,13).

[1] From the Department of Radiology, Brigham and Women's Hospital (K.H.Z., K.T., S.G.S.) and Department of Health Care Policy (K.H.Z.), Harvard Medical School, 180 Longwood Ave, Boston, MA 02115. Received September 10, 2001; revision requested October 31; revision received December 26; accepted January 21, 2002. **Address correspondence to** K.H.Z. (e-mail: *zou@bwh.harvard.edu*).

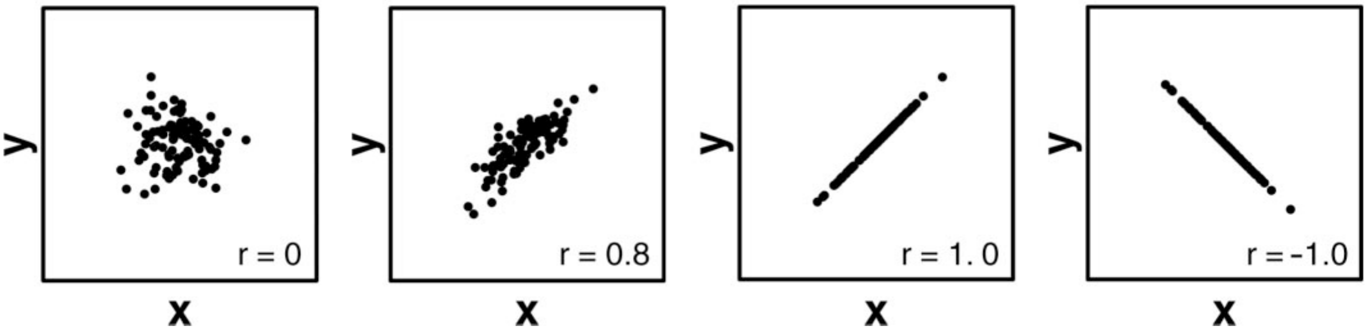


Figure 1. Scatterplots of four sets of data generated by means of the following Pearson correlation coefficients (from left to right): $r = 0$ (uncorrelated data), $r = 0.8$ (strongly positively correlated), $r = 1.0$ (perfectly positively correlated), and $r = -1$ (perfectly negatively correlated).

CORRELATION

The purpose of correlation analysis is to measure and interpret the strength of a linear or nonlinear (eg, exponential, polynomial, and logistic) relationship between two continuous variables. When conducting correlation analysis, we use the term *association* to mean "linear association" (1,2). Herein, we focus on the Pearson and Spearman ρ correlation coefficients. Both correlation coefficients take on values between −1 and +1, ranging from being negatively correlated (−1) to uncorrelated (0) to positively correlated (+1). The sign of the correlation coefficient (ie, positive or negative) defines the direction of the relationship. The absolute value indicates the strength of the correlation (Table 1, Fig 1). We elaborate on two correlation coefficients, linear (eg, Pearson) and rank (eg, Spearman), that are commonly used for measuring linear and general relationships between two variables.

Linear Correlation

The Pearson correlation coefficient is also known as the sample correlation coefficient *(r)*, product-moment correlation coefficient, or coefficient of correlation (14). It was introduced by Galton in 1877 (15,16) and developed later by Pearson (17). It measures the linear relationship between two random variables. For example, when the value of the predictor is manipulated (increased or decreased) by a fixed amount, the outcome variable changes proportionally (linearly). A linear correlation coefficient can be computed by means of the data and their sample means (Appendix A). When a scientific study is planned, the required sample size may be computed on the basis of a certain hypothesized value with the desired statistical power at a specified level of significance (Appendix B) (18).

Rank Correlation

The Spearman ρ is the sample correlation coefficient *(r_s)* of the ranks (the relative order) based on continuous data (19,20). It was first introduced by Spearman in 1904 (6). The Spearman ρ is used to measure the monotonic relationship between two variables (ie, whether one variable tends to take either a larger or smaller value, though not necessarily linearly) by increasing the value of the other variable.

Linear versus Rank Correlation Coefficients

The Pearson correlation coefficient necessitates use of interval or continuous measurement scales of the measured outcome in the study population. In contrast, rank correlations also work well with ordinal rating data, and continuous data are reduced to their ranks (Appendix C) (20,21). The rank procedure will also be illustrated briefly with our example data. The smallest value in the sample has rank 1, and the largest has the highest rank. In general, rank correlations are not easily influenced by the presence of skewed data or data that are highly variable.

Statistical Hypothesis Tests for a Correlation Coefficient

The null hypothesis states that the underlying linear correlation has a hypothesized value, ρ_0. The one-sided alternative hypothesis is that the underlying value exceeds (or is less than) ρ_0. When the sample size *(n)* of the paired data is large ($n \geq 30$ for each variable), the standard error *(s)* of the linear correlation *(r)* is approximately $s(r) = (1 - r^2)/\sqrt{n}$. The test statistic value $(r - \rho_0)/s(r)$ may be computed by means of the z test (22). If the *P* value is below .05, the null hypothesis is rejected. The *P* value based on the Spearman ρ can be found in the literature (20,21).

TABLE 1
Interpretation of Correlation Coefficient

Correlation Coefficient Value	Direction and Strength of Correlation
−1.0	Perfectly negative
−0.8	Strongly negative
−0.5	Moderately negative
−0.2	Weakly negative
0.0	No association
+0.2	Weakly positive
+0.5	Moderately positive
+0.8	Strongly positive
+1.0	Perfectly positive

Note.—The sign of the correlation coefficient (ie, positive or negative) defines the direction of the relationship. The absolute value indicates the strength of the correlation.

Limitations and Precautions

It is worth noting that even if two variables (eg, cigarette smoking and lung cancer) are highly correlated, it is not sufficient proof of causation. One variable may cause the other or vice versa, or a third factor is involved, or a rare event may have occurred. To conclude causation, the causal variables must precede the variable it causes, and several conditions must be met (eg, reversibility, strength, and exposure response on the basis of the Bradford-Hill criteria or the Rubin causal model) (23–26).

SIMPLE LINEAR REGRESSION

The purpose of simple regression analysis is to evaluate the relative impact of a predictor variable on a particular outcome. This is different from a correlation analysis, where the purpose is to examine the strength and direction of the rela-

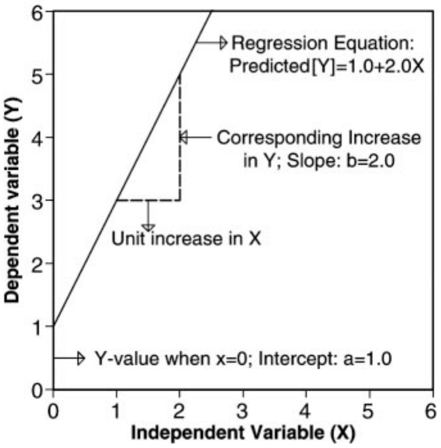


Figure 2. Simple linear regression model shows that the expectation of the dependent variable *Y* is linear in the independent variable *X*, with an intercept $a = 1.0$ and a slope $b = 2.0$.

tionship between two random variables. In this article, we deal with only linear regression of one continuous variable on another continuous variable with no gaps on each measurement scale (3). There are other types of regression (eg, multiple linear, logistic, and ordinal) analyses, which will be provided in a future article in this Statistical Concepts Series in *Radiology*.

A simple regression model contains only one independent (explanatory) variable, X_i, for $i = 1, \ldots, n$ subjects, and is linear with respect to both the regression parameters and the dependent variable. The corresponding dependent (outcome) variable is labeled. The model is expressed as

$$Y_i = a + bX_i + e_i, \qquad (1)$$

where the regression parameter *a* is the intercept (on the y axis), and the regression parameter *b* is the slope of the regression line (Fig 2). The random error term e_i is assumed to be uncorrelated, with a mean of 0 and constant variance. For convenience in inference and improved efficiency in estimation (27), analyses often incur an additional assumption that the errors are distributed normally. Transformation of the data to achieve normality may be applied (28,29). Thus, the word *line* (linear, independent, normal, equal variance) summarizes these requirements.

Typical steps for regression model analysis are the following: *(a)* determine if the assumptions underlying a normal relationship are met in the data, *(b)* obtain the equation that best fits the data, *(c)* evaluate the equation to determine the strength of the relationship for prediction and estimation, and *(d)* assess whether the data fit these criteria before the equation is applied for prediction and estimation.

Least Squares Method

The main goal of linear regression is to fit a straight line through the data that predicts *Y* based on *X*. To estimate the intercept and slope regression parameters that determine this line, the least squares method is commonly used. It is not necessary for the errors to have a normal distribution, although the regression analysis is more efficient with this assumption (27). With this regression method, a set of regression parameters are found such that the sum of squared residuals (ie, the differences between the observed values of the outcome variable and the fitted values) are minimized (14). The fitted *y* value is then computed as a function of the given *x* value and the estimated intercept and slope regression parameter (Appendix D). For example, in Equation (1), once the estimates of *a* and *b* are obtained from the regression analysis, the predicted y value at any given *x* value is calculated as $a + bx$.

Coefficient of Determination, R^2

It is meaningful to interpret the value of the Pearson correlation coefficient *r* by squaring it; hence, the term R-square (R^2) or coefficient of determination. This measure (with a range of 0–1) is the fraction of the variability in *Y* that can be explained by the variability in *X* through their linear relationship, or vice versa. That is, $R^2 = \mathrm{SS}_{\mathrm{regression}}/\mathrm{SS}_{\mathrm{total}}$, where SS stands for the sum of squares. Note that R^2 is calculated only on the basis of the Pearson correlation coefficient in the linear regression analysis. Thus, it is not appropriate to compute R^2 on the basis of rank correlation coefficients such as the Spearman ρ.

Statistical Hypothesis Tests

There are several hypotheses in the context of regression analysis, for example, to test if the slope of the regression line is $b = 0$ (hypothesis, there is no linear association between *Y* and *X*). One may also test whether intercept *a* takes on a certain value. The significance of the effects of the intercept and slope may also be computed by means of a Student *t* statistic introduced earlier in this Statistical Concepts Series in *Radiology* (30).

Limitations and Precautions

The following understandings should be considered when regression analysis is performed. *(a)* To understand whether the assumptions have been met, determine the magnitude of the gap between the data and the assumptions of the model. *(b)* No matter how strong a relationship is demonstrated with regression analysis, it should not be interpreted as causation (as in the correlation analysis). *(c)* The regression should not be used to predict or estimate outside the range of values of the independent variable of the sample (eg, extrapolation of radiation cancer risk from the Hiroshima data to that of diagnostic radiologic tests).

AN EXAMPLE: DOSE VERSUS TOTAL PROCEDURE TIME IN CT FLUOROSCOPY

We applied these statistical methods to help assess the benefit of the use of CT fluoroscopy to guide interventions in the abdomen (11). During CT fluoroscopy–guided interventions, one might postulate that the radiation dose received by a patient is related to (or correlated with) the total procedure time, because the more difficult the procedure is, the more CT fluoroscopic scanning is required, which means a longer procedure time. The rationale was to assess whether radiation dose could be estimated by simply measuring the total CT fluoroscopic procedure time, with the null hypothesis that the slope of the regression line is 0.

Earlier, we discussed two methods to target lesions with CT fluoroscopy. In one method, continuous CT scanning is used during needle placement. In the other method, short CT scanning is used to image the needle after it is placed. The latter method, the so-called quick-check method, has been adopted almost exclusively at our institution. Now, we demonstrate correlation and regression analyses based on a subset of the interventional procedures ($n = 19$). With the quick-check method, we examine the relationship between total procedure time (in minutes) and dose (in rads) on a natural log scale. We also examine the marginal ranks of the *x* (log of total time) and *y* (log of dose) components (Table 2). For convenience, the *x* data are given in ascending order.

In Table 2, each set of rank data is derived by first placing the 19 observations in each sample in ascending order and then assigning ranks 1–19. Ties are broken by means of averaging the respective adjacent ranks. Finally, the ranks are identified for the observations of each of the paired *x* and *y* samples.

The natural log (ln) transformation of the total time is used to make the data appear normal, for more efficient analysis (Appendix D), with normality verified statistically (31). However, normality is not necessary in the subsequent regression analysis. We created a scatterplot of the data, with the log of dose (ln[rad]) on the x axis and the log of total time (ln[minutes]) on the y axis (Fig 3).

For illustration purposes, we will conduct both correlation and regression analyses; however, the choice of analysis depends on the aim of research. For example, if the investigators wish to assess whether there is a relationship between time and dose, then correlation analysis is appropriate. In comparison, if the investigators wish to evaluate the impact of the total time on the resulting dose, then regression analysis is preferred.

TABLE 2
Total Procedure Time and Dose of CT Fluoroscopy–guided Procedures, by Means of the Quick-Check Method

Subject No.	*x* Data: Log Time (ln[min])	Ranks of *x* Data	*y* Data: Log Dose (ln[rad])	Ranks of *y* Data
1	3.61	1	1.48	2
2	3.87	2	1.24	1
3	3.95	3	2.08	5.5
4	4.04	4	1.70	3
5	4.06	5	2.08	5.5
6	4.11	6	2.94	10
7	4.19	7	2.24	7
8	4.20	8	1.85	4
9	4.32	9.5	2.84	9
10	4.32	9.5	3.93	16
11	4.42	11.5	3.03	11
12	4.42	11.5	3.23	13
13	4.45	13	3.87	15
14	4.50	14	3.55	14
15	4.52	15	2.81	8
16	4.57	16	4.07	17
17	4.58	17	4.44	19
18	4.61	18	3.16	12
19	4.74	19	4.19	18

Source.—Reference 11.

Note.—Paired *x* and *y* data are sorted according to the *x* component; therefore, the log of the total procedure time and the log of the corresponding rank have an increasing order. When ties are present in the data, the average of their adjacent ranks is used. Pearson correlation coefficient between log time and log dose, $r = 0.85$; Spearman $\rho = 0.84$.

Correlations

To compute the Spearman ρ with a Pearson correlation coefficient of $r = 0.85$, the marginal ranks of time and dose were derived separately; consequently, $r_s = 0.84$. Both correlation coefficients confirm that the log of total time and the log of dose are correlated strongly and positively.

Regression

We first conducted a simple linear regression analysis of the data on a log scale ($n = 19$); results are shown in Table 3. The value calculated for R^2 was 0.73, which suggests that 73% of the variability of the data could be explained by the linear regression.

The regression line, expressed in the form given in Equation (1), is $Y = -9.28 + 2.83X$, where the predictor variable X represents the log of total time, and the outcome variable Y represents the log of dose. The estimated regression parameters are $a = -9.28$ (intercept) and $b = 2.83$ (slope) (Fig 4). This regression line can be interpreted as follows: At $X = 0$, the value of Y is -9.28. For every one-unit increase in X, the value of Y will increase on average by 2.83. Effects of both the intercept and slope are statistically significant ($P < .005$) (Excel; Microsoft, Redmond, Wash); therefore, the null hypothesis (H_0, the dose remains constant as the total procedure time increases) is rejected. Thus, we confirm the alternative hypothesis (H_1, the dose increases in the total procedure time).

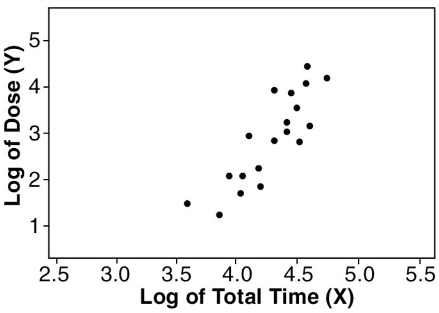


Figure 3. Scatterplot of the log of dose (y axis) versus the log of total time (x axis). Each point in the scatterplot represents the values of two variables for a given observation.

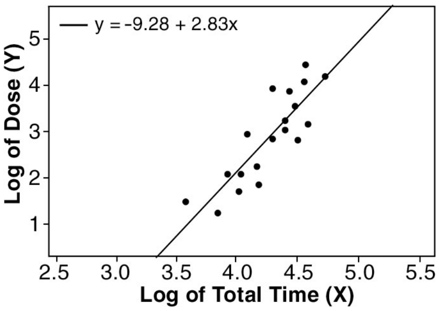


Figure 4. Scatterplot of the log of dose (y axis) versus the log of total time (x axis). The regression line has the intercept $a = -9.28$ and slope $b = 2.83$. We conclude that there is a possible association between the radiation dose and the total time of the procedure.

The regression line may be used to give predicted values of Y. For example, if in a future CT fluoroscopy procedure, the log total time is specified at $x = 4$ (translated to $e^4 = 55$ minutes, approximately), then the log dose that is to be applied is approximately $y = -9.28 + 2.83 \times 4 = 2.04$ (translated to $e^{2.04} = 7.69$ rad). On the other hand, if the log total time is specified at $x = 4.5$ (translated to $e^{4.5} = 90$ minutes, approximately), then the log dose that is to be applied is approximately $y = -9.28 + 2.83 \times 4.5 = 3.46$ (translated to $e^{3.46} = 31.82$ rad). Such prediction can be useful for future clinical practice.

TABLE 3
Results based on Correlation and Regression Analysis for Example Data

Regression Statistic	Numerical Result
Correlation coefficient *r*	0.85
R-square (R^2)	0.73
Regression parameter	
Intercept	−9.28
Slope	2.83

Source.—Reference 11.

SUMMARY AND REMARKS

Two important statistical concepts, correlation and regression, which are used commonly in radiology research, are reviewed and demonstrated herein. Additional sources of information and electronic textbooks on statistical analysis methods found on the World Wide Web are listed in Appendix E. A glossary of the statistical terms used in this article is presented in Appendix F.

When correlation analysis is conducted to measure the association between two random variables, either the Pearson linear correlation coefficient or the Spearman rank correlation coefficient ρ may be adopted. The former coefficient is used to measure the linear relationship but is not recommended for use with skewed data or data with extremely large or small values (often called the outliers). In contrast, the latter coefficient is used to measures a general association, and it is recommended for use with data that are skewed or that have outliers.

When simple regression analysis is conducted to assess the linear relationship of a dependent variable as a function of the independent variable, caution must be used when determining which of the two variables is viewed as the independent variable that makes sense clinically. A useful graphical aid is a scatterplot. Once the regression line is obtained, caution should also be used to avoid prediction of a *y* value for any value of *x* that is outside the range of the data. Finally, correlation and regression analyses do not infer causality, and more rigorous analyses are required if causal inference is to be made (23–26).

APPENDIX A

Formula for computing the Pearson correlation coefficient, *r*: The formula for computing *r* between bivariate data, X_i and Y_i values ($i = 1,\ldots,n$) is

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}},$$

where $\overline{X}$ and $\overline{Y}$ are the sample means of the X_i and Y_i values, respectively.

The Pearson correlation coefficient may be computed by means of a computer-based statistics program (Excel; Microsoft) by using the option "Correlation" under the option "Data Analysis Tools". Alternatively, it may also be computed by means of a built-in software function "Cor" (Insightful; MathSoft, Seattle, Wash [MathSoft S-Plus 4 guide to statistics, 1997; 89–96]. Available at: *www.insightful.com*) or with a free software program (R Software. Available at: *lib.stat.cmu.edu/R)*.

APPENDIX B

Total sample size based on the Pearson correlation coefficient: Specify *r* = expected correlation coefficient, $C = 0.5 \times \ln[(1 + r)/(1 - r)]$, *N* = total number of subjects required, α = type I error (ie, significance level, typically fixed at 0.05), β = type II error (ie, 1 minus statistical power, typically fixed at 0.10). Then $N = [(Z_\alpha + Z_\beta)/C]^2 + 3$, where Z_α is the inverse of the cumulative probability of a standard normal distribution with the tail probability of α. Similarly, Z_β is the inverse of the cumulative probability of a standard normal distribution with the tail probability of β. Consequently, compute the smallest integer, *n*, such that $n \geq N$, as the required sample size.

For example, an investigator wishes to conduct a clinical trial of a paired design based on a one-tailed hypothesis test of the correlation coefficient. The null hypothesis is that the correlation between two variables is $r = 0.60$ (ie, $C = 0.693$) in the population of interest. The alternative hypothesis is that the correlation is $r > 0.60$. Type I error is fixed to be 0.05 (ie, $Z_\alpha = 1.645$), while type II error is fixed to be 0.10 (ie, $Z_\beta = 1.282$). Thus, the required sample size is $N = 21$ subjects. A sample size table may also be found in reference 18.

APPENDIX C

Formula for computing Spearman ρ and Pearson r_s: Replace bivariate data, X_i and Y_i ($i = 1,\ldots,n$), by their respective ranks $R_i = \text{rank}(X_i)$ and $S_i = \text{rank}(Y_i)$. Rank correlation coefficient, r_s, is defined as the Pearson correlation coefficient between the R_i and S_i values, which can be computed by means of the formula given in Appendix A. An alternative direct formula was given by Hettmansperger (19).

The Spearman ρ may also be computed by first reducing the continuous data to their marginal ranks by using the "rank and percentile" option with Data Analysis Tools (Excel; Microsoft) or the "rank" function (Insightful; MathSoft) or the free software. Both software programs correctly rank the data in ascending order. However, the rank and percentile option in Excel ranks the data in descending order (the largest is 1). Therefore, to compute the correct ranks, one may first multiply all of the data by −1 and then apply the rank function. Excel also gives integer ranks in the presence of ties compared with the methods that yield possible noninteger ranks, as described in the standard statistics literature (19).

Subsequently, the sample correlation coefficient is computed on the basis of the ranks of the two marginal data by using the Correlation option in Data Analysis Tools (Excel; Microsoft) or by using the Cor function (Insightful; MathSoft) or the free software.

APPENDIX D

Simple regression analysis: Regression analysis may be performed by using the "Regression" option with Data Analysis Tools (Excel; Microsoft). This regression analysis tool yields the sample correlation R^2; estimates of the regression parameters, along with their statistical significance on the basis of the Student *t* test; residuals; and standardized residuals. Scatter, line fit, and residual plots may also be created. Alternatively, the analyses can be performed by using the function "lsfit" (Insightful; MathSoft) or the free software.

With either program, one may choose to transform the data or exclude outliers before conducting a simple regression analysis. A commonly used variance-stabilizing transformation is the natural log function (ln) applied to one or both variables. Other transformation (eg, Box-Cox transformation) and weighting methods in regression analysis may also be used (28,29).

APPENDIX E

Uniform resource locator, or URL, links to electronic statistics textbooks: *www.davidmlane.com/hyperstat/index.html*, *www.statsoft.com/textbook/stathome.html*, *www.ruf.rice.edu/~lane/rvls.html*, *www.bmj.com/collections/statsbk/index.shtml*, *espse.ed.psu.edu/statistics/investigating.htm*.

APPENDIX F

Glossary of statistical terms:

Bivariate data.—Measurements obtained on more than one variable for the same unit or subject.

Correlation coefficient.—A statistic between −1 and 1 that measures the association between two variables.

Intercept.—The constant *a* in the regression equation, which is the value for *y* when $x = 0$.

Least squares method.—The regression line that is the best fit to the data for which the sum of the squared residuals is minimized.

Outlier.—An extreme observation far away from the bulk of the data, often caused by faulty measuring equipment or recording error.

Pearson correlation coefficient.—Sample correlation coefficient for measuring the linear relationship between two variables.

R^2.—The square of the Pearson correlation coefficient *r,* which is the fraction of the variability in *Y* that can be explained by the variability in *X* through their linear relationship or vice versa.

Rank.—The relative ordering of the measurements in a variable, which can be noninteger numbers in the presence of ties.

Residual.—The difference between the observed values of the outcome variable and the fitted values based on a linear regression analysis.

Scatterplot.—A plot of the observed bivariate outcome variable (y axis) against its predictor variable (x axis), with a dot for each pair of bivariate observations.

Simple linear regression analysis.—A linear regression analysis with one predictor and one outcome variable.

Skewed data.—A distribution is skewed if there are more extreme data on one side of the mean. Otherwise, the distribution is symmetric.

Slope.—The constant *b* in the regression equation, which is the change in *y* that corresponds to a one-unit increase (or decrease) in *x*.

Spearman ρ.—A rank correlation coefficient for measuring the monotone relationship between two variables.

Acknowledgments: We thank Kimberly E. Applegate, MD, MS, and Philip E. Crewson, PhD, co-editors of this Statistical Concepts Series in *Radiology* for their constructive comments on earlier versions of this article.

References

1. Krzanowsk WJ. Principles of multivariate analysis: a user's perspective. Oxford, England: Clarendon, 1988; 405–432.
2. Rodriguez RN. Correlation. In: Kotz S, Johnson NL, eds. Encyclopedia of statistical sciences. New York, NY: Wiley, 1982; 193–204.
3. Applegate KE, Crewson PE. An introduction to biostatistics. Radiology 2002; 225: 318–322.
4. Goldman RN, Weinberg JS. Statistics: an introduction. Upper Saddle River, NJ: Prentice Hall, 1985; 72–98.
5. Freund JE. Mathematical statistics. 5th ed. Upper Saddle River, NJ: Prentice Hall, 1992; 494–546.
6. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904; 15:72–101.
7. Fieller EC, Hartley HO, Pearson ES. Tests for rank correlation coefficient. I. Biometrika 1957; 44:470–481.
8. Fieller EC, Pearson ES. Tests for rank correlation coefficients. II. Biometrika 1961; 48:29–40.
9. Kruskal WH. Ordinal measurement of association. J Am Stat Assoc 1958; 53:814–861.
10. David FN, Mallows CL. The variance of Spearman's rho in normal samples. Biometrika 1961; 48:19–28.
11. Silverman SG, Tuncali K, Adams DF, Nawfel RD, Zou KH, Judy PF. CT fluoroscopy-guided abdominal interventions: techniques, results, and radiation exposure. Radiology 1999; 212:673–681.
12. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7th ed. New York, NY: Wiley, 1999.
13. Altman DG. Practical statistics for medical research. Boca Raton, Fla: CRC, 1990.
14. Neter J, Wasserman W, Kutner MH. Applied linear models: regression, analysis of variance, and experimental designs. 3rd ed. Homewood, Ill: Irwin, 1990; 38–44, 62–104.
15. Galton F. Typical laws of heredity. Proc R Inst Great Britain 1877; 8:282–301.
16. Galton F. Correlations and their measurements, chiefly from anthropometric data. Proc R Soc London 1888; 45:219–247.
17. Pearson K. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Phil Trans R Soc Lond Series A 1896; 187:253–318.
18. Hulley SB, Cummings SR. Designing clinical research: an epidemiological approach. Baltimore, Md: Williams & Wilkins, 1988; appendix 13.C.
19. Hettmansperger TP. Statistical inference based on ranks. Malabar, Fla: Krieger, 1991; 200–205.
20. Kendall M, Gibbons JD. Rank correlation methods. 5th ed. New York, NY: Oxford University Press, 1990; 8–10.
21. Zou KH, Hall WJ. On estimating a transformation correlation coefficient. J Appl Stat 2002; 29:745–760.
22. Fisher RA. Frequency distributions of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 1915; 10:507–521.
23. Duncan OD. Path analysis: sociological examples. In: Blalock HM Jr, ed. Causal models in the social sciences. Chicago, Ill: Alpine-Atherton, 1971; 115–138.
24. Rubin DB. Estimating casual effects of treatments in randomized and nonrandomized studies. J Ed Psych 1974; 66:688–701.
25. Holland P. Statistics and causal inference. J Am Stat Assoc 1986; 81:945–970.
26. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996; 91:444–455.
27. Seber GAF. Linear regression analysis. New York, NY: Wiley, 1997; 48–51.
28. Carroll RJ, Ruppert D. Transformation and weighting in regression. New York, NY: Chapman & Hall, 1988; 2–61.
29. Box GEP, Cox DR. An analysis of transformation. J R Stat Soc Series B 1964; 42: 71–78.
30. Tello R, Crewson PE. Hypothesis testing II: means. Radiology 2003; 227:1–4.
31. Mudholkar GS, McDermott M, Scrivastava DK. A test of p-variate normality. Biometrika 1992; 79:850–854.

Curtis P. Langlotz, MD, PhD

Index terms:
Diagnostic radiology
Statistical analysis

Published online
10.1148/radiol.2281011106
Radiology 2003; 228:3–9

Abbreviation:
ROC = receiver operating characteristic

Fundamental Measures of Diagnostic Examination Performance: Usefulness for Clinical Decision Making and Research[1]

Measures of diagnostic accuracy, such as sensitivity, specificity, predictive values, and receiver operating characteristic curves, can often seem like abstract mathematic concepts that have a minimal relationship with clinical decision making or clinical research. The purpose of this article is to provide definitions and examples of these concepts that illustrate their usefulness in specific clinical decision-making tasks. In particular, nine principles are provided to guide the use of these concepts in daily radiology practice, in interpreting clinical literature, and in designing clinical research studies. An understanding of these principles and of the measures of diagnostic accuracy to which they apply is vital to the appropriate evaluation and use of diagnostic imaging examinations.
© RSNA, 2003

The bulk of the radiology literature concerns the assessment of examination performance, which is sometimes referred to as diagnostic accuracy. Despite the proliferation of such research on examination performance, it is still difficult to assess new imaging technologies, in part because such initial assessments are not always performed with an eye for how the results will be used clinically (1). The goal of this article is to describe nine fundamental principles (Appendix) to help answer specific clinical questions by using the radiology literature.

Consider the following clinical scenario: A referring physician calls you about the findings of a diagnostic mammogram that you interpreted yesterday. In the upper outer quadrant of the left breast you identified a cluster of suspicious microcalcifications—not the kind that suggests definite cancer but rather that which indicates the need for a more definitive work-up. The referring physician relays to you the patient's desire to explore the possibility of breast magnetic resonance (MR) imaging.

In this article, I will use this clinical example to illustrate the basic concepts of examination performance. To supplement previously published introductory material (2–4), I will relate the nine fundamental principles to the specific clinical scenario just described to illustrate the strengths and weaknesses of using them for clinical decision making and clinical research. I plan to answer the following questions in the course of this discussion: Which descriptors of an examination are the best intrinsic measures of performance? Which are the most clinically important? What are the limitations of sensitivity and specificity in the assessment of diagnostic examinations? What are receiver operating characteristic (ROC) curves, and why is their clinical usefulness limited? Why are predictive values more clinically relevant, and what are the pitfalls associated with using them? The ability of radiologists to understand the answers to these questions is critical to improving the application of the radiology literature to clinical practice.

TWO-BY-TWO CONTINGENCY TABLE: A SIMPLE AND UNIVERSAL TOOL

One of the most intuitive methods for the analysis of diagnostic examinations is the two-by-two table. This simple device can be jotted on the back of an envelope yet is quite

[1] From the Departments of Radiology and Epidemiology, University of Pennsylvania, 3400 Spruce St, Philadelphia, PA 19104. Received June 25, 2001; revision requested August 2; revision received May 3, 2002; accepted May 15. **Address correspondence to** the author (e-mail: *langlotz@rad.upenn.edu*).

versatile and powerful, both in the analysis of a diagnostic examination and in increasing our understanding of examination performance.

Simplifying Assumptions

The use of two-by-two tables (a more general term is *contingency tables*) requires certain simplifying assumptions and prerequisites. The first assumption that I will make is that the examination in question must be compared with a reference-standard examination—that is, one with results that yield the truth with regard to the presence or absence of the disease. In the past, this reference standard has commonly been called a "gold standard"—a term that is falling out of favor, perhaps because of the recognition that even some of the best reference standards are imperfect. For example, even clinical diagnoses supplemented by the results of the most effective histopathologic analyses are fallible (5).

A second major assumption that I will make is that the examination result must be considered either positive or negative. This is perhaps the least appealing assumption with regard to a two-by-two table, because many examinations have continuous result values, such as the degree of stenosis in a vessel or the attenuation of a liver lesion. As we will see later, this is one of the first assumptions that I will discard when advanced concepts such as ROC curves are discussed (6–8). The final assumption is that one assesses examination performance with respect to the presence or absence of a single disease and not several diseases.

Example Use of a Two-by-Two Table

Table 1 is a prototypical form of a two-by-two table. Across the top, we see two center columns, one for all cases (or patients) in which the disease is truly present (D+) and the other for all cases in which the disease is truly absent (D−). In the far left column of the table, we see the two possible examination results: positive, indicating disease presence, and negative, indicating disease absence. This table summarizes the relationship between the examination result and the reference-standard examination and defines four distinct table cells (ie, true-positive, false-positive, true-negative, and false-negative examination results). In the first row (T+), we see that a positive examination result can be either true-positive or false-positive, depending on whether the disease is present or absent, respectively. The second row (T−) shows that a negative examination result can be either false-negative or true-negative, again depending on whether the disease is present or absent, respectively.

Data in the D+ column show how the examination performs (ie, yields results that indicate the true presence or true absence of a given disease) in patients who have the disease in question. Data in the D− column show how the examination performs in patients who do not have the disease (ie, who are "healthy" with respect to the disease in question). The total numbers of patients who actually do and do not have the disease according to the reference-standard examination results are listed at the bottom of the D+ and D− columns, respectively.

The datum in the first row at the far right (TP + FP) is the total number of patients who have positive examination results; the datum in the second row at the far right (FN + TN) is the total number of patients who have negative examination results. The overall total *(N)* is the total number of patients who participated in the study of examination performance.

The example data in Table 2 are interim data from an experiment to evaluate the accuracy of breast MR imaging in patients with clinically or mammographically suspicious lesions. Like the patient with suspicious microcalcifications who is considering undergoing MR imaging in the hypothetical scenario described earlier, all patients in this experiment had suspicious lesions and were about to undergo open excisional biopsy. Prior to biopsy, each woman underwent dynamic contrast material–enhanced MR imaging of the breast. The results of histopathologic examination of the specimen obtained at subsequent excisional biopsy were used as the reference standard for disease. (A more detailed description of the experimental methodology and a more recent report of the data are published elsewhere [9].) As shown in Table 2, a total of 182 women were enrolled in the study at the time the table was constructed. Seventy-four of these women had cancer, and 108 did not. There were a total of 99 positive examination results: 71 were true-positive and 28 false- positive. The 83 negative examination results comprised three false-negative and 80 true-negative results.

In the following sections, I describe the important quantitative measures of examination performance that can be computed from a two-by-two table.

TABLE 1
Shorthand Two-by-Two Table Describing Diagnostic Examination Performance

Examination Result	D+	D−	Total
T+	TP	FP	TP + FP
T−	FN	TN	FN + TN
Total	TP + FN	FP + TN	*N*

Note.—D+ = all cases or patients in which disease is truly present (ie, according to reference-standard examination results), D− = all cases or patients in which disease is truly absent, FN = number of cases or patients with false-negative examination results, FP = number of cases or patients with false-positive examination results, *N* = overall total number of cases or patients, T+ = positive examination result, T− = negative examination result, TN = number of cases or patients with true-negative examination results, TP = number of cases or patients with true-positive examination results.

TABLE 2
Patient Data in Experiment to Study Breast MR Imaging

MR Imaging Result	Malignant	Benign	Totals
Positive	71	28	99
Negative	3	80	83
Total	74	108	182

Note.—Data are numbers of women with malignant or benign breast tumors.

SENSITIVITY AND SPECIFICITY: INTRINSIC MEASURES OF EXAMINATION PERFORMANCE

Sensitivity: Examination Performance in Patients with the Disease in Question

Principle 1: Sensitivity is a measure of how a diagnostic examination performs in a population of patients who have the disease in question. The value can be defined as how often the examination will enable detection of the disease when it is present: TP/(TP + FN).

Given the data in two-by-two Table 2, sensitivity is computed by using the numbers in the "Malignant" column. Of the 74 women who actually had cancer, 71 had a positive MR imaging result. Thus, the sensitivity of breast MR imaging in the sample of women undergoing breast biopsy was 96%—that is, 71 of 74 women with cancer were identified by using MR imaging.

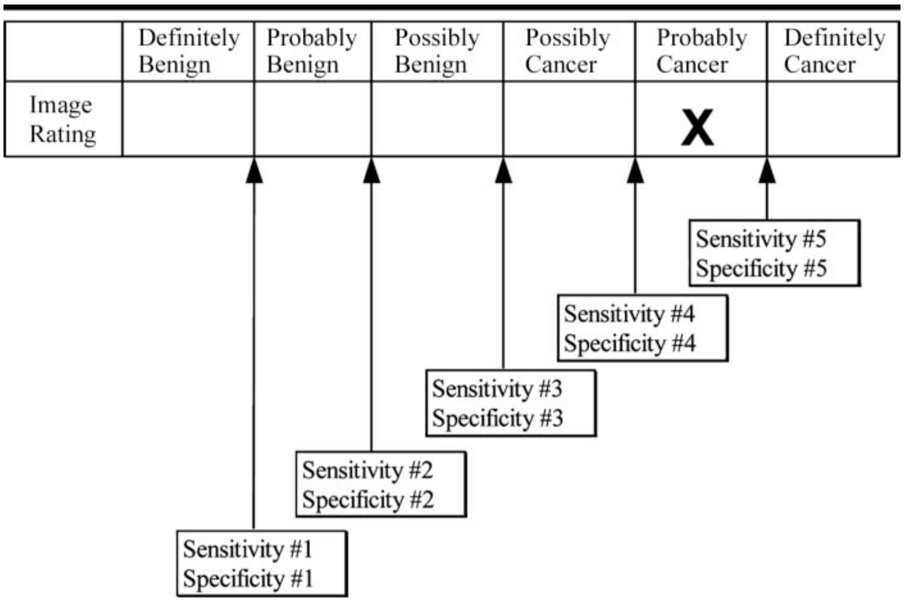


Figure 1. Six-category scale for rating the presence or absence of breast cancer. By varying a cutoff for rating categories, one can create five two-by-two tables of data from which sensitivity and specificity values can be calculated (sensitivity and specificity *#1–#5*).

Specificity: Examination Performance in Patients without the Disease in Question

Principle 2: Specificity is a measure of how a diagnostic examination performs in a population of patients who do not have the disease (ie, healthy subjects)—in other words, a value of the ability of an examination to yield an appropriately negative result in these patients. Specificity can be defined as how often a healthy patient will have a normal examination result: TN/(FP + TN).

In the example scenario, specificity is calculated by using the numbers in the "Benign" column of two-by-two Table 2. Of the 108 women who had benign lesions, 80 had negative MR imaging results. Thus, the specificity of breast MR imaging in the sample of women undergoing breast biopsy was 74%—that is, 80 of 108 women without cancer were identified by using MR imaging.

Relative Importance of Sensitivity and Specificity

How can sensitivity and specificity values be used directly to determine whether an examination might be useful in a specific clinical situation? Which value is more important? A quantitative analysis of these questions (4) is beyond the scope of this article. However, here are two qualitative rules of thumb, which together make up principle 3: A sensitive examination is more valuable in situations where false-negative results are more undesirable than false-positive results. A specific examination is more valuable in situations where false-positive results are more undesirable than false-negative results.

For example, with regard to a woman with a suspicious breast mass, we must consider how we would feel if we were to miss a cancer owing to a false-negative examination. Because we would regret this outcome, we place appropriate emphasis on developing and enhancing the sensitivity of breast MR imaging to avoid missing cancers that may progress during the follow-up interval after a false-negative MR imaging examination. We would also feel uncomfortable about referring a patient for excisional biopsy of a benign lesion, but perhaps less so, since this result would occur even if MR imaging was never performed. Consequently, this principle leads us to the conclusion that sensitivity is more important than specificity with respect to breast MR imaging in this clinical setting. Because the main potentially beneficial role of breast MR imaging in this clinical setting is to allow some women without cancer to avoid excisional biopsy, this principle also highlights the greater importance of sensitivity compared with specificity in this case. Pauker and Kassirer (10) provide a quantitative discussion of how this principle functions.

Limitations

Sensitivity and specificity are important because they are diagnostic examination descriptors that do not vary greatly among patient populations. A detailed analysis of the limitations of these measures is described elsewhere (11). Let us return to the woman with a suspicious lesion on the mammogram. She wants to know whether breast MR imaging might help her. Now that we have computed the sensitivity and specificity of MR imaging by using the data in the two-by-two table, we can convey to her the following: "If you have cancer, the chance that your MR imaging examination will be positive is 96%. If you don't have cancer, the chance that your MR imaging examination will be negative is 74%." Statements of this kind are often difficult for patients and health care providers to incorporate into their clinical reasoning. Thus, a key weakness of sensitivity and specificity values is that they do not yield information about a diagnostic examination in a form that is immediately relevant to a specific clinical decision-making task. Therefore, while the diagnostic imaging literature may contain a great deal of information about the measured sensitivity and specificity of a given examination, it often contains few data that help us assess the optimal clinical role of the examination (12).

Principle 4: The sensitivity and specificity of a diagnostic examination are related to one another. An additional important weakness of sensitivity and specificity is that these two measures cannot always be used to rank the accuracy of two examinations (or two radiologists). This weakness is particularly evident when one examination has a higher sensitivity but a lower specificity than another. The reason that examination comparison difficulties often arise with regard to sensitivity and specificity is that these two values are inherently related: You cannot evaluate one without the other. As sensitivity increases, specificity tends to decrease, and vice versa. We see this phenomenon every day when two colleagues interpret the same images differently.

Consider, for example, how two radiologists decide whether congestive heart failure is depicted on a chest radiograph. One reader may use strict criteria for the presence of congestive heart failure and thus interpret fewer findings as positive, facilitating decreased sensitivity and increased specificity. The other reader may use much more flexible criteria and thus interpret more image findings as positive for congestive heart failure, facilitating increased sensitivity but decreased specificity.

	Definitely Benign	Probably Benign	Possibly Benign	Possibly Cancer	Probably Cancer	Definitely Cancer	Totals
Cancer Cases	2	3	5	10	30	50	100
Non-Cancer Cases	50	30	10	5	3	2	100
Totals	52	33	15	15	33	52	100

	D+	D-	Totals
T+	98	50	148
T-	2	50	52
Totals	100	100	200

Sensitivity #1 = 98%
Specificity #1 = 50%

Figure 2. Sample two-by-six table showing the results of an ROC study of breast cancer identification in 200 patients. The two-by-two table at the bottom can be created by setting a cutoff between the ratings of definitely benign and probably benign. This cutoff corresponds to sensitivity *#1* and specificity *#1* in Figure 1. Sensitivity and specificity values are calculated by using the two-by-two table data. As expected, use of the more flexible criteria leads to high sensitivity but low specificity.

ROC CURVES

Comprehensive Comparisons among Diagnostic Examinations

Principle 5: ROC curves provide a method to compare diagnostic examination accuracy independently of the diagnostic criteria (ie, strict or flexible) used. When one examination is more sensitive but another is more specific, how do we decide which examination provides better diagnostic information? Or, are the accuracies of the two examinations really similar, with the exception that one involves the use of stricter criteria for a positive result? ROC curves are important and useful because they can answer these questions by explicitly representing the inherent relationship between the sensitivity and specificity of an examination. As such, ROC curves are designed to illustrate the overall information yielded by an imaging examination, regardless of the criteria a reader uses to interpret the images. Therefore, ROC curves specifically address situations in which examinations cannot be compared on the basis of sensitivity and specificity alone. A detailed discussion of ROC methodology is published elsewhere (7).

For the discussion of ROC curves, I will "relax" one of the assumptions made earlier—that of a two-value (ie, positive or negative) examination result. Instead, the readers of the images generated in the two examinations will be allowed to specify their results on a scale. Figure 1 is an illustration of a six-point rating scale for imaging-based identification of breast cancer. When using this scale, the reader of breast images is asked to specify the interpretation in terms of one of six finding categories: definitely cancer, probably cancer, possibly cancer, possibly benign, probably benign, or definitely benign. One then tabulates the ratings by using the two-by-six table shown in Figure 2.

The rating scale and the table produced by using it provide multiple opportunities to measure sensitivity and specificity. For example, we can assume that the examination is positive only when "definitely cancer" is selected and is negative otherwise. Next, we can assume that the probably cancer and definitely cancer ratings both represent positive examination results and that the remaining ratings represent negative results. When the two-by-six table is collapsed in this manner, a new two-by-two table is formed comprising higher sensitivity and lower specificity values than the first two-by-two table (because less strict imaging criteria were used and more image findings were rated as positive).

We can repeat this process five times, concluding with a two-by-two table such as that shown in Figure 2 (bottom table), in which only the definitely benign ratings are considered to represent negative results and the remaining ratings are considered to represent positive results. This approach would result in low specificity and high sensitivity. Figure 3 shows a plot of all five sensitivity-specificity pairs that can be derived from the two-by-six table data in Figure 2 and the ROC curve defined by these points.

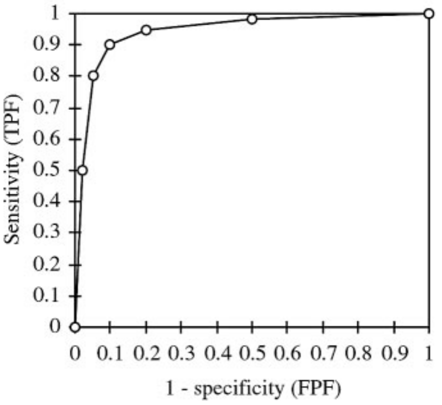


Figure 3. Sample ROC curve. This curve is a plot of sensitivity versus (1 − specificity). The 0,0 point and 1,1 point are included by default to represent the situation in which all images are considered to be either negative or positive, respectively. *FPF* = false-positive fraction, *TPF* = true-positive fraction.

Clinical Limitations

Several methods to quantitatively assess an examination on the basis of the ROC curve that it yields have been designed. The most popular method is to measure the area under the ROC curve, or the A_z. In general, the larger the area under the ROC curve, the better the diagnostic examination. Despite the advantages of these measurements of area under the ROC curve for research and analysis, they do not provide information that is useful to the clinician or patient in clinical decision making. Thus, the value of the area under the ROC curve has no intrinsic clinical meaning. Also, there is no cutoff above or below which one can be certain of the usefulness a diagnostic examination.

Should a patient with an abnormal mammogram be satisfied if she is told that the area under the ROC curve for breast MR imaging is 0.83? Although the area under the ROC curve is helpful for comparing two examinations, it has limited usefulness in facilitating the clinical decisions of patients and referring physicians.

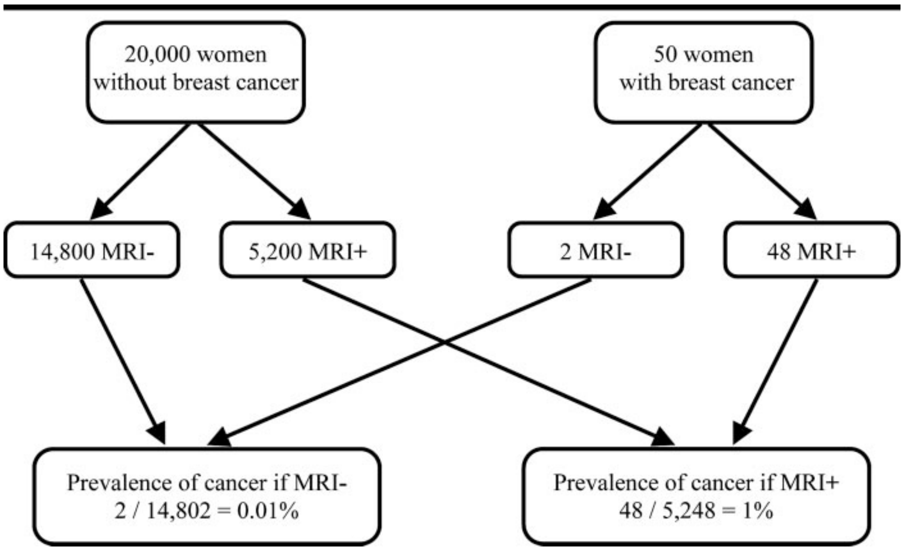


Figure 4. Flowchart depicts a simulated population of 20,050 low-risk asymptomatic women who might be screened with breast MR imaging. The assumed prevalence of cancer in this screening population of 50 women with and 20,000 women without cancer is 0.25%. *MRI+* and *MRI−* = positive and negative MR imaging results, respectively.

PREDICTIVE VALUES

Measuring Postexamination Likelihood of Disease

Because sensitivity, specificity, and ROC curves provide an incomplete picture of the clinical usefulness of an imaging examination, I will now shift back to the original two-by-two table for breast MR imaging (Table 2) to examine two additional measurements that have much greater clinical relevance and intuitive appeal: positive and negative predictive values. These quantities emphasize an important principle of diagnostic examinations—principle 6: A diagnostic examination causes a change in our belief about the likelihood that disease is truly present.

For example, the data in two-by-two Table 2 indicate that the probability of cancer in the women with suspicious mammograms was 41% (74 of 182 women). Thus, simply on the basis of her referral for excisional biopsy (without knowing the specific mammographic appearance of the lesion), the chance of cancer in the woman in the hypothetical scenario is about 41%. How can MR imaging help modify this likelihood to benefit the patient? The predictive values can help answer this question.

Principle 7: The positive predictive value indicates the likelihood of disease given a positive examination. Positive predictive value is defined as the probability of disease in a patient whose examination result is abnormal: TP/(TP + FP). Thus, we can compute the positive predictive value by using only the numbers in the first row (“Positive”) of two-by-two Table 2. A total of 99 patients had positive MR imaging results. Seventy-one of these patients actually had cancer. Thus, the positive predictive value of breast MR imaging is 72%—in other words, 71 of the 99 women with positive MR imaging results had cancer. These values are sometimes referred to as “posttest likelihoods” or “posttest probabilities of disease,” because the predictive value simply reflects the probability of disease after the examination result is known. The positive predictive value tells us, as expected, that a positive breast MR imaging result increases the probability of disease from 41% to 72%.

Principle 8: The negative predictive value indicates the likelihood of no disease given a negative examination. The negative predictive value is the negative analog of the positive predictive value. The negative predictive value can be defined as the probability that disease is absent in a patient whose examination result is negative: TN/(FN + TN). Thus, the negative predictive value can be computed solely by using the values in the second row (“Negative”) of two-by-two Table 2. A total of 83 patients in the sample had negative MR imaging results. Eighty of these patients actually had benign lesions; there were three false-negative results. Thus, the negative predictive value of breast MR imaging was 96%—in other words, 80 of the 83 women with negative MR imaging results did not have cancer. (Note: It is coincidence that the sensitivity and negative predictive value are approximately equivalent in this case.) The probability of disease after a negative examination is simply 100% minus the negative predictive value, or 4% in this case. This computation tells us—as expected—that a negative examination decreases the probability of disease in this case from 41% to 4%.

The clinical usefulness of the predictive values is best illustrated by the first question that the patient in our hypothetical scenario might ask after she has undergone MR imaging: “Do I have cancer?,” which in the uncertain world of medicine, can be translated as “How likely is it that I have cancer?” The predictive values, in contrast to sensitivity and specificity, answer this question and therefore are helpful in the clinical decision-making process. Knowing that a negative MR imaging result decreases the chance of cancer to 4% raises several additional questions for the clinician and patient to consider: Is a 4% likelihood low enough that excisional biopsy could be deferred during a short period of follow-up? Is it worth it to trade the potential harm of tumor progression during short-interval follow-up for the potential benefit of not undergoing excisional biopsy? These trade-offs can be considered explicitly by using decision analysis (13) but are routinely considered implicitly by referring physicians, patients, and other medical decision makers.

Limitations

Although predictive values have substantial clinical usefulness, a discussion of their weaknesses is warranted. The most important weakness is the dependence of predictive values on the preexamination probability, or the prevalence of disease in the imaged population. As emphasized earlier, a diagnostic examination causes a change in our belief about the likelihood of disease. And, as expected, the higher the preexamination probability of disease, the higher the postexamination probability of disease. Thus, predictive values are directly dependent on the population in which the given examination is performed.

Consider the following example, which illustrates this dependence: Since breast MR imaging may depict some cancers that were missed with mammography, why not use MR imaging as a screening tool to detect cancer in low-risk asymptomatic women? This approach has some intuitive appeal since breast MR imaging is a highly sensitive examination and might depict a greater number of cancers than would be detected with screening mammography. To illustrate the implications of this approach, a simulation of what would occur if a group of low-risk asymptomatic women were screened with MR imaging is provided (Fig 4). To approximate the prevalence of cancer in a low-risk screening population and to simplify the calculations, I will use 0.25% as the prevalence of breast cancer. For simplicity, I will consider a screening population of 20,050 women, 50 of whom have occult cancer.

Since we have established that MR imaging is 96% sensitive, 48 of the 50 women who actually have cancer will have positive MR imaging examinations ($50 \times .96 = 48$). The other two women will have false-negative MR imaging examinations and undetected cancer, just as if they had never undergone screening MR imaging. With the 74% specificity for breast MR imaging computed earlier, 14,800 women will have normal MR imaging examinations ($20{,}000 \times .74 = 14{,}800$). The remaining 5,200 women will have false-positive examinations. Table 3 is a two-by-two table containing these data.

Because the true disease status of these women will be unknown at the time of examination, clinical inferences must be drawn from the examination results and predictive values. According to principle 8, the negative predictive value indicates the clinical implications of a negative examination. Since there are 14,802 women with negative examinations, only two of whom have cancer, the negative predictive value is 0.01% (two of 14,802 women). This value represents a decrease from 0.25% and has no real clinical importance in a screening population; however, it does have some potential reassurance value.

There are 5,248 women with positive examinations in the simulation, and 48 of them actually have cancer, so the likelihood of cancer is approximately 1.00% (48 of 5,248 women). Thus, a positive MR imaging examination increases the likelihood of cancer from 0.25% to 1.00%. This group of women represents a clinical problem, however: Are we willing to perform 100 biopsies to find one cancer? Probably not. Should these women be followed up with a special more intensive regimen? Perhaps, but there are lingering questions regarding the cost-effectiveness of this program, which would likely cost tens of millions of dollars and lead to a substantial increase in the number of negative excisional biopsies.

This example of screening MR imaging illustrates clearly that the predictive values for breast MR imaging are vastly different in a screening population with a much smaller prevalence of disease. Likewise, the values of clinical usefulness of breast MR imaging as a screening examination are in stark contrast to the analogous measures of MR imaging performed in women with mammographically suspicious lesions. This contrast illustrates a weakness of predictive values: They vary according to the population in which the given examination is performed. Although the predictive values for breast MR imaging performed in women with suspicious mammograms are appealing, the predictive values for this examination performed in a screening population suggest that it has little value for asymptomatic women with low breast cancer risk. Another realistic clinical example of this phenomenon is described elsewhere (14).

Despite these limitations, predictive values can be determined mathematically from sensitivity, specificity, and prevalence data. Because sensitivity and specificity values are often published, a clinician can compute the predictive values for a particular population of interest by using the prevalence of disease in that population and the sensitivity and specificity values provided in the literature.

LIKELIHOOD RATIO

Quantifying Changes in Disease Likelihood

Principle 9: Likelihood ratios enable calculation of the postexamination probability of disease from the preexamination probability of disease. A brief description of the likelihood ratio (15,16) is relevant here because this measurement is not affected by disease prevalence and can yield clinically useful information. Likelihood ratio is defined as the probability that a person with a disease will have a particular examination result divided by the probability that a person with no disease will have that same result. Positive likelihood ratio (LR+), sometimes expressed as λ, is defined as the likelihood, or probability, that a person with a disease will have a positive examination divided by the likelihood that a person with no disease will have a positive examination: LR+ = sensitivity/(1 − specificity). Negative likelihood ratio (LR−) is defined as the probability that a person with a disease will have a negative examination divided by the probability that a person without the disease will have a negative examination: LR− = (1 − sensitivity)/specificity.

TABLE 3
Patient Data in a Hypothetical Group of Women Undergoing Screening Breast MR Imaging

MR Imaging Result	Malignant	Benign	Total
Positive	48	5,200	5,248
Negative	2	14,800	14,802
Total	50	20,000	20,050

Note.—Data are numbers of women with malignant or benign breast tumors.

To illustrate the use of likelihood ratios, consider the breast MR imaging example: The sensitivity of breast MR imaging is 96%, and the specificity is 74%. Therefore, the positive likelihood ratio is calculated by dividing the sensitivity by (1 − specificity). Thus, LR+ = 0.96/(1 − 0.74) = 3.7. The negative likelihood ratio is calculated by dividing (1 − sensitivity) by the specificity. Thus, LR− = (1 − 0.96)/0.74 = 0.055.

Once the likelihood ratios have been calculated, they can be used to calculate the postexamination probability of disease given the preexamination probability of disease, or *P*(D+|T+). For this calculation, one first must convert probabilities of disease to odds of disease. Odds of disease, Odds(D+), is defined as the probability that disease is present, *p*(D+), divided by the probability that disease is absent, *p*(D−): Odds(D+) = *p*(D+)/*p*(D−).

To compute the postexamination probability of disease given a positive examination result for any preexamination probability value and examination result, multiply the preexamination odds of disease, Odds(D+), by the positive likelihood ratio for the examination result, LR+, to obtain the postexamination odds, Odds(D+|T+). In other words, the postexamination odds of having a given disease given a positive examination result is equal to the positive likelihood ratio multiplied by the preexamination odds of the disease: Odds(D+|T+) = LR+ · Odds(D+).

Finally, to determine the postexamination probability of disease, *P*(D+|T+), convert the postexamination odds back to a postexamination probability value. The postexamination probability of disease given the examination result can be computed from the postexamination odds as follows:

$$P(D+|T+) = \frac{\text{Odds}(D+|T+)}{1+\text{Odds}(D+|T+)}$$

With these three formulas, the likelihood ratio can be used to determine the postexamination probability of disease (or predictive value) from any preexamination probability of disease.

Limitations of the Likelihood Ratio

The likelihood ratio has several properties that limit its usefulness in describing diagnostic examinations. First, it functions only on the basis of the odds of disease rather than the more intuitive probability of disease. Accordingly, the likelihood ratio is best considered on a logarithmic scale: Likelihood ratios of less than 1.0 indicate that the examination result will decrease disease likelihood, and ratios of greater than 1.0 indicate that the examination result will increase disease likelihood. To many, it is not obvious that a likelihood ratio of 4.0 increases the likelihood of disease to the same degree that a likelihood ratio of 0.25 decreases the likelihood. Furthermore, it is sometimes counterintuitive that the same likelihood ratio causes different absolute changes in probability, depending on the preexamination probability. Despite these weaknesses, the likelihood ratio is probably underused in the radiology literature today as a measure of examination performance.

CONCLUSION

In this article, I describe nine principles (Appendix) that guide the evaluation and use of diagnostic imaging examinations in clinical practice. These principles are helpful when choosing measures to describe the capabilities of a diagnostic examination. As I have discussed, sensitivity and specificity are relatively invariant descriptors of examination accuracy. However, they have limited clinical usefulness and often cannot be used directly to compare two diagnostic examinations. ROC curves can be used to directly compare examinations independently of reader temperament or varying image interpretation criteria, but they yield little information that is useful to the clinical decision maker. Positive and negative predictive values yield useful information for clinical decision makers by facilitating explicit consideration of the trade-offs at hand, but they are intrinsically dependent on the preexamination likelihood of disease and therefore on the population of patients in whom the given examination is performed. Finally, the likelihood ratio can be used to calculate the postexamination likelihood of disease from the preexamination likelihood of disease, but the associated use of odds and the logarithmic scale are counterintuitive for some. An understanding of the described fundamental measures of examination performance and how they are clinically useful is vital to the appropriate evaluation and use of diagnostic imaging examinations.

APPENDIX

Nine principles that are helpful when choosing measures to describe the capabilities of a diagnostic examination:

1. Sensitivity is a measure of how a diagnostic examination performs in a population of patients who have the disease in question.
2. Specificity is a measure of how a diagnostic examination performs in a population of patients who do not have the disease in question (ie, healthy subjects).
3. A sensitive examination is more valuable in situations where false-negative results are more undesirable than false-positive results. A specific examination is more valuable in situations where false-positive results are more undesirable than false-negative results.
4. The sensitivity and specificity of a diagnostic examination are related to one another.
5. ROC curves provide a method to compare diagnostic examination accuracy independently of the diagnostic criteria (ie, strict or flexible) used.
6. A diagnostic examination causes a change in our belief about the likelihood that disease is truly present.
7. The positive predictive value indicates the likelihood of disease given a positive examination.
8. The negative predictive value indicates the likelihood of no disease given a negative examination.
9. Likelihood ratios enable calculation of the postexamination probability of disease from the preexamination probability of disease.

Acknowledgments: My special thanks to Harold Kundel, MD, Kimberly Applegate, MD, MS, and Philip Crewson, PhD, for their thoughtful comments on earlier versions of this manuscript.

References

1. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11:88–94.
2. Brismar J, Jacobsson B. Definition of terms used to judge the efficacy of diagnostic tests: a graphic approach. AJR Am J Roentgenol 1990; 155:621–623.
3. Black WC. How to evaluate the radiology literature. AJR Am J Roentgenol 1990; 154:17–22.
4. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med 1975; 293:211–215.
5. Burton E, Troxclair D, Newman W. Autopsy diagnoses of malignant neoplasms: how often are clinical diagnoses incorrect? JAMA 1998; 280:1245–1248.
6. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
7. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
8. Brismar J. Understanding receiver-operating-characteristic curves: a graphic approach. AJR Am J Roentgenol 1991; 157:1119–1121.
9. Nunes LW, Schnall MD, Siegelman E, et al. Diagnostic performance characteristics of architectural features revealed by high spatial-resolution MR imaging of the breast. AJR Am J Roentgenol 1997; 169:409–415.
10. Pauker S, Kassirer J. The threshold approach to clinical decision making. N Engl J Med 1980; 302:1109–1117.
11. Ransohoff D, Feinstein A. Problems of spectrum bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 229:926–930.
12. Hillman BJ. Outcomes research and cost-effectiveness analysis for diagnostic imaging. Radiology 1994; 193:307–310.
13. Hrung J, Langlotz C, Orel S, Fox K, Schnall M, Schwartz J. Cost-effectiveness of magnetic resonance imaging and needle core biopsy in the pre-operative workup of suspicious breast lesions. Radiology 1999; 213:39–49.
14. Filly RA. The "lemon" sign: a clinical perspective. Radiology 1988; 167:573–575.
15. Thornbury JR, Fryback DG, Edwards W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiology 1975; 114:561–565.
16. Black WC, Armstrong P. Communicating the significance of radiologic test results: the likelihood ratio. AJR Am J Roentgenol 1986; 147:1313–1318.

Harold L. Kundel, MD
Marcia Polansky, ScD

Index terms:
Diagnostic radiology, observer performance
Statistical analysis

Published online before print
10.1148/radiol.2282011860
Radiology 2003; 228:303–308

Measurement of Observer Agreement[1]

Statistical measures are described that are used in diagnostic imaging for expressing observer agreement in regard to categorical data. The measures are used to characterize the reliability of imaging methods and the reproducibility of disease classifications and, occasionally with great care, as the surrogate for accuracy. The review concentrates on the chance-corrected indices, κ and weighted κ. Examples from the imaging literature illustrate the method of calculation and the effects of both disease prevalence and the number of rating categories. Other measures of agreement that are used less frequently, including multiple-rater κ, are referenced and described briefly.

The statistical analysis of observer agreement in imaging is generally performed for three reasons. First, observer agreement provides information about the reliability of imaging diagnosis. A reliable method should produce good agreement when used by knowledgeable observers. Second, observer agreement can be used to check the consistency of a method for classification of an abnormality that indicates the extent or severity of disease (1) and to determine the reliability of various signs of disease (2). It can also be used to compare the performance of humans and computers (3). Third, observer agreement can provide a general estimate of the value of an imaging technique when an independent method of proving the diagnosis precludes the measurement of sensitivity and specificity or the more general receiver operating characteristic curve. In many clinical situations, imaging provides the best evidence of abnormality. Furthermore, even if an independent method for obtaining proof exists, it may be difficult to use. For every suspected lesion, a biopsy cannot be performed to obtain a specific tissue diagnosis. As we will demonstrate, currently popular measures of agreement do not necessarily reflect accuracy. However, there are statistical techniques for use of the agreement of multiple expert readers (4) or the agreement of multiple tests (5) to estimate the underlying accuracy of the test.

We illustrate the standard methods for description of agreement in regard to categorical data and point out the advantages and disadvantages of the use of these methods. We refer to some of the less common, although not less important, methods but do not describe them. Then we describe some current developments in methods for use of agreement to estimate accuracy. The discussion is limited to data that can be assigned to categories, such as positive or negative; high, medium, or low; class I–V. Data, such as lesion volume or heart size, that are collected on a continuous scale are more appropriately analyzed with methods of correlation.

MEASUREMENT OF AGREEMENT OF TWO READERS

Consider readings of the same 150 images that are reported as either positive or negative by two readers. The results are shown in Table 1 as joint agreement in a 2 × 2 format, with the responses of each reader as marginal totals. Three general indices of agreement can be derived from Table 1. The overall proportion of agreement, which we will call p_o, is calculated as follows:

$$p_o = \frac{7 + 121}{150} = 0.85.$$

The proportion is useful for calculations, but the result is usually expressed as a percentage. A p_o of 0.85 indicates that the two readers agree in regard to 85% of their interpretations. If the number of negative readings is large relative to the number of positive readings, the agreement in regard to negative readings will dominate the value of

[1] From the Department of Radiology (H.L.K.) and MCP Hahnemann School of Public Health (M.P.), University of Pennsylvania Medical Center, 3600 Market St, Suite 370, Philadelphia, PA 19104. Received November 21, 2001; revision requested January 29, 2002; revision received March 4; accepted March 11. Supported by grant P01 CA 53141 from the National Cancer Institute, National Institutes of Health, U.S. Public Health Service, Department of Health and Human Services. **Address correspondence to** H.L.K. (e-mail: *kundel@rad.upenn.edu*).

p_o and may give a false impression of performance. For example, suppose that 90% of the cases are actually negative, and two readers agree about all of the negative interpretations but disagree about the positive interpretations. The overall agreement will be at least 90% and may be greater depending on the number of positive interpretations on which they agree. As an alternative to the overall agreement, the positive and negative agreement can be estimated separately. This will give an indication of the type of decision on which readers disagree. The positive agreement, which we will call p_{pos}, is the number of positive readings that both readers agree on divided by all of the positive readings for both readers. For the data in Table 1, the positive agreement is calculated with the following equation:

$$p_{pos} = \frac{7 + 7}{(10 + 7) + (12 + 7)} = 0.39.$$

The negative agreement, which we will call p_{neg}, can be calculated in a similar way as follows:

$$p_{neg} = \frac{121 + 121}{(10 + 121) + (12 + 121)} = 0.92.$$

In the example given in Table 1, although the two readers agree 85% of the time overall, they only agree on positive interpretations 39% of the time, whereas they agree on negative interpretations 92% of the time. The advantage of calculation of p_{pos} and p_{neg} is that any imbalance in the proportion of positive and negative responses becomes apparent, as in the example. The disadvantage is that CIs cannot be calculated.

COHEN κ

Some of the observer agreement concerning findings of imaging tests can be caused by chance. For example, chance agreement occurs when the readers know in advance that most of the cases are negative and they adopt a reading strategy of reporting a case as negative whenever they are in doubt. Both will have a large percentage of negative agreements because of prior knowledge of the prevalence of negative cases, not because of information obtained from viewing of the images. An index called κ has been developed as a measure of agreement that is corrected for chance. The κ is calculated by subtracting the proportion of the readings that are expected to agree by chance, which we will call p_e, from the overall agreement, p_o, and dividing the remainder by the number of cases on which agreement is not expected to occur by chance. This is demonstrated in Equation (1) as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \qquad (1)$$

TABLE 1
Joint Judgment of Two Readers about Same 150 Images

	Second Reader		
First Reader	Positive for Disease	Negative for Disease	Total
Positive for disease	7	10	17
Negative for disease	12	121	133
Total	19	131	150

TABLE 2
Guidelines for Strength of Agreement Indicated with κ Values

κ Value	Strength of Agreement beyond Chance
<0	Poor
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Note.—Data are from Landis and Koch (8).

Another way to view κ is that if the readers read different images and the readings were paired, some agreement, namely p_o, would be observed. The observed agreement would occur purely by chance. The agreement that is expected to occur by chance, which we shall designate p_e, can be calculated. When the readings of different images are compared, the observed value, namely the p_o, should equal the expected value, p_e, because there is no agreement beyond chance and κ is zero.

The joint agreement that is expected because of chance is calculated for each combination with multiplication of the total responses of each reader contained in the marginal totals of the data table. From Table 1, the agreement expected by chance for the joint positive and joint negative responses is calculated with the following equation:

$$p_e = \left(\frac{17}{150} \cdot \frac{19}{150}\right) + \left(\frac{133}{150} \cdot \frac{131}{150}\right) = 0.79.$$

TABLE 3
Joint Judgment of Two Readers about Position of Tubes and Catheters on 100 Portable Chest Images

	Second Reader		
First Reader	Malpositioned	Correctly Positioned	Total
Malpositioned	3	3	6
Correctly positioned	2	92	94
Total	5	95	100

TABLE 4
Joint Judgment of Two Readers about Presence of Signs of Congestive Heart Failure on 100 Portable Chest Images

	Second Reader		
First Reader	CHF	No CHF	Total
CHF	20	12	32
No CHF	8	60	68
Total	28	72	100

Note.—CHF = congestive heart failure.

The value for κ is 0.31, as is calculated with this equation:

$$\kappa = \frac{0.85 - 0.79}{1 - 0.79} = 0.31.$$

The standard error, which we will call SE, of κ for a 2 × 2 table can be estimated with the following equation:

$$\mathrm{SE} = \sqrt{\frac{p_o(1 - p_o)}{n(1 - p_e)^2}},$$

$$\mathrm{SE}(\kappa) = \sqrt{\frac{0.85(1 - 0.85)}{150(1 - 0.79)^2}} = 0.14. \qquad (2)$$

A more accurate and more complicated equation for the standard error of κ can be found in most books about statistics (6,7).

The 95% CIs of κ can be calculated as follows:

$$\mathrm{CI}_{95\%} = \kappa \pm 1.96 \cdot \mathrm{SE}(\kappa). \qquad (3)$$

For example, the 95% CIs are 0.31 − 1.96 × 0.14 = 0.04 and 0.31 + 1.96 × 0.14 = 0.58.

Thus, what is the meaning of a κ of 0.31, together with an overall agreement of 0.85? The calculated value of κ can range from −1.00 to +1.00, but for practical purposes the range from zero to +1.00 is of interest. A κ of zero means that there is no agreement beyond chance, and a κ of

TABLE 5
Indices of Agreement for Readings of Two Radiologists Regarding Portable Chest Images for Position of Tubes and Catheters and Signs of Congestive Heart Failure

Agreement Index	Type of Agreement	Tubes and Catheters	Congestive Heart Failure
p_o	Overall	0.95	0.80
p_{pos}	Positive	0.54	0.67
p_{neg}	Negative	0.97	0.86
p_e	Chance	0.90	0.57
κ	Chance corrected	0.52	0.52

TABLE 6
Comparison of Unweighted and Weighted p_o and κ Calculated by Using Four-, Three-, and Two-Response Categories

	Unweighted		Quadratic Weighting	
Categories	p_o	κ	$p_o(w)$	$\kappa(w)$
Four-response	0.55	0.37	0.93	0.76
Three-response	0.66	0.48	0.92	0.71
Two-response	0.82	0.62	0.82	0.62

Note.—Values were calculated for data from Table A1.

1.00 means that there is perfect agreement. Interpretations of intermediate values are subjective. Table 2 shows the strength of agreement beyond chance for various ranges of κ that were suggested by Landis and Koch (8). The choice of intervals is entirely arbitrary but has become ingrained with frequent usage. The values calculated from Table 1 show that there is good overall agreement (p_o = 0.85) but only fair chance-corrected agreement (κ = 0.31). This paradoxical result is caused by the high prevalence of negative cases. Prevalence effects can lead to situations in which the values of κ do not correspond with intuition (9,10). This is illustrated with the data in Tables 3 and 4 that were extrapolated, with a bit of adjustment to make the numbers come out even, from a data set collected during a study of readings in regard to portable chest images obtained in a medical intensive care unit (11). Table 3 shows the agreement of the reports of two of the readers concerning the position of tubes and catheters. An incorrectly positioned tube or catheter was defined as a positive reading. Table 4 shows the agreement in regard to the reports of the same two readers about the presence of radiographic signs of congestive heart failure. The example was chosen because the actual values of κ for the two diagnoses were very close.

The agreement indices for the two types of readings are shown in Table 5. The overall agreement (95%) for the position of tubes and catheters is very high, but so is the agreement according to chance (90%) calculated from the marginal values in Table 3. This results in a low κ of 0.52, which happens to be the same κ as that for congestive heart failure. The result is not intuitively appealing, because a relatively simple decision such as that about the location of a catheter tip should have a higher index of agreement than a more difficult decision such as that concerning a diagnosis of congestive heart failure. Feinstein and Cicchetti (9) have pointed out the paradox of high overall agreement and low κ, and Cicchetti and Feinstein (10) suggest that when investigators report the results of studies of agreement they should include the three indices of κ, positive agreement, and negative agreement. We agree that this is a useful way of showing agreement data, because it provides more details about where disagreements occur and alerts the reader to the possibility of effects caused by prevalence or prior knowledge.

WEIGHTED κ FOR MULTIPLE CATEGORIES

The κ can be calculated for two readers who report results with multiple categories. As the number of categories increases, the value of κ decreases because there is more room for disagreement with more categories. However, when findings are reported by using a ranked variable, the relative importance of disagreement between categories may not be the same for adjacent categories as it is for distant categories. Two readers who consistently disagree about minimal and moderate categories would have the same value for κ calculated in the usual way as would two readers who consistently disagree about minimal and severe categories. A method for calculation of κ has been developed that allows for differences in the importance of disagreements. The usual approach is to assign weights between 1.00 and zero to each agreement pair, where 1.00 represents perfect agreement and zero represents no agreement. Assignment of weights can be very subjective and can confuse comparison of κ values between studies in which different weights were used. For theoretical reasons, Fleiss (7) suggests assignment of weights as follows:

$$w_{ij} = 1 - \frac{(i - j)^2}{(k - 1)^2}, \tag{4}$$

where w represents weight, i is the number of the row, j is the number of the column, and k is the total number of categories. The weighting is called quadratic because of the squared terms. An example of the method for calculation of weighted κ by using four categories is presented in the Appendix. In the example in the Appendix, the categories of absent, minimal, moderate, and severe are used. The weighted and unweighted values for p_o and κ are included in Table 6. The calculations were repeated by collapsing the data for four categories first into three and then into two categories: First, minimal and moderate categories were combined, and then minimal, moderate, and severe categories were combined, and these two combinations would be equivalent to normal and abnormal categories, respectively. Table 6 shows that the value of κ increases as the number of categories is decreased, thus indicating better agreement when the fine distinctions are eliminated. The weighted κ is greater than the unweighted κ when multiple categories are used and is the same as the unweighted κ when only two categories are used. Some investigators prefer to use multiple categories because they are a better reflection of actual clinical decisions, and if sensible weighting can be achieved, the weighted κ may reflect the actual agreement better than does the unweighted κ.

ESTIMATION OF κ FOR MULTIPLE READERS

When multiple readers are used, some authors calculate the values of κ for pairs of readers and then compute an average κ for all possible pairs (12–14). Fleiss (7) describes a method for calculation of a κ index for multiple readers. It has not been used very much in diagnostic imaging, although it has been reported in some studies along with values for weighted κ (15).

ADVANTAGES AND DISADVANTAGES OF THE κ INDEX

κ has the advantage that it is corrected for agreement with statistical chance, and there is an accepted method for computing confidence limits and for statistical testing. The main disadvantage of κ is that the scale is not free of dependence on disease prevalence or the number of rating categories. As a consequence, it is difficult to interpret the meaning of any absolute value of κ, although it is still useful in experiments in which a control for prevalence and for the number of categories is used. The prevalence bias makes it difficult to compare the results of clinical studies where disease prevalence may vary; for example, this may occur in studies about the screening and diagnosis of breast cancer. The disease prevalence should always be reported when κ is used to prevent misunderstanding when one is trying to make generalizations.

RELATIONSHIP BETWEEN AGREEMENT AND ACCURACY

High accuracy implies high agreement, but high agreement does not necessarily imply high accuracy. There is no direct way to infer the accuracy in regard to an image-reading task from reader agreement. Accuracy can only be implied from agreement, with the assumption that when readers agree they must be correct. We frequently make this assumption by seeking a consensus diagnosis or by obtaining a second opinion, but it is not always correct. The κ has been shown to be inconsistent with accuracy as measured by the area under the receiver operating characteristic curve (16) and should not be used as a surrogate for accuracy. Different areas under the receiver operating characteristic curve can have the same κ, and the same areas under the receiver operating characteristic curve can have different κ values. For example, Taplin et al (14) studied the accuracy and agreement of single- and double-reading screening mammograms by using the area under the receiver operating characteristic curve and κ. The study included 31 radiologists who read 120 mammograms. The mean area under the receiver operating characteristic curve for single-reading mammograms was 0.85, and that for double-reading mammograms was 0.87. However, the average unweighted κ for patients with cancer was 0.41 for single-reading mammograms and 0.71 for double-reading mammograms. The average unweighted κ for patients without cancer was 0.26 for single-reading mammograms and 0.34 for double-reading mammograms. Double reading of mammograms resulted in better agreement but not in better accuracy.

If we assume that agreement implies accuracy, then we can use measurements of observed agreement to set a lower limit for accuracy. Suppose two readers agree with respect to interpretation in 50% of the cases; then, by implication, they are both correct with respect to interpretation in 50% of the cases about which they agree and one of them is correct with respect to interpretation in half (25% of the total) of the cases about which they disagree. Therefore, the overall accuracy of the readings is 75%. Typically, in radiology, observed between-reader agreement is 70%–80%, implying an accuracy that is 85%–90% (ie, 70% + 30%/2 to 80% + 20%/2).

Some new approaches to estimation of accuracy from agreement have been proposed. These approaches are based on the assumption that when a majority of readers agree about a diagnosis they are likely to be right (4,17). We have proposed the use of a technique called mixture distribution analysis (4,18). At least five readers report the cases by using either a yes-no response or a rating scale. The agreement of the group of readers about each case is fit to a mathematic model, with the assumption that the sample was drawn from a population that consists of easy normal, easy abnormal, and hard cases. With the computer program, the population that best fits the sample is located, and an overall measure of performance that we call the relative percentage agreement is calculated. We have found that the relative percentage agreement has values similar to those obtained

TABLE A1
Frequency of Responses of Two Readers Who Rated a Disease as Absent, Minimal, Moderate, or Severe

	Reader 1				
Reader 2	Absent	Minimal	Moderate	Severe	Total
Absent	34	10	2	0	46
Minimal	6	8	8	2	24
Moderate	2	5	4	12	23
Severe	0	1	2	14	17
Total	42	24	16	28	110

Note.—The frequencies in Table A1 are converted into proportions in Table A2 by dividing by the total number of cases.

TABLE A2
Proportion of Responses of Two Readers Who Rated a Disease as Absent, Minimal, Moderate, or Severe

	Reader 1				
Reader 2	Absent	Minimal	Moderate	Severe	Total
Absent	0.31	0.09	0.02	0	0.42
Minimal	0.05	0.07	0.07	0.02	0.22
Moderate	0.02	0.05	0.04	0.11	0.21
Severe	0	0.01	0.02	0.13	0.15
Total	0.38	0.22	0.15	0.25*	1.00

* Value was rounded.

TABLE A3
Quadratic Weights for 4 × 4 Table

	Absent, 1	Minimal, 2	Moderate, 3	Severe, 4
Absent, 1	1.0	0.89	0.56	0
Minimal, 2	0.89	1.00	0.89	0.56
Moderate, 3	0.56	0.89	1.00	0.89
Severe, 4	0	0.56	0.89	1.00

Note.—Numbers 1–4 are weighting factors that correspond to the respective category.

TABLE A4
Weighted Proportion of Observed and Expected Responses

Disease Rating Category	Observed Weighted Proportions for Disease Rating Category				Expected Weighted Proportions for Disease Rating Category			
	Absent	Minimal	Moderate	Severe	Absent	Minimal	Moderate	Severe
Absent	0.31	0.08	0.01	0	0.16	0.08	0.03	0
Minimal	0.05	0.07	0.06	0.01	0.07	0.05	0.03	0.03
Moderate	0.01	0.04	0.04	0.10	0.04	0.04	0.03	0.05
Severe	0	0.01	0.02	0.13	0	0.02	0.02	0.04

by using receiver operating characteristic curve analysis with proved cases (18,19).

CONCLUSION

Formal evaluations of imaging technology by using reader agreement started in 1947 with the publication of an article about tuberculosis case finding by using four different chest imaging systems (20). The author of an editorial that accompanied the article expressed surprise that there was so much disagreement (21). History repeated itself when an article about agreement in screening mammography that showed considerable reader variability (22) was published; this article was accompanied by an editorial in which the author expressed surprise in regard to the extent of disagreement (23). The consensus of a group of physicians is frequently the only basis for determination of a difficult diagnostic decision. Studies of pathologists who classify cancer have shown levels of disagreement are similar to those associated with hard decisions in radiology (24). Agreement usually results from informal discussion; however, the method used to obtain agreement can have a large influence on the decision outcome (25). Formal procedures that are used to achieve agreement have been proposed (26); although they can minimize individual bias in achieving a consensus, they are rarely used. We hope that this brief review will stimulate greater use of existing statistics for characterization of agreement and further exploration of new methods.

APPENDIX

Consider a data set in Table A1 that consists of four categories. The frequencies in Table A1 are converted into proportions, which are included in Table A2, by dividing the data by the total number of cases.

Table A3 shows the quadratic weights calculated by using Equation (4), as presented earlier:

$$w_{ij} = 1 - \frac{(i - j)^2}{(k - 1)^2},$$

where w represents weight, i is the number of the row, j is the number of the column, and k is the total number of categories. It is assumed that disagreement between adjacent categories (ie, disagreement for absent to minimal is 0.89) is not as important as that between distant categories (ie, disagreement for absent to severe is zero).

The weighted observed agreement is calculated by multiplying the proportion of responses in each cell of the 4 × 4 table by the corresponding weighting factor. The calculations for the first row are as follows: 0.31 × 1.00 = 0.31, 0.09 × 0.89 = 0.08, 0.02 × 0.56 = 0.01, and 0 × 0 = 0.

The results for observed weighted proportions are presented in Table A4. The expected agreement is calculated by multiplying the row and column total for each cell of the 4 × 4 table by the corresponding weighting factor. The calculations for the first row are as follows: (0.42 × 0.38) × 1.00 = 0.16, (0.42 × 0.22) × 0.89 = 0.08, (0.42 × 0.15) × 0.56 = 0.03, and (0.42 × 0.25) × 0 = 0.

The results for expected weighted proportions are presented in Table A4. The sum of all of the cells in regard to observed weighted proportions (sum, 0.93) in Table A4 is the weighted observed agreement, which we call $p_o(w)$, and the sum of all of the cells in regard to expected weighted proportions (sum, 0.70) in Table A4 is the weighted expected agreement, which we call $p_e(w)$. When we apply the equation for κ to the weighted values, we get a weighted κ index of 0.76, which is calculated with the following equation:

$$\kappa(w) = \frac{p_o(w) - p_e(w)}{1 - p_e(w)}.$$

An unweighted κ can be calculated by using the sum of the diagonal cells in Table A2, or 0.31 + 0.07 + 0.04 + 0.13 = 0.55, to calculate the observed agreement and the sum of the diagonal cells in Table A4 with regard to expected weighted proportions, or 0.16 + 0.05 + 0.03 + 0.04 = 0.28, to calculate the expected agreement. The unweighted κ is 0.37.

The calculation of the appropriate standard error and the use of the standard error for testing either the hypothesis that κ is different from zero or that κ is different from a value other than zero is beyond the scope of this article but is in most basic statistical texts (6,7).

GLOSSARY

Below is a list of common terms and definitions related to the measurement of observer agreement.

Accuracy.—This value is the likelihood of the interpretation being correct when compared with an independent standard.

Agreement.—This term represents the likelihood that one reader will indicate the same responses as another reader.

Attributes.—An attribute is a categorical variable that represents a property of the object being imaged (eg, tumor descriptors such as mass, calcification, and architectural distortion).

Categorical variables.—Categorical variables are variables that can be assigned to specific categories. Categorical variables can be either ranked variables or attributes.

κ.—The κ value is an overall measure of agreement that is corrected for agreement by chance. It is sensitive to disease prevalence.

Marginal sums.—A marginal sum is the sum of the responses in a single row or column of the data table, and it represents the total response of one of the readers.

Measurement variable.—Measurement variables are variables that can be measured or counted. They are generally divided into continuous variables (eg, lesion diameter or volume) and discrete variables (eg, number of lesions, expressed as whole numbers but never as decimal fractions).

Prevalence.—Prevalence is the proportion of a particular class of cases in the population being studied.

Ranked variables.—Ranked variables are categorical variables that have a natural order, such as stage of a disease, histologic grade, or discrete severity index (ie, mild, moderate, or severe).

Reliability.—Reliability is the likelihood that one reader will provide the same responses as those provided by a large consensus group.

Weighted κ.—The weighted κ is an overall measure of agreement that is corrected for agreement by chance; a weighting factor is applied to each pair of disagreements to account for the importance of the disagreement.

References

1. Baker JA, Kornguth PJ, Floyd CE. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. AJR Am J Roentgenol 1996; 166:773–778.
2. Markus JB, Somers S, Franic SE, et al. Interobserver variation in the interpretation of abdominal radiographs. Radiology 1989; 171:69–71.
3. Tiitola M, Kivisaari L, Tervahartiala P, et al. Estimation or quantification of tumour volume? CT study on irregular phantoms. Acta Radiol 2001; 42:101–105.
4. Polansky M. Agreement and accuracy: mixture distribution analysis. In: Beutel J, VanMeter R, Kundel H, eds. Handbook of imaging physics and perception. Bellingham, Wash: Society of Professional Imaging Engineers, 2000; 797–835.
5. Henkelman RM, Kay I, Bronskill MJ. Receiver operating characteristic (ROC) analysis without truth. Med Decis Making 1990; 10:24–29.
6. Agresti A. Categorical data analysis. New York, NY: Wiley, 1990; 366–370.
7. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY: Wiley, 1981; 212–236.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159–174.
9. Feinstein A, Cicchetti D. High agreement but low kappa. I. The problem of two paradoxes. J Clin Epidemiol 1990; 43:543–549.
10. Cicchetti D, Feinstein A. High agreement but low kappa. II. Resolving the paradoxes. J Clin Epidemiol 1990; 43:551–558.
11. Kundel HL, Gefter W, Aronchick J, et al. Relative accuracy of screen-film and computed radiography using hard and soft copy readings: a receiver operating characteristic analysis using bedside chest radiographs in a medical intensive care unit. Radiology 1997; 205:859–863.
12. Epstein DM, Dalinka MK, Kaplan FS, et al. Observer variation in the detection of osteopenia. Skeletal Radiol 1986; 15:347–349.
13. Herman PG, Khan A, Kallman CE, et al. Limited correlation of left ventricular end-diastolic pressure with radiographic assessment of pulmonary hemodynamics. Radiology 1990; 174:721–724.
14. Taplin SH, Rutter CM, Elmore JG, et al. Accuracy of screening mammography using single versus independent double interpretation. AJR Am J Roentgenol 2000; 174:1257–1262.
15. Robinson PJ, Wilson D, Coral A, et al. Variation between experienced observers in the interpretation of accident and emergency radiographs. Br J Radiol 1999; 72:323–330.
16. Swets JA. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychol Bull 1986; 99:100–117.
17. Uebersax JS. Modeling approaches for the analysis of observer agreement. Invest Radiol 1992; 27:738–743.
18. Kundel HL, Polansky M. Mixture distribution and receiver operating characteristic analysis of bedside chest imaging using screen-film and computed radiography. Acad Radiol 1997; 4:1–7.
19. Kundel HL, Polansky M. Comparing observer performance with mixture distribution analysis when there is no external gold standard. In: Kundel HL, ed. Medical imaging 1998: image perception. Bellingham, Wash: Society of Professional Imaging Engineers, 1998; 78–84.
20. Birkelo CC, Chamberlain WE, Phelps PS, et al. Tuberculosis case finding: a comparison of the effectiveness of various roentgenographic and photofluorographic methods. JAMA 1947; 133:359–366.
21. The “personal equation” in the interpretation of a chest roentgenogram (editorial). JAMA 1947; 133:399–400.
22. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists’ interpretation of mammograms. N Engl J Med 1994; 331:1493–1499.
23. Kopans DB. Accuracy of mammographic interpretation (editorial). N Engl J Med 1994; 331:1521–1522.
24. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977; 33:363–374.
25. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. Invest Radiol 1983; 18:194–198.
26. Hillman BJ, Hessel SJ, Swensson RG, Herman PG. Improving diagnostic accuracy: a comparison of interactive and Delphi consultations. Invest Radiol 1977; 12:112–115.

Kimberly E. Applegate, MD, MS
Richard Tello, MD, MSME, MPH
Jun Ying, PhD

Index term:
Statistical analysis

Published online before print
10.1148/radiol.2283021330
Radiology 2003; 228:603–608

Hypothesis Testing III: Counts and Medians[1]

Radiology research involves comparisons that deal with the presence or absence of various imaging signs and the accuracy of a diagnosis. In this article, the authors describe the statistical tests that should be used when the data are not distributed normally or when they are categorical variables. These nonparametric tests are used to analyze a 2 × 2 contingency table of categorical data. The tests include the χ^2 test, Fisher exact test, and McNemar test. When the data are continuous, different nonparametric tests are used to compare paired samples, such as the Mann-Whitney *U* test (equivalent to the Wilcoxon rank sum test), the Wilcoxon signed rank test, and the sign test. These nonparametric tests are considered alternatives to the parametric *t* tests, especially in circumstances in which the assumptions of *t* tests are not valid. For radiologists to properly weigh the evidence in the literature, they must have a basic understanding of the purpose, assumptions, and limitations of each of these statistical tests.

The purpose of hypothesis testing is to allow conclusions to be reached about groups of people by examining samples from the groups. The data collected are analyzed by using statistical tests, which may be parametric or nonparametric, depending on the nature of the data to be analyzed. Statistical methods that require specific distributional assumptions are called parametric methods, whereas those that require no assumptions about how the data are distributed are nonparametric methods. Nonparametric tests are often more conservative tests compared with parametric ones. This means that the test has less power to reject the null hypothesis (1). Nonparametric tests can be used with discrete variables or data based on weak measurement scales, consisting of rankings (ordinal scale) or classifications (nominal scale).

The purpose of this article is to discuss different nonparametric or distribution-free tests and their applications with continuous and categorical data. For the analysis of continuous data, many radiologists are familiar with the *t* test, a parametric test that is used to compare two means. However, misuse of the *t* test is common in the medical literature (2). To perform *t* tests properly, we need to make sure the data meet the following two critical conditions: *(a)* The data are continuous, and *(b)* the populations are distributed normally. In this article, we introduce the application of nonparametric statistical methods when these two assumptions are not met. These methods require less stringent assumptions of the population distributions than those for the *t* tests. When two populations are independent, the Mann-Whitney *U* test can be used to compare the two population distributions (3). An additional advantage of the Mann-Whitney *U* test is that it can be used to compare ordinal data, as well as continuous data. When the observations are in pairs from the same subject, we can use either the Wilcoxon signed rank test or the sign test to replace the paired *t* test.

For categorical data, the χ^2 test is often used. The χ^2 test for goodness of fit is used to study whether two or more mutually independent populations are similar (or homogeneous) with respect to some characteristic (4–12). Another application of the χ^2 test is a test of independence. Such a test is used to determine whether two or more characteristics are associated (or independent). In our discussion, we will also introduce some extensions of the χ^2 test, such as the Fisher exact test (13,14) for small samples and the McNemar test for paired data (15).

[1] From the Department of Radiology, Riley Hospital for Children, Indianapolis, Ind (K.E.A.); Department of Radiology, Boston University, 88 E Newton St, Atrium 2, Boston, MA 02118 (R.T.); and Departments of Radiology and Biostatistics, Indiana University School of Medicine, Indianapolis (J.Y.). Received October 16, 2002; revision requested February 18, 2003; final revision received March 13; accepted April 1. **Address correspondence to** R.T. (e-mail: *tello@alum.mit.edu*).

CATEGORICAL DATA

In many cases, investigators in radiology are interested in comparing two groups of count data in a 2 × 2 contingency table. One of the most commonly used statistical tests to analyze categorical data is the χ^2 test (16). If two groups of subjects are sampled from two independent populations and a binary outcome is used for classification (eg, positive or negative imaging result), then we use the χ^2 test of homogeneity. Sometimes radiologists are interested in analyzing the association between two criteria of classification. This results in the test of independence by using a similar 2 × 2 contingency table and χ^2 statistic. When sample sizes are small, we prefer to use the Fisher exact test. If we have paired measurements from the same subject, we use the McNemar test to compare the proportions of the same outcome between these two measurements in the 2 × 2 contingency table.

TABLE 1
χ^2 Test of Homogeneity

Participants	Positive MR Imaging Result	Negative MR Imaging Result	Total
Patients	50	50	100
Control subjects	28	72	100
Total	78	122	200

Note.—Data are the number of participants. χ^2 statistic, 10.17; P = .001. The null hypothesis is that the two populations are homogeneous. We reject the null hypothesis and conclude that the two populations are different.

TABLE 2
χ^2 Test of Independence

Presence of Contrast Enhancement	MR Imaging Finding		
	Malignant Mass	Benign Mass	Total
Enhancement	14	3	17
No enhancement	3	45	48
Total	17	48	65

Note.—Data are the number of masses. χ^2 statistic, 34.65; $P < .001$. The null hypothesis is that contrast enhancement of a renal mass at MR imaging is not associated with the presence of a malignant tumor, and the alternative hypothesis is that enhancement and malignancy are associated. We reject the null hypothesis and conclude that the presence of contrast enhancement is associated with renal malignancy.

χ^2 Test

The χ^2 test allows comparison of the observed frequency with its corresponding expected frequency, which is calculated according to the null hypothesis in each cell of the 2 × 2 contingency table (Eq [A1], Appendix A). If the expected frequencies are close to the observed frequencies, the model according to the null hypothesis fits the data well; thus, the null hypothesis should not be rejected. We start with the analysis of a 2 × 2 contingency table by considering the following two examples. The same χ^2 formula is used in both examples, but they are different in the sense that the data are sampled in different ways.

Example 1: test of homogeneity between two groups.—One hundred patients and 100 healthy control subjects are enrolled in a magnetic resonance (MR) imaging study. The MR imaging result can be classified as either "positive" or "negative" (Table 1). The radiologist is interested in finding out if the proportion of positive findings in the patient group is the same as that in the control group. In other words, the null hypothesis is that the proportion of positive findings is the same in the two groups. The alternative hypothesis is that they are different. We call this a test of homogeneity. In this first example, the two groups (patients and subjects) are in the rows, and the two outcomes of positive and negative test results are in the columns. In the statistical analysis, only one variable, the imaging result (classified as positive or negative), is considered.

The results in Table 1 show that 50 patients and 28 control subjects are categorized as having positive findings. The χ^2 statistic is calculated and yields a P value of .001 (17). Typically, we reject the null hypothesis if the P value is less than .05 (the significance level). In this example, we conclude that there is no homogeneity between the two groups, since the proportions of positive imaging results are different.

Example 2: test of independence between two variables in one group.—A radiologist studies gadolinium-based contrast material enhancement of renal masses at MR imaging in 65 patients (18). Table 2 shows that there are 17 patients with enhancing renal masses, with 14 malignant masses and three benign masses at pathologic examination. Among the 48 patients with nonenhancing renal masses, three masses are malignant and 45 are benign at pathologic examination. In this example, the presence or absence of contrast enhancement is indicated in the rows, and the malignant and benign pathologic findings are in the columns. In this second example, only the total number of 65 patients is fixed; the presence or absence of contrast enhancement is compared with the pathologic result (malignant or benign). The question of interest is whether these two variables are associated. In other words, the null hypothesis is that contrast enhancement of a renal mass is not associated with the presence of a malignant tumor, and the alternative hypothesis is that enhancement and malignancy are associated. In this example, the χ^2 statistic yields a P value less than .001. We reject the null hypothesis and conclude that the presence of contrast enhancement at MR imaging is associated with renal malignancy.

One potential issue with the χ^2 test is that the χ^2 statistic is discrete, since the observed frequencies in the 2 × 2 contingency table are counts. However, the χ^2 distribution itself is continuous. In 1934, Yates (12) proposed a procedure to correct for this possible bias. Although there is controversy about whether to apply this correction, it is sometimes used when the sample size is small. In the first example discussed earlier, the χ^2 statistic was 10.17, and the P value was .001. The Yates corrected χ^2 statistic is 9.27 with a P value of .002. This corrected χ^2 statistic yields a smaller χ^2 statistic, and the P value is larger after Yates correction. This indicates that the Yates corrected χ^2 test is less powerful in rejecting the null hypothesis. Some applications of Yates correction in medicine are discussed in the statistical textbook by Altman (19).

TABLE 3
Fisher Exact Test for Small Samples

Participants	Positive CT Result	Negative CT Result	Total
Patients	10	10	20
Control subjects	4	16	20
Total	14	26	40

Note.—Data are the number of participants. (Two-sided) Fisher exact test result, $P = .10$. For small samples, the Fisher exact test is used. The null hypothesis is that the two populations of patients and control subjects are homogeneous at CT—that is, they have the same number of positive results. We retain the null hypothesis because the *P* value does not indicate a significant difference, and we conclude that these two groups are homogeneous. If we incorrectly used the χ^2 test for this comparison, the conclusion would have been the opposite: $\chi^2 = 3.96$, $P = .04$.

TABLE 4
McNemar Test for Paired Comparisons: Angiography versus CT Results in the Diagnosis of Coronary Bypass Graft Thrombosis

CT Result	Positive Angiography Result	Negative Angiography Result	Total
Positive	71	30	101
Negative	13	86	99
Total	84	116	200

Note.—Data are the number of CT results. McNemar χ^2 result, 5.95; $P = .02$. The *P* value indicates a significant difference, and therefore, we reject the null hypothesis and conclude that there is a difference between these two modalities.

TABLE 5
Incorrect Use of the χ^2 Test for Paired Data for the Evaluation of Angiography versus CT (when paired data are incorrectly treated as independent)

Modality	Positive Result	Negative Result	Total
CT	101	99	200
Angiography	84	116	200

Note.—Data are the number of results. For the χ^2 test with the assumption of two independent samples, $P = .09$. We would incorrectly conclude that there is no significant difference between these two modalities.

Fisher Exact Test

When sample sizes are small, the χ^2 test yields poor results, and the Fisher exact test is preferred. A general rule of thumb for its use is when either the sample size is less than 30 or the expected number of observations in any one cell of a 2 × 2 contingency table is fewer than five (20). The test is called an "exact" test because it allows calculation of the exact probability (rather than an approximation) of obtaining the observed results or results that are more extreme. Although radiologists may be more familiar with the traditional χ^2 test, there is no reason not to use the Fisher exact test in its place, given the ease of use and availability of computer software today.

In example 1, the *P* value resulting from use of the χ^2 test was .001, whereas the *P* value for the same data tested by using the Fisher exact test was .002. Both tests lead to the same conclusion of lack of homogeneity between the patient and control groups. Intuitively, the *P* value derived by using the Fisher exact test is the probability of positive results becoming more and more discrepant between the two groups. Most statistical software packages provide computation of the Fisher exact test (Appendix B).

Example 3: Fisher exact test.—A radiologist enrolls 20 patients and 20 healthy subjects in a computed tomographic (CT) study. The CT result is classified as either "positive" or "negative." Table 3 shows that 10 patients and four healthy subjects have positive findings at CT. The null hypothesis is that the two populations are homogeneous in the number of positive findings seen at CT.

In this example, the sample sizes in both the patient and control groups are small. The Fisher exact test yields a *P* value of .10. We retain the null hypothesis because the *P* value does not indicate a significant difference, and we conclude that these two groups are homogeneous. If we use the χ^2 test incorrectly, the *P* value is .05, which suggests the opposite conclusion—that the proportions of positive CT results are different in these two groups.

McNemar Test for Paired Data

A test for assessment of paired count data is the McNemar test (15). This test is used to compare two paired measurements from the same subject. When the sample size is large, the McNemar test follows the same χ^2 distribution but uses a slightly different formula. Radiology research often involves the comparison of two paired imaging results from the same subject. In a 2 × 2 table, the results of one imaging test are labeled "positive" and "negative" in rows, and the results of another imaging test are labeled similarly in columns. An interesting property of this table is that there are two concordant cells in which the paired results are the same (both positive or both negative) and two discordant cells in which the paired results are different for the same subject (positive-negative or negative-positive).

We are interested in analyzing whether these two imaging tests show equivalent results. The McNemar test uses only the information in the discordant cells and ignores the concordant cell data. In particular, the null hypothesis is that the proportions of positive results are the same for these two imaging tests, versus the alternative hypothesis that they are not the same. Intuitively, the null hypothesis is retained if the discordant pairs are distributed evenly in the two discordant cells. The following example illustrates the problem in more detail.

Example 4: McNemar test for paired data.—There are 200 patients enrolled in a study to compare CT and conventional angiography of coronary bypass grafts for the diagnosis of graft patency (Table 4). Seventy-one patients have positive results with both conventional angiography and CT angiography, 86 have negative results with both, 30 have positive CT results but negative conventional angiographic results, and 13 have negative CT results but positive conventional angiographic results. The McNemar test compares the proportions of the discordant pairs (13 of 200 vs 30 of 200). The *P* value of the McNemar statistic is .02, which suggests that the proportion of positive results is significantly different for the two modalities. Therefore, we conclude that the ability of these two modalities to demonstrate graft patency is different.

Some radiologists may incorrectly summarize the data in a way shown in Table 5 and perform a χ^2 test, as discussed in example 1 (21). This is a common mistake in the medical literature. In example 1, the proportions compared are 101 of 200 versus 84 of 200. The problem is the assumption that CT angiography and conventional angiography results are independent, and thus, the paired relationship between these two imaging tests is ignored (2,21). The χ^2 test has less power to reject the null hypothesis than does the McNemar test in this situation and results in a *P* value of .09. We would incorrectly conclude that there is no significant difference in the ability of these two modalities to demonstrate graft patency.

HYPOTHESIS TESTING BY USING MEDIANS

The unpaired and paired *t* tests require that the population distributions be normal or approximately so. In medicine, however, we often do not know whether a distribution is normal, or we know that the distribution departs substantially from normality.

Nonparametric tests were developed to deal with situations where the population distributions are either not normal or unknown, especially when the sample size is small (<30 samples). These tests are relatively easy to understand and simple to apply and require minimal assumptions about the population distributions. However, this does not mean that they are always preferred to parametric tests. When the assumptions are met, parametric tests have higher testing power than their nonparametric counterparts; that is, it is more likely that a false null hypothesis will be rejected.

Three commonly encountered nonparametric tests include the Mann-Whitney *U* test (equivalent to the Wilcoxon rank sum test), the Wilcoxon signed rank test, and the sign test.

Comparison of Two Independent Samples: Mann-Whitney *U* Test

The Mann-Whitney *U* test is used to compare the difference between two population distributions and assumes the two samples are independent (22). It does not require normal population distributions, and the measurement scale can be ordinal.

The Mann-Whitney *U* test is used to test the null hypothesis that there is no location difference between two population distributions versus the alternative hypothesis that the location of one population distribution differs from the other. With the null hypothesis, the same location implies the same median for the two populations. For simplicity, we can restate the null hypothesis: The medians of the two populations are the same. Three alternative hypotheses are available: *(a)* The population medians are not equal, *(b)* the population median of the first group is larger than that of the second, or *(c)* the population median of the second group is larger than that of the first. If we put the two random samples together and rank them, then, according to the null hypothesis, which holds that there is no difference between the two populations medians, the total rank of one sample would be close to the total rank of the other. On the other hand, if all the ranks of one sample are smaller than the ranks of the other, then we know almost surely that the location of one population is shifted relative to that of the other.

TABLE 6
Mann-Whitney *U* Test for Ordinal Data

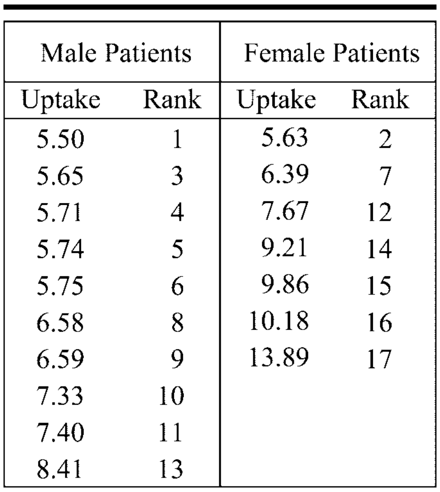
Fat Saturation	Score <25	25–75	>75	Total
No	8	14	2	24
Yes	3	12	7	22
Total	11	26	9	46

Note.—Data are scores from T2-weighted fast spin-echo MR images obtained with or without fat saturation. Wilcoxon rank sum test, $P = .03$. The null hypothesis is that the median scores for the two types of MR images are the same. We reject the null hypothesis and conclude that they are not the same. If we incorrectly used the χ^2 test, we would conclude the opposite: $\chi^2 = 5.13$, $P = .08$.

We give two examples of the application of the Mann-Whitney *U* test, one involving continuous data and the other involving ordinal data.

Example 5: Mann-Whitney U test for continuous data.—The uptake of fluorine 18 choline (hereafter, "fluorocholine") by the kidney can be considered approximately distributed normally (23). Let us say that some results of hypothetical research suggest that fluorocholine uptake above 5.5 (percentage dose per organ) is more common in men than in women. If we are only interested in the patients whose uptake is over 5.5, the distribution is no longer normal but becomes skewed. The Figure shows the uptake over 5.5 in 10 men and seven women sampled from populations imaged with fluorocholine for tumor surveillance. We are interested in finding out if there are any differences in these populations on the basis of patient sex.

We can quickly exclude use of the *t* test in this example, since the fluorocholine uptakes we have selected are no longer distributed normally. The null hypothesis is that the medians for the men and women are the same. By using the Mann-Whitney *U* test, the *P* value is .06, so we retain the null hypothesis. We conclude that the medians of these two populations are the same at the .05 significance level, and therefore, men and women have similar renal uptake of fluorocholine. If we had incorrectly used the *t* test, the *P* value would be .02, and we would conclude the opposite.



Male Patients		Female Patients	
Uptake	Rank	Uptake	Rank
5.50	1	5.63	2
5.65	3	6.39	7
5.71	4	7.67	12
5.74	5	9.21	14
5.75	6	9.86	15
6.58	8	10.18	16
6.59	9	13.89	17
7.33	10		
7.40	11		
8.41	13		

Mann-Whitney *U* test result for continuous data (fluorocholine uptake over 5.5 [percentage dose per organ] in human kidneys), $P = .06$. We retain the null hypothesis that there is no difference in the medians, and we conclude that the fluorocholine renal uptake in men and women is similar at the .05 significance level (the marginal *P* value suggests a trend toward a significant difference). If we had incorrectly used the *t* test, we would have concluded the opposite: $P = .02$.

Example 6: Mann-Whitney U test for ordinal data.—A radiologist wishes to know which of two different MR imaging sequences provides better image quality. Twenty-four patients undergo MR imaging with a T2-weighted fast spin-echo sequence, and 22 other patients are imaged with the same sequence but with the addition of fat saturation (Table 6). The image quality is measured by using a standardized scoring system, ranging from 1 to 100, where 100 is the best image quality. The null hypothesis is that the median scores are the same for the two populations. In the group imaged with the first MR sequence, the images of eight subjects are scored under 25, those of 14 subjects are scored between 25 and 75, and those of two subjects are scored above 75. In the group imaged with the fat-saturated MR sequence, there are three, 12, and seven subjects in these three score categories, respectively.

In this example, each patient's image score is classified into one of three ordinal categories. Since the observations are discrete rather than continuous, the *t* test cannot be used. Some researchers might consider the data in Table 6 to be a 2 × 3 contingency table and use a χ^2 statistic to compare the two groups. The *P* value corresponding to the χ^2 statistic is .08, and we would conclude that the two groups have similar image quality. The problem

TABLE 7
Wilcoxon Signed Rank Test for Paired Comparisons of Ring Enhancement between Two Spin-Echo MR Sequences

	MR Sequence*				
Case No.	T1-weighted	Fat-saturated T1-weighted	Difference†	Absolute Value of Difference‡	Rank†
1	45	43	−2	2	5
2	45	42	−3	3	8
3	49	47	−2	2	5
4	50	47	−3	3	8
5	49	48	−1	1	3
6	44	50	6	6	13.5
7	42	98	56	56	20
8	49	47	−2	2	5
9	39	44	5	5	11
10	42	42	0	0	1.5
11	44	54	10	10	18
12	47	53	6	6	13.5
13	42	53	11	11	19
14	45	54	9	9	16.5
15	44	48	4	4	10
16	41	47	6	6	13.5
17	45	54	9	9	16.5
18	50	47	−3	3	8
19	51	51	0	0	1.5
20	42	48	6	6	13.5

Note.—Wilcoxon signed rank test, $P = .02$. The null hypothesis is that the median of the paired population differences is zero. The Wilcoxon signed rank test result indicates a significant difference, and therefore, we conclude that the enhanced fat-saturated T1-weighted MR sequence showed ring enhancement better than did the conventional enhanced T1-weighted MR sequence. If we had incorrectly used the paired *t* test, the *P* value would be .07, and we would have had the opposite conclusion. Like the *t* test, the sign test produced a *P* value of .50, and the conclusion would be that the two sequences are the same.

* Data are image quality scores on MR images after contrast material administration.

† Difference in enhancement values between MR sequences.

‡ Rank of the absolute values of the differences; when there is a tie in the ranking, an average ranking is assigned—for example, rank 16.5 rather than ranks 16 and 17 for the tied case numbers 14 and 17; and rank 13.5 for the four cases (6, 12, 16, and 20) that compose ranks 12, 13, 14, and 15.

is that the three image quality score categories are only treated as nominal variables, and their ordinal relationship is not accounted for in the χ^2 test. An alternative test that allows us to use this information is the Mann-Whitney *U* test. The Mann-Whitney *U* test yields a *P* value of .03. We reject the null hypothesis and conclude that the median image scores are different.

Comparison of Paired Samples: Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is an alternative to the paired *t* test. Each paired sample is dependent, and the data are continuous. The assumption needed to use the Wilcoxon signed rank test is less stringent than the assumptions needed for the paired *t* test. It requires only that the paired population be distributed symmetrically about its median (24).

The Wilcoxon signed rank test is used to test the null hypothesis that the median of the paired population differences is zero versus the alternative hypothesis that the median is not zero. Since the distribution of the differences is symmetric about the mean, it is equivalent to using the mean for the purpose of hypothesis testing, as long as the sample size is large enough (at least 10 rankings).

We rank the absolute values of the paired differences from the sample. With the null hypothesis, we would expect the total rank of the pairs whose differences are negative to be comparable to the total rank of the pairs whose differences are positive. The following example shows the application of the Wilcoxon signed rank test.

Example 7: paired data.—A sample of 20 patients is used to compare ring enhancement between T1-weighted spin-echo MR images and fat-saturated T1-weighted spin-echo MR images obtained after contrast material administration (Table 7). We notice that the image quality scores on fat-saturated T1-weighted spin-echo MR images in case 7 is 98, which is much higher than the others. As a result, the difference in values between the two sequences is also much higher than that for the other paired differences. It would be unwise to use a paired *t* test in this case, since the *t* test is sensitive to extreme values in a sample and tends to incorrectly retain a false null hypothesis as a consequence. The nonparametric tests are more robust to data extremes, and thus, the Wilcoxon signed rank test is preferred in this case. The null hypothesis states that the median of the paired MR sequence differences is zero. The Wilcoxon signed rank test provides a *P* value of .02, so we reject the null hypothesis. We conclude that the fat-saturated MR sequence showed ring enhancement better than did the MR sequence without fat saturation. If we had incorrectly used the paired *t* test, the *P* value would be .07, and we would have arrived at the opposite conclusion.

Comparing Paired Samples: Sign Test

The sign test is another nonparametric test that can be used to analyze paired data. Unlike the Wilcoxon signed rank test or the paired *t* test, this test requires neither a symmetric distribution nor a normal distribution of the variable of interest. The only assumption underlying this test is that the data are continuous. Since the distribution is arbitrary with the sign test, the hypothesis of interest focuses on the median rather than the mean as a measure of central tendency. In particular, the null hypothesis for comparing paired data is that the median difference is zero. The alternative hypothesis is that the median difference is not, is greater than, or is less than zero. This simplistic test considers only the signs of the differences between two measurements and ignores the magnitudes of the differences. As a result, it is less powerful than the Wilcoxon signed rank test, and a false null hypothesis is often not rejected (25).

SUMMARY

Hypothesis testing is a method for developing conclusions about data. Radiology research often produces data that require nonparametric statistical analyses. Nonparametric tests are used for hypothesis testing when the assumptions about the data distributions are not valid or when the data are categorical. We have discussed the most common of these statistical tests and provided examples to demonstrate how to perform them. For radiologists to properly weigh the evidence in our literature, we need a basic understanding of the purpose, assumptions, and limitations of each of these statistical tests. Understanding how and when these methods are used will strengthen our ability to evaluate the medical literature.

APPENDIX A

The χ^2 formula is based on the following equation:

$$\chi^2 = \Sigma\left[\frac{(\mathrm{Fo} - \mathrm{Fe})^2}{\mathrm{Fe}}\right], \qquad (\mathrm{A1})$$

where Fo is the frequency observed in each cell, and Fe is the frequency expected in each cell, which is calculated by multiplying the row frequency by the quotient of the column frequency divided by total sample size.

APPENDIX B

Examples of statistical software that is easily capable of calculating χ^2 and McNemar statistics include SPSS, SAS, StatXact 5, and EpiInfo (EpiInfo allows calculation of the Fisher exact test and may be downloaded at no cost from the Centers for Disease Control and Prevention Web site at *www.cdc.gov*). Other statistical Web sites include *fonsg3.let.uva.nl/Service/Statistics.html, department.obg.cuhk.edu.hk/ResearchSupport/WhatsNew.asp,* and *www.graphpad.com/quickcalcs/Contingency1.cfm* (all Web sites accessed January 30, 2003).

Acknowledgments: The authors thank Miriam Bowdre for secretarial support in preparation of the manuscript and George Parker, MD, Phil Crewson, PhD, and an anonymous statistical reviewer for comments on earlier drafts of the manuscript.

References

1. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995; 526.
2. Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. BMJ 1977; 1:85–87.
3. Wilcoxon F. Individual comparisons by ranking methods. Biometrics 1945; 1:80–83.
4. Fisher LD, Belle GV. Biostatistics: a methodology for the health sciences. New York, NY: Wiley, 1993.
5. Ott RL. An introduction to statistical methods and data analysis. 4th ed. Belmont, Calif: Wadsworth, 1993.
6. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995.
7. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall, 1991.
8. Chase W, Bown F. General statistics. 4th ed. New York, NY: Wiley, 2000.
9. Conover WJ. Practical nonparametric statistics. 3rd ed. New York, NY: Wiley, 1999.
10. Rosner B. Fundamentals of biostatistics. 4th ed. Boston, Mass: Duxbury, 1995; 370–565.
11. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY: Wiley, 1981; 112–125.
12. Yates F. Contingency tables involving small numbers and the chi-square test. J Royal Stat Soc Ser B 1934; (suppl 1):2179–2235.
13. Fisher RA. Statistical methods for research workers. 5th ed. Edinburgh, Scotland: Oliver & Boyd, 1934.
14. Fisher RA. The logic of inductive inference. J Royal Stat Soc Ser A 1935; 98:39–54.
15. Cochran WG. Some methods for strengthening the common chi-square tests. Biometrics 1954; 10:417–451.
16. Agresti A. Categorical data analysis. New York, NY: Wiley, 1990.
17. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995; 691.
18. Tello R, Davidson BD, O'Malley M, et al. MR imaging of renal masses interpreted on CT to be suspicious. AJR Am J Roentgenol 2000; 174:1017–1022.
19. Altman D. Practical statistics for medical research. London, England: Chapman & Hall, 1991; 252.
20. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995; 537.
21. Dwyer AJ. Matchmaking and McNemar in the comparison of diagnostic modalities. Radiology 1991; 178:328–330.
22. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947; 18:50–60.
23. DeGrado TR, Reiman R, Price DT, Wang S, Coleman RE. Pharmacokinetics and radiation dosimetry of F-fluorocholine. J Nucl Med 2002; 43:92–96.
24. Altman D. Practical statistics for medical research. London, England: Chapman & Hall, 1991; 194–197.
25. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995; 569–578.

Nancy A. Obuchowski, PhD

Index terms:
Diagnostic radiology
Statistical analysis

Published online
10.1148/radiol.2291010898
Radiology 2003; 229:3–8

Abbreviations:
FPR = false-positive rate
MS = multiple sclerosis
ROC = receiver operating characteristic

[1] From the Department of Biostatistics and Epidemiology/Wb4, Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195. Received May 8, 2001; revision requested June 11; revision received August 1; accepted August 2. **Address correspondence to** the author (e-mail: *nobuchow@bio.ri.ccf.org*).

Receiver Operating Characteristic Curves and Their Use in Radiology[1]

Sensitivity and specificity are the basic measures of accuracy of a diagnostic test; however, they depend on the cut point used to define "positive" and "negative" test results. As the cut point shifts, sensitivity and specificity shift. The receiver operating characteristic (ROC) curve is a plot of the sensitivity of a test versus its false-positive rate for all possible cut points. The advantages of the ROC curve as a means of defining the accuracy of a test, construction of the ROC, and identification of the optimal cut point on the ROC curve are discussed. Several summary measures of the accuracy of a test, including the commonly used percentage of correct diagnoses and area under the ROC curve, are described and compared. Two examples of ROC curve application in radiologic research are presented.

Sensitivity and specificity are the basic measures of the accuracy of a diagnostic test. They describe the abilities of a test to enable one to correctly diagnose disease when disease is actually present and to correctly rule out disease when it is truly absent. The accuracy of a test is measured by comparing the results of the test to the true disease status of the patient. We determine the true disease status with the reference standard procedure.

Consider as an example the test results of 100 patients who have undergone mammography (Table 1). According to biopsy results and/or 2-year follow-up results (ie, the reference standard procedures), 50 patients actually have a malignant lesion and 50 patients do not. If these 100 test results were from 100 asymptomatic women without a personal history of breast cancer, then we might define a positive test result as any that represents a "suspicious" or "malignant" finding and a negative test result as any that represents a "normal," "benign," or "probably benign" finding. We have used a cut point for defining positive and negative test results. The cut point is located between the suspicious and probably benign findings. The estimated sensitivity with this cut point is (18 + 20)/50 = 0.76, and the specificity is (15 + 3 + 18)/50 = 0.72.

Alternatively, if these 100 test results were from 100 asymptomatic women with a personal history of breast cancer, then we might use a different cut point, such that a positive test result represents a probably benign, suspicious, or malignant finding and a negative test result represents a normal or benign finding. The estimates of sensitivity and specificity would change (ie, they would now be 0.96 and 0.36, respectively).

Important point: Sensitivity and specificity depend on the cut point used to define positive and negative test results. As the cut point shifts, the sensitivity increases while the specificity decreases, or vice versa.

COMBINED MEASURES OF SENSITIVITY AND SPECIFICITY

It is often useful to summarize the accuracy of a test by using a single number; for example, when comparing two diagnostic tests, it is easier to compare a single number than to compare both the sensitivity and the specificity values of the tests. There are several such summary measures; I will describe a popular but easily misinterpreted one that is usually referred to simply as accuracy. Using the second cut point in Table 1, we can compute accuracy as the percentage of correct diagnoses in the entire sample—that is, (48 + 18)/100 = 0.66, or 66%. The strength of this measure of accuracy is its simple computa-

TABLE 1
Results from Mammography Study with 100 Patients

Cut Point and Reference Standard Result	Radiologist's Interpretation					
	Normal	Benign	Probably Benign	Suspicious	Malignant	Total
Cut point 1*						
Reference standard result						
Cancer present	2	0	10	18†	20†	50
Cancer absent	15	3	18	13‡	1‡	50
Cut point 2*						
Reference standard result						
Cancer present	2	0	10†	18†	20†	50
Cancer absent	15	3	18‡	13‡	1‡	50

Note.—Data are numbers of patients with the given result in a fictitious study of mammography in which 50 patients had a malignant lesion and 50 did not.

* For cut point 1, a positive result is defined as a test score of suspicious or malignant; for cut point 2, a positive result is defined as a test score of probably benign, suspicious, or malignant.

† Test results considered true-positive (for estimating sensitivity) with this cut point.

‡ Test results considered false-positive (for estimating the false-positive rate [FPR] or specificity) with this cut point.

TABLE 2
Effect of Prevalence on Accuracy

Reference Standard Result	Radiologist's Interpretation					
	Normal	Benign	Probably Benign	Suspicious	Malignant	Total
Cancer present	2	0	10	18	20	50
Cancer absent	285	57	342	247	19	950

Note.—Data are numbers of patients with the given result in a fictitious study of mammography with 1,000 patients. This data set represents a modification of the data set in Table 1 so that the prevalence of cancer is 5%. When cut point 2 (described in the note to Table 1) is used with this data set, the estimated sensitivity ([10 + 18 + 20]/50 = 0.96) and specificity ([285 + 57]/950 = 0.36) are the same as with the data set in Table 1. However, one commonly used estimate of overall accuracy is the percentage of correct diagnoses in the sample. With this data set it is 39% ([10 + 18 + 20 + 285 + 57]/1,000 = 0.39), which is not the same as with the data set in Table 1.

tion. It has several limitations, however: Its magnitude varies as the prevalence of disease varies in the sample, it is calculated on the basis of only one cut point, and false-positive and false-negative results are treated as if they are equally undesirable. As an illustration of the first limitation, note that in Table 2 the prevalence of disease is 5% instead of the 50% in Table 1. The sensitivity and specificity values are the same in Tables 1 and 2, yet the estimated accuracy value in Table 2 drops to (48 + 342)/1,000 = 0.39, or 39%.

Important point: A measure of test accuracy is needed that combines sensitivity and specificity but does not depend on the prevalence of disease.

RECEIVER OPERATING CHARACTERISTIC CURVE

In 1971, Lusted (1) described how receiver operating characteristic (ROC) curves could be used to assess the accuracy of a test. An ROC curve is a plot of test sensitivity (plotted on the y axis) versus its FPR (or 1 − specificity) (plotted on the x axis). Each point on the graph is generated by using a different cut point. The set of data points generated from the different cut points is the empirical ROC curve. We use lines to connect the points from all the possible cut points. The resulting curve illustrates how sensitivity and the FPR vary together.

Figure 1 illustrates the empirical ROC curve for the mammography example. Since in our example there are five categories for the test results, we can compute four cut points for the ROC curve. The two endpoints on the ROC curve are 0,0 and 1,1 for FPR, sensitivity. The points labeled 1 and 2 on the curve correspond to the first and second cut points, respectively, that are defined in the note to Table 1. Estimations of the other points are provided in Table 3.

The ROC plot has many advantages over single measurements of sensitivity and specificity (2). The scales of the curve—that is, sensitivity and FPR—are the basic measures of accuracy and are easily read from the plot; the values of the cut points are often labeled on the curve as well. Unlike the measure of accuracy defined in the previous section (ie, the percentage of correct diagnoses), the ROC curve displays all possible cut points. Because sensitivity and specificity are independent of disease prevalence, so too is the ROC curve. The curve does not depend on the scale of the test results (ie, we can alter the test results by adding or subtracting a constant or taking the logarithm or square root without any change to the ROC curve) (3). Lastly, the ROC curve enables a direct visual comparison of two or more tests on a common set of scales at all possible cut points.

It is often convenient to make some assumptions about the distribution of the test results and then to draw the ROC curve on the basis of the assumed distribution (ie, assumed model). The resulting curve is called the fitted or smooth ROC curve. The fitted curve for the mammography study is plotted in Figure 1; it was constructed from a binormal distribution (ie, two normal distributions: one for the test results of patients without breast cancer and another for test results of patients with breast cancer) (Fig 2). The binormal distribution is the most

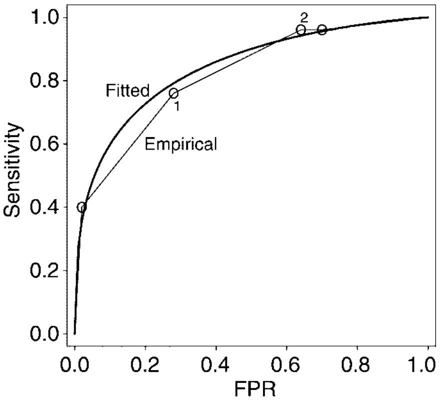


Figure 1. Graph of the empirical and fitted ROC curves for the mammography study. The points on the empirical curve are marked with open circles and are estimated in Table 3. The points labeled *1* and *2* on the curve correspond to the first and second cut points, respectively, that are defined in the note to Table 1.

TABLE 3
Construction of Empirical ROC Curve for Mammography Study

Cut Point	Sensitivity*	FPR†
Between normal and benign	0.96 (48/50)	0.70 (35/50)
Between benign and probably benign	0.96 (48/50)	0.64 (32/50)
Between probably benign and suspicious	0.76 (38/50)	0.28 (14/50)
Between suspicious and malignant	0.40 (20/50)	0.02 (1/50)

Note.—These data represent estimations of the points on the empirical ROC curve marked with open circles and depicted in Figure 1. The ROC curve in Figure 1 was constructed on the basis of the data in Table 1, with sensitivity and the FPR estimated at each possible cut point.

* Data in parentheses are those used to calculate the sensitivity value.

† Data in parentheses are those used to calculate the FPR (or 1 − specificity) value.

commonly used distribution for estimating the smooth ROC curve. There are computer programs (for example, *www-radiology.uchicago.edu/sections/roc/software.cgi*) for estimating the smooth ROC curve on the basis of the binormal distribution; these programs make use of a statistical method called maximum likelihood estimation.

An ROC curve can be constructed from objective measurements of a test (eg, serum glucose level as a test for diabetes), objective evaluation of image features (eg, the computed tomographic [CT] attenuation coefficient of a renal mass relative to normal kidney), or subjective diagnostic interpretations (eg, the five-category Breast Imaging Reporting and Data System scale used for mammographic interpretation) (5). The only requirement is that the measurements or interpretations can be meaningfully ranked in magnitude. With objective measurements the cut point is explicit, so one can choose from an infinite number of cut points along the continuum of the test results. For diagnostic tests whose results are interpreted subjectively, the cut points are implicit or latent in that they only exist in the mind of the observer (6). Furthermore, it is assumed that each observer has his or her own set of cut points.

The term *receiver operating characteristic curve* comes from the idea that, given the curve, we, the receivers of the information, can use (or operate at) any point on the curve by using the appropriate cut point. The clinical application determines which cut point is used. For example, for evaluating women with a personal history of breast cancer, we need a cut point with good sensitivity (eg, cut point 2 in Table 1), even if the FPR is high. For evaluating women without a personal history of breast cancer, we require a lower FPR. For each application the optimal cut point (2,7) can be determined by finding the sensitivity and specificity pair that maximizes the function sensitivity − *m*(1 − specificity), where *m* is the slope of the ROC curve as follows:

$$m = \frac{\mathrm{Prob}_{\mathrm{Norm}}}{\mathrm{Prob}_{\mathrm{Dis}}} \times \frac{(C_{\mathrm{FP}} - C_{\mathrm{TN}})}{(C_{\mathrm{FN}} - C_{\mathrm{TP}})},$$

$\mathrm{Prob}_{\mathrm{Norm}}$ is the probability that the patient's condition is normal before the test is performed, $\mathrm{Prob}_{\mathrm{Dis}}$ is the probability that the patient has the disease before the test is performed, C_{FP} is the cost (ie, the financial cost and/or health "cost") of a false-positive result, C_{TN} is the cost of a true-negative result, C_{FN} is the cost of a false-negative result, and C_{TP} is the cost of a true-positive result.

MEASURES OF ACCURACY BASED ON THE ROC CURVE

One of the most popular measures of the accuracy of a diagnostic test is the area under the ROC curve. The ROC curve area can take on values between 0.0 and 1.0. A ROC curve with an area of 1.0 is shown in Figure 3. A test with an area under the ROC curve of 1.0 is perfectly accurate because the sensitivity is 1.0 when the FPR is 0.0. In contrast, a test with an area of 0.0 is perfectly inaccurate. That is, all patients with disease are incorrectly given negative test results and all patients without disease are incorrectly given positive test results. With such a test it would be better to convert it into a test with perfect accuracy by reversing the interpretation of the test results. The practical lower bound for the ROC curve area is then 0.5. The line segment from 0,0 to 1,1 has an area of 0.5; it is called the chance diagonal (Fig 3). If we relied purely on guessing to distinguish patients with from patients without disease, then the ROC curve would be expected to fall along this diagonal line. Diagnostic tests with ROC curve areas greater than 0.5 have at least some ability to discriminate between patients with and those without disease. The closer the ROC curve area is to 1.0, the better the diagnostic test. One method (8) of estimating the area under the empirical ROC curve is described and illustrated in the Appendix. There are other methods (9,10) of estimating the area under the empirical ROC curve and its variance; all of these methods rely on nonparametric statistical methods.

The ROC curve area has several interpretations: *(a)* the average value of sensitivity for all possible values of specificity, *(b)* the average value of specificity for all possible values of sensitivity (11,12), and *(c)* the probability that a randomly selected patient with disease has a test result that indicates greater suspicion than a randomly chosen patient without disease (9).

In Figure 1 the area under the empirical ROC curve for mammography is 0.82; that is, if we select two patients at random—one with breast cancer and one without—the probability is 0.82 that the patient with breast cancer will have a more suspicious mammographic result. The area under the fitted curve is slightly larger at 0.84. When the number of cut points is small, the area under the empirical ROC curve is usually smaller than the area under the fitted curve.

The ROC curve area is a good summary measure of test accuracy because it does not depend on the prevalence of disease or the cut points used to form the curve.

However, once a test has been shown to distinguish patients with disease from those without disease well, the performance of the test for particular applications (eg, diagnosis, screening) must be evaluated. At this stage, we may be interested in only a small portion of the ROC curve. Furthermore, the ROC curve area may be misleading when one is comparing the accuracies of two tests. Figure 4 illustrates the ROC curves of two tests with equal area. At the clinically important FPR range (for example, 0.0–0.2), however, the curves are different: ROC curve A demonstrates higher sensitivity than does ROC curve B. Whenever the ROC curves of two tests cross (regardless of whether or not their areas are equal), it means that the test with superior accuracy (ie, higher sensitivity) depends on the FPR range; a global measure of accuracy, such as the ROC curve area, is not helpful here.

Important point: There are situations where we need a more refined measure of diagnostic test accuracy than the area under the ROC curve.

One alternative is to use the ROC curve to estimate sensitivity at a fixed FPR (or, as appropriate, we could use the FPR at a fixed sensitivity). As an example, in Figure 1 the sensitivity at a fixed FPR of 0.10 is 0.60. This measure of accuracy allows us to focus on the portion of the ROC curve that is of clinical relevance.

Another alternative measure of accuracy is the partial area under the ROC curve. It is defined as the area between two FPRs, e_1 and e_2 (or, as appropriate, the area between two false-negative rates). If $e_1 = 0$ and $e_2 = 1$, then the area under the entire ROC curve is specified. If $e_1 = e_2$, then the sensitivity at a fixed FPR is given. The partial area measure is thus a "compromise" between the entire ROC curve area and the sensitivity at a fixed FPR.

To interpret the partial area, we must consider its maximum possible value. The maximum area is equal to the width of the interval—that is, $e_2 - e_1$ (13). McClish (13) and Jiang et al (14) recommend standardizing the partial area by dividing it by its maximum value. Jiang et al (14) refer to this standardized partial area as the partial area index. The partial area index is interpreted as the average sensitivity for the range of FPRs examined (or the average FPR for the range of sensitivities examined). As an example, in Figure 1, the partial area in the FPR range of 0.00–0.20 is 0.112; the partial area index is 0.56. In other words, when the FPR is between 0.00 and 0.20, the average sensitivity is 0.56.

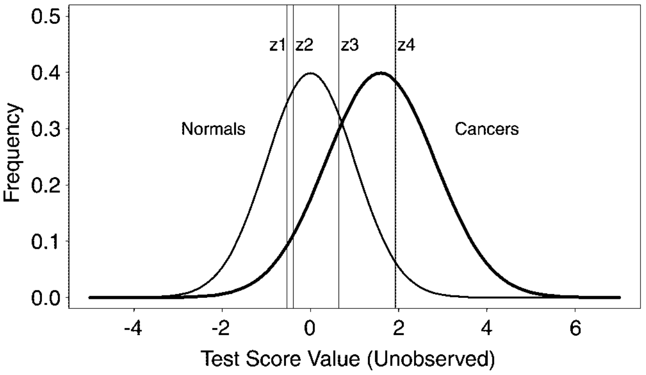


Figure 2. Graph shows the binormal distribution that best fits the mammography study data. By convention, the distribution of unobserved variables for the patients without cancer is centered at zero (ie, $\mu_1 = 0$) with variance (σ_1^2) equal to 1. For these data, the center of the distribution of the unobserved variables for the patients with cancer is estimated to be 1.59 (ie, $\mu_2 = 1.59$) with variance (σ_2^2) estimated to be 1.54. The binormal distribution can be described by its two parameters (4), a and b, as $a = (\mu_1 - \mu_2)/\sigma_2$ and $b = \sigma_1/\sigma_2$. The four cut points z_1, z_2, z_3, and z_4 define the five categories of test results. That is, a variable with a value below the point defined by z_1 indicates a normal result; a variable with a value between z_1 and z_2, a benign result; a variable with a value between z_2 and z_3, a probably benign result; a variable with a value between z_3 and z_4, a suspicious result; and a variable with a value above the point defined by z_4, a malignant result. Note that the binormal variables exist only in the mind of the reader (ie, they are unobserved). When the reader applies the cut points z_1, z_2, z_3, and z_4 to the unobserved variables, we obtain the observed five categories of test results.

EXAMPLES OF ROC CURVES IN RADIOLOGY

There are many examples of the application of ROC curves in radiologic research. I present two examples here. The first example illustrates the comparison of two diagnostic tests and the identification of a useful cut point. The second example describes a multireader study of the differences in diagnostic accuracy of two tests and differences in reader performance.

The first example is the study of Mushlin et al (15) of the accuracy of magnetic resonance (MR) imaging for detecting multiple sclerosis (MS). Three hundred three patients suspected of having MS underwent MR imaging and CT of the head. The images were read separately by two neuroradiologists without knowledge of the clinical course of or final diagnosis given to the patients. The images were scored as definitely showing MS, probably showing MS, possibly showing MS, probably not showing MS, or definitely not showing MS. The reference standard consisted of results of a review of the clinical findings by a panel of MS experts, results of follow-up for at least 6 months, and results of other diagnostic tests; the results of CT and MR imaging were not included to avoid bias.

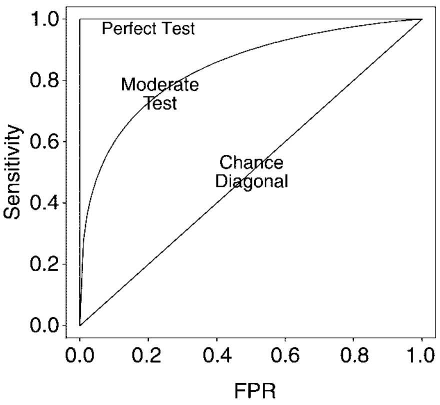


Figure 3. Graph shows comparison of three ROC curves. A perfect test has an area under the ROC curve of 1.0. The chance diagonal has an ROC area of 0.5. Tests with some discriminating ability have ROC areas between these two extremes.

The estimated ROC curve area for MR imaging was 0.82, indicating a good, but not definitive, test. In contrast, the estimated ROC curve area of CT was only 0.52; this estimated area was not significantly different from 0.50, indicating that CT results were no more accurate than guessing for diagnosing MS. The authors concluded that a "definite MS"

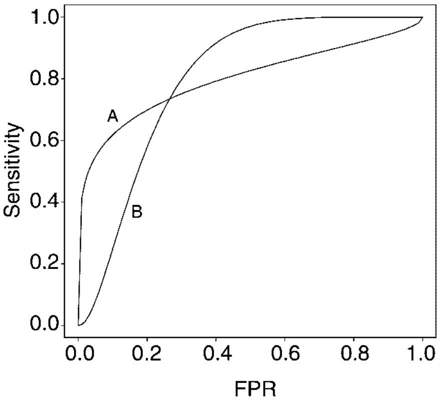


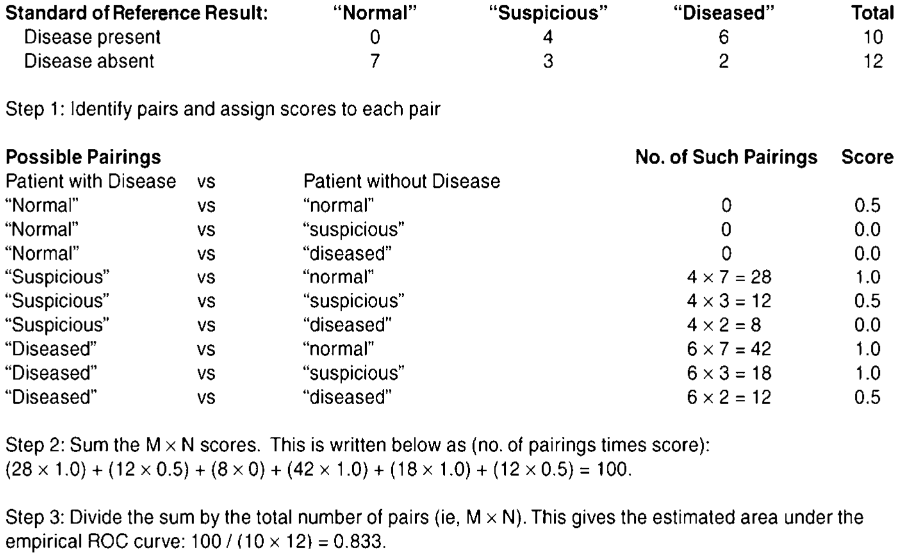
Figure 4. Graph shows two crossing ROC curves. The ROC areas of the two tests are the same at 0.80; however, for the clinically important range (ie, an FPR of less than 0.20), test *A* is preferable to test *B*.

reading at MR imaging essentially establishes the diagnosis of MS (MR images in only two of 140 patients without MS were scored as definitely showing MS, for an FPR of 1%). However, a normal MR imaging result does not conclusively exclude the diagnosis of MS (MR images in 35 of 163 patients with MS were scored as definitely not showing MS, for a false-negative rate of 21%).

In the second example, Iinuma et al (16) compared the accuracy of conventional radiography and digital radiography for the diagnosis of gastric cancers. One hundred twelve patients suspected of having gastric cancer underwent conventional radiography, and 113 different patients with similar symptoms and characteristics underwent digital radiography. Six readers interpreted the images from all 225 patients; the readers were blinded to the clinical details of the patients. The images were scored with a six-category scale, in which a score of 1 indicated that cancer was definitely absent; a score of 2, cancer was probably absent; a score of 3, cancer was possibly absent; a score of 4, cancer was possibly present; a score of 5, cancer was probably present; and a score of 6, cancer was definitely present. The diagnostic standard consisted of the findings of a consensus panel of three radiologists (not the same individuals as the six readers) who examined the patients and were told of the findings of other tests, such as endoscopy and histopathologic examination after biopsy.

The ROC curve areas of the six readers were all higher with digital radiography than with conventional radiography; the average ROC curve areas with digital and conventional radiography were 0.93 and 0.80, respectively. By plotting the fitted ROC curve areas of each of the six readers, the authors determined that for five of the six readers, digital radiography resulted in higher sensitivity for all FPRs; for the sixth reader, digital radiography resulted in considerably higher sensitivity only at a low FPR.

In summary, the ROC curve has many advantages as a measure of the accuracy of a diagnostic test: *(a)* It includes all possible cut points, *(b)* it shows the relationship between the sensitivity of a test and its specificity, *(c)* it is not affected by the prevalence of disease, and *(d)* from it we can compute several useful summary measures of test accuracy (eg, ROC curve area, partial area). The ROC curve alone cannot provide us with the optimal cut point for a particular clinical application; however, given information about the pretest probability of disease and the relative costs of diagnostic test errors, we can find the optimal cut point on the ROC curve. There are many study design issues (eg, patient and reader selection, verification and diagnostic standard bias) that need to be considered when one is conducting and interpreting the results of a study of diagnostic test accuracy. Many of these issues will be covered in a future article.



Standard of Reference Result:	"Normal"	"Suspicious"	"Diseased"	Total
Disease present	0	4	6	10
Disease absent	7	3	2	12

Step 1: Identify pairs and assign scores to each pair

Possible Pairings			No. of Such Pairings	Score
Patient with Disease	vs	Patient without Disease		
"Normal"	vs	"normal"	0	0.5
"Normal"	vs	"suspicious"	0	0.0
"Normal"	vs	"diseased"	0	0.0
"Suspicious"	vs	"normal"	4 × 7 = 28	1.0
"Suspicious"	vs	"suspicious"	4 × 3 = 12	0.5
"Suspicious"	vs	"diseased"	4 × 2 = 8	0.0
"Diseased"	vs	"normal"	6 × 7 = 42	1.0
"Diseased"	vs	"suspicious"	6 × 3 = 18	1.0
"Diseased"	vs	"diseased"	6 × 2 = 12	0.5

Step 2: Sum the M × N scores. This is written below as (no. of pairings times score):
(28 × 1.0) + (12 × 0.5) + (8 × 0) + (42 × 1.0) + (18 × 1.0) + (12 × 0.5) = 100.

Step 3: Divide the sum by the total number of pairs (ie, M × N). This gives the estimated area under the empirical ROC curve: 100 / (10 × 12) = 0.833.

Figure A1. Fictitious data set and example of how to calculate the area under the empirical ROC curve.

APPENDIX

The area under the empirical ROC curve can be estimated as follows: First, consider every possible pairing of patients with disease and patients without disease. Give each pair a score of 1.0 if the test result for the patient with disease is higher (ie, more suspicious for disease), a score of 0.5 if the test results are the same, and a score of 0.0 if the test result for the patient with disease is lower (ie, less suspicious for disease). Second, take the sum of these scores. If there are *N* nondiseased patients and *M* diseased patients in the sample, then there are $M \times N$ scores. Finally, divide the sum of these scores by ($M \times N$). This gives the estimate of the area under the empirical ROC curve.

Figure A1 depicts a fictitious data set. The process described and illustrated in the figure can be written mathematically as follows (8): Let X_j denote the test score of the *j*th patient with disease and Y_k denote the test score of the *k*th patient without disease. Then,

$$A = \frac{1}{(M \times N)} \sum_{(j=1)}^{M} \sum_{(k=1)}^{N} \text{score}(X_j, Y_k),$$

where *A* is the estimate of the area under the empirical ROC curve and score(X_j, Y_k) is the score assigned to the pair composed of the *j*th patient with disease and the *k*th patient without disease. The score equals 1 if X_j is greater than Y_k, equals $\frac{1}{2}$ if X_j is equal to Y_k, and equals 0 if X_j is less than Y_k. The symbol in the following formula

$$\sum_{(k=1)}^{N} c_k$$

is called a summation sign. It means that we take the sum of all of the c_k values, where *k* is from 1 to *N*. So, if *N* is equal to 12, then

$$\sum_{(k=1)}^{N} c_k = c_1 + c_2 + c_3 + \cdots + c_{12}.$$

References

1. Lusted LB. Signal detectability and medical decision-making. Science 1971; 171: 1217–1219.
2. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39:561–577.
3. Campbell G. General methodology I: advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat Med 1994; 13:499–508.
4. Dorfman DD, Alf E Jr. Maximum likelihood estimation of parameters of signal detection theory: a direct solution. Psychometrika 1968; 33:117–124.
5. Dwyer AJ. In pursuit of a piece of the ROC. Radiology 1997; 202:621–625.
6. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging 1989; 29:307–335.
7. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–298.
8. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol 1975; 12:387–415.
9. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837–844.
11. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
12. Metz CE. Some practical issues of experimental design and data analysis in radiologic ROC studies. Invest Radiol 1989; 24: 234–245.
13. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989; 9:190–195.
14. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745–750.
15. Mushlin AI, Detsky AS, Phelps CE, et al. The accuracy of magnetic resonance imaging in patients with suspected multiple sclerosis. JAMA 1993; 269:3146–3151.
16. Iinuma G, Ushio K, Ishikawa T, Nawano S, Sekiguchi R, Satake M. Diagnosis of gastric cancers: comparison of conventional radiography and digital radiography with a 4 million-pixel charge-coupled device. Radiology 2000; 214:497–502.

Primer on Multiple Regression Models for Diagnostic Imaging Research[1]

Ilana F. Gareen, PhD
Constantine Gatsonis, PhD

Index term:
Statistical analysis

Published online
10.1148/radiol.2292030324
Radiology 2003; 229:305–310

Abbreviations:
FTE = full-time equivalent
ICU = intensive care unit
OR = odds ratio

This article provides an introduction to multiple regression analysis and its application in diagnostic imaging research. We begin by examining why multiple regression models are needed in the evaluation of diagnostic imaging technologies. We then examine the broad categories of available models, notably multiple linear regression models for continuous outcomes and logistic regression models for binary outcomes. The purpose of this article is to elucidate the scientific logic, meaning, and interpretation of multiple regression models by using examples from the diagnostic imaging literature.
© RSNA, 2003

Readers of the diagnostic imaging literature increasingly encounter articles with the results of multiple regression analyses. Typically, these analyses are performed to examine the relation of an outcome and several explanatory factors for the purpose of quantifying the effect of the explanatory factors on the outcome and/or predicting the outcome. For example, multiple regression modeling is used to study which combinations of imaging features are important predictors of the presence of disease. In this article and in the statistics literature, the explanatory variables are also referred to as covariates or *independent* variables, and the outcome variable is also referred to as the response or *dependent* variable. If the outcome is represented by a continuous variable, such as cost of care, then linear regression is often used. If the outcome is a dichotomous variable, such as presence or absence of disease, then logistic regression is commonly used. These modeling techniques provide an important tool in medical research. They enhance our ability to disentangle the nature of the relation between multiple factors that affect a single outcome. In this article, we examine why investigators choose to use multiple regression methods and how analyses with these methods should be interpreted. We use examples from the radiology literature as illustrations and focus on the meaning and interpretation of these models rather than on the methods and software used for building them.

WHAT IS REGRESSION ANALYSIS?

Regression analysis provides a quantitative approach to the assessment of the relation between two or more variables, one of which is considered the dependent or response variable, and the others are considered the independent or explanatory variables (also called "covariates"). The purpose of the analysis may be to estimate the effect of a covariate or to predict the value of the response on the basis of the values of the covariates. In both cases, a regression model is developed to predict the value of the response. The way in which a model is built depends on the specific research question and the nature of the data. Each regression model incorporates assumptions regarding the nature of the data. If these assumptions are incorrect, the model may be invalid, and the interpretation of the data that is based on that model may be incorrect. The process of fitting such a model involves the specification of a shape or curve for the expected value of the response and the examination of how closely the data fit this specified shape. For example, the simple linear regression model assumes a straight-line relation between the single independent variable and the expected value of the dependent variable. The slope and intercept of this straight line are estimated from the data. The fitted model can then be used to estimate the effect of the independent variable and can also be used to predict future values of the dependent variable, which correspond to specific values of the independent variable.

[1] From the Center for Statistical Sciences, Brown University, Box G-H, 167 Angell St, Providence, RI 02912. Received February 19, 2003; revision requested March 25; revision received May 9; accepted May 20. Supported in part by National Cancer Institute grant CA79778. Address correspondence to I.F.G. (e-mail: *igareen@stat.brown.edu*).

WHY ARE MULTIPLE REGRESSION MODELS USED IN DIAGNOSTIC IMAGING?

Multiple Factors of Interest

Multiple regression analyses may be used by investigators to examine the impact of multiple factors (independent variables) on a single outcome of interest (dependent variable). Sunshine and Burkhardt (1), for example, assembled survey data from 87 radiology groups on practice patterns. The authors used these data to examine the relation between the number of procedures each radiologist performed per year (dependent variable) and several independent variables, including academic status of the group, annual hours worked by each full-time radiologist, group size, and the proportion of “high-productivity” procedures (Table 1). High-productivity procedures are defined by the authors as those that require more mental effort, stress, physical effort, and training than do other types of procedures. Such procedures include CT, MR, and interventional or angiographic procedures. On the basis of the results in Table 1, it appears that workload, as measured by the dependent variable, is significantly lighter in academic groups than that in non-academic groups and decreases marginally with increasing group size. In this model, it appears that workload is not associated with the other two factors, namely group size and proportion of high-productivity procedures. The authors also report that this regression model explains only a modest amount of the variability in the dependent variable and “did not yield very accurate results.” We return to the interpretation of Table 1 later in this article.

TABLE 1
Results of Multiple Linear Regression Analysis to Examine the Number of Annual Procedures per FTE Radiologist in Diagnostic Radiology Groups

Variable	Coefficient (β)	Standard Error*	*P* Value
Intercept (β_0)	10,403	2,154	.001
Academic status (X_1)	−2,238	1,123	.05
Annual hours per FTE (X_2)	0.43	1.11	.70
Group size (FTE) (X_3)	−59.7	32.5	.07
Proportion of high productivity procedures (X_4)†	−4,782	11,975	.69

Note.—Adapted and reprinted, with permission, from reference 1.

* Standard error of the estimated coefficient.

† High-productivity procedures included computed tomography (CT) and magnetic resonance (MR) imaging, and interventional or angiographic procedures that required more mental effort, stress, physical effort, and training than did other types of procedures.

Adjustment for Potential Confounding

Multiple regression techniques may also be used to adjust analyses for potential confounding factors so that the influence of these extraneous factors is quantified and removed from the evaluation of the association of interest. Extraneous or confounding factors are those that are associated with both the exposure and the outcome of interest but are not consequences of the exposure. Consider, for example, a study to compare two diagnostic procedures (independent variable) on the basis of their impact on a patient outcome (death within a fixed follow-up period in this example). A factor such as patient age may play a role in decisions about which procedure will be performed but may also be related to the outcome. In this case, age would be a potential confounder. Clearly, confounding is a major consideration in observational studies as contrasted with randomized clinical trials, because in the former, participants are not randomly assigned to one imaging modality or another.

Goodman et al (2) compared results in patients evaluated for suspected pulmonary embolism with helical CT with those evaluated with ventilation-perfusion scintigraphy to determine whether there is a difference in the number of pulmonary embolisms and deaths in the 90 days following the diagnostic imaging evaluation. The population of patients evaluated with CT was more likely to have been referred from the intensive care unit (ICU) (and, hence, more likely to have severe co-morbid disease conditions), was older, and was at increased risk of pulmonary embolism due to patient immobilization and patient malignancy than were those evaluated with ventilation-perfusion scintigraphy. To adjust for these potential confounders, the authors included them as independent variables in a logistic regression analysis and evaluated the association between imaging methods and death within 90. As a result of this adjustment, the magnitude of the estimated effect of imaging on mortality, as measured by means of the OR, changed from 3.42 to 2.54. We will return to this point later in the discussion of logistic regression.

Prediction

Regression models are also used to predict the value of a response variable using the explanatory variables. For example, to develop optimal imaging strategies for patients after trauma, Blackmore et al used clinical data from trauma patients seen in the emergency room to predict the risk of cervical spine fracture (3). The authors evaluated 20 potential clinical predictors of cervical spine injury. Their final prediction model includes four of these factors: the presence of focal neurologic deficit, presence of severe head injury, cause of injury, and patient age. These independent variables predict cervical spine fracture with a high degree of accuracy (area under the receiver operating characteristic curve = 0.87) (3).

GENERAL FORM OF REGRESSION MODELS

Regression models with multiple independent variables have been constructed for a variety of types of response variables, including continuous and discrete variables. A large class of such models, and the models used most commonly in the medical literature, are the so-called generalized linear models (4). In these models, a linear relation is postulated to exist between the independent variables and the expected value of the dependent variable (or some transformed value of that expected value, such as the logarithm). The observed value of the dependent variable (response) is then the sum of its expected value and an error term. Multiple regression models for continuous, binary, and other discrete dependent variables are discussed in the following sections.

MODELING OF CONTINUOUS OUTCOME DATA

First, let us consider the situation in which a continuous dependent variable and a single independent variable are available. Such would be the case, for example, if in the practice pattern data discussed earlier, only the number of annual procedures per radiologist as the dependent variable and the radiology group size as the independent variable were considered. An earlier article in this series (5) introduced the concept of a simple linear regression model in which a linear relation is assumed between the mean of the response and the independent variable. This model is represented as $Y_i = \beta_0 + \beta_1 X_{1i} + e_i$, where Y_i is the term representing the value of the dependent variable for the ith case, X_{1i} represents the value of the independent random variable, β_0 represents the intercept, β_1 represents the slope of the linear relation between the mean of the dependent and the independent variables, and e_i denotes the random error term, which has a mean of zero. The expected value of the dependent variable is then equal to $\beta_0 + \beta_1 X_{1i}$, and the error term is what is left unexplained by the model.

When several independent variables are considered, such as in the analysis of the practice pattern data, multiple regression models are used. For example, assume that in addition to X_1, independent variables $X_{2,} \ldots, X_p$ are to be included in the analysis. A *linear* multiple regression model would then be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + e_i.$$

The parameters $\beta_0, \beta_1, \ldots, \beta_p$ from this equation are referred to as the regression coefficients. To interpret the coefficients, again consider first the simple linear regression model. In this model, the parameter β_0 represents the intercept, that is, the expected value of the *dependent* variable when the value of the *independent* variable is set to zero. The parameter β_1 represents the slope of the regression line and measures the average change in the dependent variable Y that corresponds to an increase of one unit in the independent variable X_1.

In multiple regression, the relation between the dependent variable Y and the independent variables $X_{1,} \ldots, X_p$ is somewhat more complex. The intercept β_0 represents the mean value of the response when *all* of the independent variables are set to zero (that is, $X_1 = 0$, $X_2 = 0, \ldots, X_p = 0$).

The slopes of the independent variables in the multiple linear regression model are interpreted in the following way. The slope β_j of the jth independent variable measures the change in the dependent variable that corresponds to an increase of one unit in X_j, if all other independent variables are held fixed (that is, the values of the other covariates do not change). For example, the results of a multiple linear regression analysis of the practice pattern data reported by Sunshine and Burkhardt (1) are shown in Table 1. From this survey of 87 radiology groups, the dependent variable Y is the number of procedures performed annually per full-time equivalent (FTE) radiologist. X_1 is a dichotomous variable and an indicator of academic status (coded 1 if the group is academic or 0 if the group is non-academic). The remaining variables can be treated as approximately continuous. X_2 is the number of annual hours worked by each FTE radiologist, X_3 is the practice group size, and X_4 is the percentage of procedures that are high productivity. Suppressing the notation for cases, this model is written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4.$$

The terms can be substituted into this model such that it is represented as the following equation: $Y = \beta_0 + \beta_1$ (academic status) $+ \beta_2$ (annual hours per FTE) $+ \beta_3$ (group size) $+ \beta_4$ (percentage high-productivity procedures).

The estimated coefficient of X_1, the indicator of academic status in the regression model, is −2,238 (Table 1). Because X_1 takes the values of 1 (academic group) or 0 (non-academic group), this coefficient estimate implies that, if all other independent variables remain fixed, academic groups would, on an annual basis, be expected to have 2,238 procedures per FTE radiologist less than those performed in non-academic groups (the number decreases because the coefficient is a negative number).

The interpretation of coefficients for continuous independent variables is similar. For example, the model estimates that, if all other independent variables were fixed, an increase of one unit in group size would correspond to an average decrease of 59.7 in the number of procedures performed annually by each FTE radiologist in a group practice. Thus, if all other independent variables remained fixed, and practice size increased by five, the number of procedures per FTE radiologist would be expected to decrease by 5 × 59.7 = 298.5, and so on. One caveat in the interpretation of coefficients is that it is not always possible to give them a direct physical interpretation. In this example, the intercept term in the model does not have a direct interpretation because it corresponds to a setting of all the independent variables to zero, which would be impossible to do. It may also be argued that it is not possible to fix some of the independent variables, such as annual hours per FTE radiologist, while allowing others, such as practice size, to vary.

In Table 1, the standard error for each coefficient provides a measure of the degree of statistical uncertainty about the estimate. The fitting of models to data with a lot of scatter and small sample sizes can lead to large standard errors for the estimated coefficients. The standard error can be used to construct a confidence interval (CI) for the coefficient. The P values in Table 1 correspond to tests of the null hypothesis that a particular coefficient is equal to zero (that is, the hypothesis of "no association" between the particular independent variable and the dependent variable).

MODELING OF DICHOTOMOUS OUTCOMES

Logistic regression is commonly used to analyze dichotomous outcomes (dependent variable). The independent variables in these models may be continuous, categoric, or a combination of the two. For simplicity, let us assume that the dichotomous dependent variable is coded as 0 or 1. For example, a dichotomous outcome of interest is whether each patient is dead or alive at the end of the study observation period: $Y = 1$ if a patient died during the follow-up interval or $Y = 0$ if a patient was alive at the end of the follow-up interval.

In the logistic model, the expected value of the response Y is equal to the probability that $Y = 1$, that is, the probability that an event (such as death) occurs. The form of the model, however, is more complex than that in the linear model for continuous responses. In particular, the *logit* of the expected value, rather than the expected value of Y, is assumed to be a linear function of the covariates. If p denotes the probability that an event will occur, the logit of p is defined as the logarithm of the odds, that is, logit $p = \log[p/(1 - p)]$.

Formally, the logistic model with multiple independent variables is written as

$$\text{logit}\ p(Y = 1) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

or, equivalently, as

$$[p(Y = 1)] = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}.$$

In the logistic model, β_j measures the change in log-odds for $Y = 1$ that corresponds to an increase of one unit in X_j, if all of the other independent variables remain fixed. In contrast to the linear model for continuous responses, the corresponding change in actual odds is multiplicative. Hence, $\exp(\beta_j)$ measures the odds ratio (OR) that corresponds to an increase of one unit in X_j. The OR is a frequently used measure of association in the epidemiology literature and is a common way of expressing the logistic regression results (6). The OR measures the odds of an outcome in the index group compared with the odds of the same outcome in the comparison group.

TABLE 2
Results of Multiple Logistic Regression Analysis to Examine Death within 90 Days of Evaluation for Pulmonary Embolism

Variable	Regression Coefficient (β)	SD	Odds Ratio	95% CI	*P* Value
Underwent CT (X_1)	0.93	0.32	2.54	1.36, 4.80	.004
Referral from intensive care unit (X_2)	1.78	0.32	5.93	3.09, 11.0	.001
Age older than 67 years (X_3)	0.75	0.34	2.12	1.12, 4.14	.024
Immobilization (X_4)	1.26	0.39	3.52	1.59, 7.58	.002
Malignancy (X_5)	0.87	0.34	2.39	1.21, 4.63	.012

Note.—Adapted and reprinted, with permission, from reference 2.

For example, in the study of Goodman et al (2), Y indicates whether the patient is dead ($Y = 1$) or alive ($Y = 0$) 3 months after admission for pulmonary embolism work-up. The primary independent variable of interest, diagnostic method, would be represented by X_1. Potential confounders (covariates) would be represented by $X_2, \ldots, X_p$. In Table 2, X_2 is an indicator that the patient was referred from an ICU, X_3 is an indicator that the patient was older than 67 years, X_4 is an indicator of immobilization, and X_5 is an indicator of malignancy.

Substitution of these covariates into this model would result in the following representation: logit $P(Y = 1) = \beta_0 + \beta_1$ (underwent CT: yes/no) $+ \beta_2$ (ICU referral: yes/no) $+ \beta_3$ (age older than 67 years: yes/no) $+ \beta_4$ (immobilization: yes/no) $+ \beta_5$ (malignancy: yes/no).

For X_1, the indicator of whether or not helical CT was performed, the estimate of the coefficient was 0.93. Therefore, the estimated odds of death among patients who were evaluated with helical CT compared with that among those who were evaluated with lung scintigraphy was $\exp(0.93) = 2.54$. That is, if all other independent variables were fixed, the odds of death within 90 days for patients who underwent CT to diagnose pulmonary embolism were 2.54 times as high as the odds for patients who underwent lung scintigraphy to diagnose pulmonary embolism. Note that this value represents an OR that was "adjusted" for the presence of potential confounders. The "unadjusted" estimate (computed from data presented in reference 2) was 3.42. Because we cannot know the counterfactual occurrence (the number of patients evaluated with CT who would have died had they been evaluated with ventilation-perfusion scintigraphy), we cannot say whether the adjustment was successful, and the OR is unbiased. That there is some difference between the unadjusted OR (3.42) and the adjusted OR (2.54) provides an indication that the potential confounders controlled for in the analysis may have been confounding the association between imaging modality and death within 90 days. However, a strong association remains between imaging method and risk of death. A CI for the OR can be obtained (6). In the example, the 95% CI for the OR is (1.36, 4.80).

The authors (2) report that "the patients in the CT imaging group had more than twice the odds of dying within 90 days as those in the [ventilation-perfusion] scintigraphy group." They also noted that "the prevalence of clinically apparent pulmonary embolism after a negative helical CT scan was low (1.0%) and minimally different from that after a normal ventilation-perfusion scan (0%)" (2). Part of this association may be due to residual confounding in the analysis. In particular, it is likely that there was confounding by indication in this sample. That is, patients with a higher likelihood of dying from pulmonary embolism were referred selectively for CT. Other more sophisticated statistical techniques may be needed to adjust for this type of confounding (7).

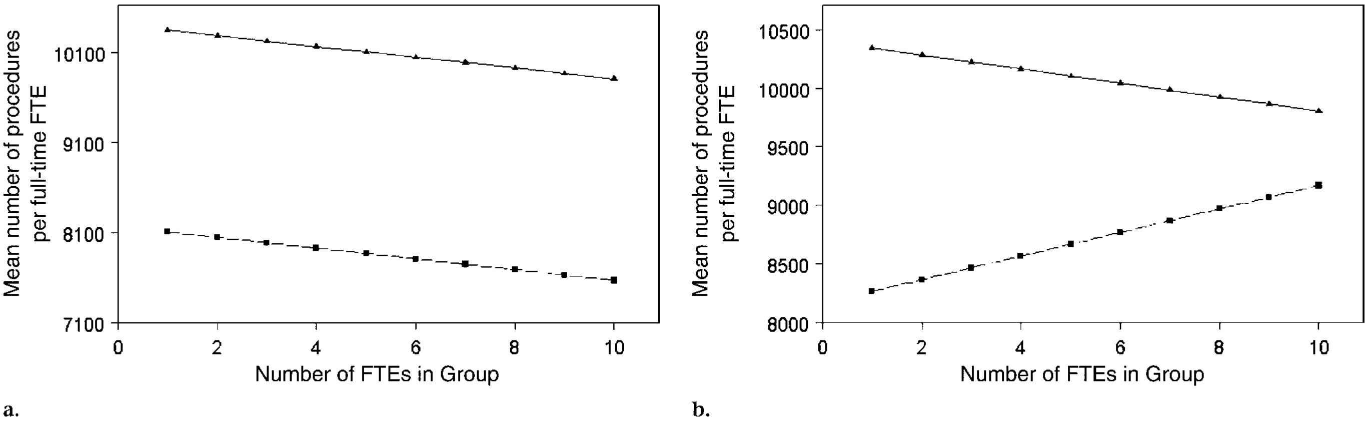
In multiple logistic regression models, the intercept β_0 measures the baseline log-odds for $Y = 1$, that is, the log-odds for $Y = 1$ for cases in which all independent variables have a value of zero. In the pulmonary embolism example, this would correspond to the subset of patients with all independent variables set to "no," that is, patients who did not undergo CT, were not referred from the ICU, were 67 years old or younger, were not immobilized, and did not have a malignancy. Note that if all covariates are centered by means of subtraction of the average population value, then β_0 measures the log-odds for $Y = 1$ for an "average" case.

POLYNOMIAL TERMS

The models discussed earlier assumed a linear relation between the independent variables and the expected value of the dependent variable. If the relation is thought to follow a non-linear form, alternative models can be considered that involve transformations of the dependent and/or independent variables. Herein, we discuss transformations of the independent variables. In a simple model with a continuous dependent variable and a continuous independent variable, if the slope of the relation appears to change with the value of the independent variable X, then a polynomial in X may be used instead of a straight line. With the example from Sunshine and Burkhardt (Table 1), if the association between the average number of procedures per radiologist and group size was not linear but seemed to be parabolic in nature, with extremes in each tail, then inclusion of a term X_3^2 might more fully describe the observed data. The addition of higher-order (ie, X^3, X^4) terms may also enhance model fit (8). In addition to polynomial functions, models with other non-linear functions of the independent variables are available (8).

MODEL INTERPRETATION: INTERACTIONS

In both linear and logistic regression, the association between the dependent variable and one of the independent variables may vary across values of another independent variable. To graphically depict this concept, the example from the Sunshine and Burkhardt article (1) is used. The relation between the number of procedures per FTE radiologist and group size for academic and nonacademic groups with no interaction terms is shown in the Figure, part a. Note that



Examples of fitted regression lines of the relation between the number of procedures per FTE radiologist and group size for academic and nonacademic groups, based on the analyses presented by Sunshine and Burkhadrt (1), show **(a)** no statistical interaction and **(b)** statistical interaction.

TABLE 3
Results of Multiple Regression Analysis to Examine Coronary Restenosis

Variable*	Odds Ratio	95% CI	*P* Value
Intercept coefficient (β_0) = 0.12	. . .	. . .	. . .
Stent use (X_1)	0.83	0.72, 0.97	.0193
Lesion length (X_2)	1.05	1.04, 1.06	<.001
PMLD (X_3)	0.53	0.46, 0.61	<.001
Previous CABG (X_4)	0.69	0.53, 0.9	.006
Diabetes mellitus (X_5)	1.33	1.16, 1.54	<.001
Stent use * PMLD (X_6)*	0.34	0.31, 0.39	.002

Source.—Reference 9.

Note.—Table presents results from a multiple logistic regression analysis to examine coronary restenosis as a function of medical treatment and other selected patient characteristics. CABG = coronary artery bypass graft, PMLD = post-procedural maximum lumen diameter.

* In this term, the * indicates that this is an interaction term in which stent use is multiplied by PMLD.

the two lines, which correspond to academic and nonacademic group status, are parallel. Now, suppose that the authors want to examine whether, in fact, the two lines are parallel or not. In other words, they want to know whether the relation between the number of procedures and group size depends on academic group status. To examine this question, the authors would consider a model with an interaction term between academic group status and group size. The Figure, part b, shows how statistical interaction with another variable (academic status) might influence the relation between the number of procedures and group size.

Another example is drawn from the article by Mercado et al (9). In this article, the authors examine whether placement of a stent influences the association between post-procedural minimal lumen diameter and restenosis (Table 3). Questions of this type can be addressed by including appropriate "interaction" terms in the regression model. In the restenosis data, a model with an interaction between post-procedural lumen diameter and stent use can be written as follows:

$$\text{logit}\, p(Y = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1 * X_3,$$

where $Y = 1$ if restenosis occurs and 0 otherwise, X_1 = stent use, X_2 = lesion length, X_3 = post-procedural maximum lumen diameter (PMLD), X_4 = previous coronary artery bypass graft (CABG), and X_5 = diabetes mellitus. $X_1 * X_3$ is a (multiplicative) interaction term between X_1 and X_3, in which * indicates that X_1 is multiplied by X_3. With the addition of the interaction term, the model would be represented as follows: logit $P(Y = 1) = \beta_0 + \beta_1$ (stent use) + β_2 (lesion length) + β_3 (PMLD) + β_4 (previous CABG) + β_5 (diabetes mellitus) + β_6 (stent use * PMLD).

The presence of a significant interaction suggests that the effect of X_1 depends on the actual level of X_3 and conversely. For example, the OR for the maximal diameter size would be $\exp(\beta_3 + \beta_6 X_1)$. Thus, for patients who did not receive a stent, an increase of one unit in the maximal diameter would multiply the odds of restenosis by $\exp(\beta_3)$. However, for patients who received a stent, the odds of restenosis would be multiplied by $\exp(\beta_3 + \beta_6)$. Hence, *in the presence of interactions, the main effects cannot be interpreted by themselves* (4,6,7).

OTHER FORMS OF MULTIPLE REGRESSION MODELING

Dependent variables that are neither continuous nor dichotomous may also be analyzed by means of specialized multiple regression techniques. Most commonly seen in the radiology literature are ordinal categoric outcomes. For example, in receiver operating characteristic studies, the radiologist's degree of suspicion about the presence of an abnormality is often elicited on the five-point ordinal categoric scale, in which 1 = definitely no abnormality present, 2 = probably no abnormality present, 3 = equivocal, 4 = probably abnormality present, and 5 = definitely abnormality present. Ordinal regression models are available for the study of ordinal categoric outcomes. Such models can be used to fit receiver operating characteristic curves and to estimate the effect of covariates such as patient, physician, or other factors. Examples and further discussion of ordinal regression are available in articles by Tosteson and Begg (10) and Toledano and Gatsonis (11).

RECENTERING AND RESCALING OF VARIABLES

As noted earlier, in some cases an independent variable cannot possibly take the value of 0, thus making it difficult to interpret the intercept of a regression model. For example, gestational age and age at menarche cannot be meaningfully set to zero. This difficulty can be addressed by subtracting some value from the independent variable before it is used in the model. In practice, the average value of the independent variable is often used, and the "centered" form of the variable now represents the deviation from that average. When independent variables are centered at their averages, the intercept represents the expected response for an "average" case, that is, a case in which all independent variables have been set to their average values.

The rescaling of variables may also enhance the interpretability of the model. Often the units in which data are presented are not those of clinical interest. By rescaling variables, each unit of increase may represent either a more clinically understandable or a more meaningful difference. For example, if gestational age is measured in days, then it may be rescaled by dividing the value for each observation by seven, which yields gestational age measured in weeks. In this case, β_1, the regression coefficient, would then represent the difference in risk per unit increase in gestational age in weeks rather than in days.

MODEL SELECTION

A detailed discussion of model selection is beyond the scope of this article. We note, however, that selection of the independent variables to include in a model is based on both subject matter and formal statistical considerations. Generally, certain independent variables will be included in the model even if they are not significantly associated with the response because they are known a priori to be related to both the exposure and the outcome of interest or to be potential confounders of the association of interest. Additional independent variables of interest are then evaluated for their contribution to an explanation of the observed variation. Models are sometimes built in a forward "stepwise" fashion in which new independent variables are added in a systematic manner, with additional terms being entered only if their contribution to the model is above a certain threshold. Alternatively, "backward elimination" may be used, starting with all potential independent variables of interest and then sequentially deleting covariates if their contribution to the model is below a fixed threshold. The validity and utility of stepwise procedures for model selection is a matter of debate and disagreement in the statistics literature (6).

In addition to the selection of pertinent independent variables for inclusion in the model, it is essential to ensure that the form of the model is appropriate. A variety of regression diagnostics are available to help the analyst determine the adequacy of the postulated form of the model. Such diagnostics generally focus on examination of the residuals, which are defined as the difference between the observed and the predicted values of the response. The analyst then examines the residuals to detect the presence of patterns that suggest poor model fit (4,8).

CONCLUSION

Multiple regression models offer great utility to radiologists. These models assist radiologists in the examination of multifactorial etiologies, adjustment for multiple confounding factors, and development of predictions of future outcomes. These models are adaptable to continuous, dichotomous, and other types of data, and their use may enhance the radiologist's understanding of complex imaging utilization and clinical issues.

References

1. Sunshine JH, Burkhardt JH. Radiology groups' workload in relative value units and factors affecting it. Radiology 2000: 214:815–822.
2. Goodman LR, Lipchik RJ, Kuzo RS, Liu Y, McAuliffe TL, O'Brien DJ. Subsequent pulmonary embolism: risk after a negative helical CT pulmonary angiogram—prosepctive comparison with scintigraphy. Radiology 2000; 215:535–542.
3. Blackmore CC, Emerson S, Mann F, Koepsell T. Cervical spine imaging in patients with trauma: determination of fracture risk to optimize use. Radiology 1999; 211:759–765.
4. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied regression analysis and other multivariable methods. 3rd ed. Boston, Mass: Duxbury, 1998.
5. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology 2003; 227:617–628.
6. Hosmer DW, Lemeshow S. Applied logistic regression. New York, NY: Wiley, 1989.
7. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998; 17:2265–2281.
8. Neter J, Kutner M, Nachtscheim C, Wasserman W. Applied linear statistical models. 4th ed. Chicago, Ill: Irwin/McGraw-Hill, 1996.
9. Mercado N, Boersma E, Wijns W, et al. Clinical and quantitative coronary angiographic predictors of coronary restenosis: a comparative analysis from the balloon-to-stent era. J Am Coll Cardiol 2001; 38: 645–652.
10. Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. Med Decision Making 1988; 8:204–215.
11. Toledano A, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. Stat Med 1995 15:1807–1826.

Special Topics III: Bias[1]

Nancy A. Obuchowski, PhD

Index term:
Statistical analysis

Published online before print
10.1148/radiol.2293010899
Radiology 2003; 229:617–621

Researchers, manuscript reviewers, and journal readers should be aware of the many potential sources of bias in radiologic studies. This article is a review of the common biases that occur in selecting patient and reader samples, choosing and applying a reference standard, performing and interpreting diagnostic examinations, and analyzing diagnostic test results. Potential implications of various biases are discussed, and practical approaches to eliminating or minimizing them are presented.

There are many potential sources of bias in radiologic studies. For those of us who perform studies, review manuscripts, or read the literature, it is important to be aware of these biases, and for the investigators in studies, it is important to know how to avoid or minimize bias. The term *bias* refers to the situation in which measurements from a study (eg, measurement of a test's sensitivity and specificity) do not correspond to the values that we would obtain if we performed the test in all patients in the relevant population. Of course, we can never perform the test in all patients in the population, so it is imperative that we do our best to design studies without bias.

Bias can occur in selecting the patients, images, and/or readers (ie, radiologists) for a study, in choosing and applying the reference-standard procedure, in performing and interpreting the tests, and in analyzing the results. Many of the biases encountered in radiologic studies have been given names, but there are many unnamed biases that we can identify and avoid by using common sense.

It is best to recognize the potential sources of bias while in the process of designing a study. Then, solutions to the bias problem, or at least ways to minimize the effect of the bias, can be implemented in the study. Note that having a large sample size may reduce the variability (ie, random error) (Table) of our estimates, but it is never a solution to bias (ie, systematic error).

BIAS IN SELECTING THE PATIENT SAMPLE

The objectives of a study determine the type of patients that we recruit. If we have a new test and want to determine if it has any diagnostic value, then we might select a group of patients with clinically evident disease and a group of volunteers who do not have the disease for comparison. Sox et al (3) refer to the patients in such studies as "the sickest of the sick" and "the wellest of the well." If the test does not yield different results for these two groups, then it probably does not have any diagnostic value. In these patients, we cannot measure, without bias, other variables such as the test's sensitivity and specificity or the differences in contrast material uptake. The reason is that our study sample is missing important types of patients—for example, patients with latent disease and control patients with confounding illnesses. Variables such as sensitivity and specificity will most likely be different for these patients.

Selection bias occurs when external factors influence the composition of the sample to the extent that the sample does not represent the population (eg, in terms of patient types, the frequency of the patient types, or both). Spectrum bias (4) is a type of selection bias; it exists when the sample is missing important subgroups. A classic example of spectrum bias is that encountered in screening mammography studies to compare the accuracy of full-field digital mammography with that of conventional mammography. A very large sample size is required to perform a comparison of these two modalities because the prevalence of breast cancer in screening populations is very low. One strategy to reduce the sample size is to consider women who have positive conventional mammography results. These women return for biopsy, and at that time, full-field digital mammography can be performed. However, there is a serious problem with this strategy: The patients with

[1] From the Departments of Biostatistics and Epidemiology and Radiology/Wb4, the Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195. Received May 8, 2001; revision requested June 4; revision received June 21; accepted July 9. **Address correspondence to** the author (e-mail: *nobuchow@bio.ri.ccf.org*).

negative conventional mammography results (true- and false-negative cases) have been selected out. The consequence is that the sensitivity of conventional mammography will be greatly overestimated—making full-field digital mammography seem inferior—and the specificity of conventional mammography will be greatly underestimated—making full-field digital mammography seem superior.

Definition of Common Terms

Term	Definition
Random error	Variation in measurements due to inherent differences between patients (or readers) and natural fluctuations within a patient (or reader)
Systematic error	Pattern of variation in measurements attributable to an external factor
Generalizeable	Situation in which a study's results can be assumed to represent and/or predict the situation at another clinical center (1)
Operational standards	Set of definitions and/or rules used to conduct a research study (eg, definitions of presence and absence of disease)
Misclassification	Incorrect diagnosis; examples are false-positive cases (ie, disease-free patients classified as having disease) and false-negative cases (ie, patients with disease classified as being disease free)
Blinding	Process of withholding information (eg, results of the reference standard procedure) from a technician and/or a reader for the study purpose of determining the value (eg, accuracy) of a test per se
Randomize	Process of assigning patients (or images) to groups or positions in a list (eg, a list of images to be read in a study) by using various methods that allow each patient (or image) to have a known (usually equal) chance of being assigned to a group or position, but the group or position cannot be predicted (2)

Once a new diagnostic test is shown to be capable of yielding different results for "the sickest of the sick" and "the wellest of the well," it is time to challenge the test. We challenge a test by performing it in study patients for whom making a diagnosis is difficult (4). From the results of these studies, we can determine if the test will be reliable in a clinical population that includes both patients for whom it is easy and patients for whom it is difficult to make a diagnosis. However, because of spectrum bias, we still cannot measure, without bias, other variables such as the test's sensitivity and specificity.

Suppose now that we have a well-established test that we know from previous studies is reliable even for difficult-to-diagnose cases. We want to measure, for example, the test's sensitivity and specificity for a particular population of patients. Ideally, we would select our study patients by taking a random sample from the population of patients who present to their primary physician with a certain set of signs and symptoms. In fact, a random sample is the basis of the interpretation of *P* values calculated in statistical analyses. We then perform the well-established test in these patients and measure the test's sensitivity and specificity. These measurements will be generalizable (Table) to similar patients who present to their primary physicians with the same signs and symptoms.

Sometimes this ideal study design is not workable. Alternatively, for this well-established test, suppose we select our study patients from a population of individuals who are referred to the radiology department for the test. Such a sample is called a referred or convenience sample. These patients have been selected to undergo the test. Other patients from the population may not have been referred for the test, or they may have been referred at a different rate. It is usually impossible to determine the factors that influenced the evaluating physicians' referral patterns. Thus, the measurements taken from a referred sample are generalizable only to the referring physicians in the study since other physicians will select different patients.

If we must use a referred sample, for example, to minimize costs, then we should at least carefully collect and record important patient characteristics—important in the sense that the measurements taken in the study might vary according to these characteristics—and the relative frequency of these characteristics. We should report the measurements obtained in patients with various characteristics (eg, report the test's sensitivity and specificity for patients with and those without symptoms). This will allow others to compare the characteristics of their patient population with the characteristics of the study sample to determine how generalizable the study results are to their radiology practice.

BIAS IN SELECTING THE READER SAMPLE

In some studies a sample of readers is needed. For example, when studying the diagnostic accuracy of tests such as chest radiography, computed tomography (CT), magnetic resonance (MR) angiography, and mammography, we must recognize that accuracy is a function of both the imaging unit and the reader who uses it (5). Since readers differ in cognitive and perceptual abilities, it is important to include multiple readers in such studies and to make sure that these readers represent the population of radiologists in whom you are interested. Too often radiologic research is performed at tertiary care hospitals by radiology subspecialists who are experts in their given specialties. Thus, the reported estimates of diagnostic test accuracy may be high, but they might not be generalizable to community hospitals where general radiologists practice.

It can be challenging to obtain a truly representative sample of readers for studies. The problem is illustrated in the mammography study performed by Beam et al (6). They identified all of the American College of Radiology–accredited mammography centers in the United States. There were 4,611 such centers in the United States at the time of the study. Then they randomly sampled 125 of the 4,611 centers and mailed letters to these centers to assess their willingness to participate. Only 50 centers (40%) agreed to take part in the study. One hundred eight radiologists from these 50 centers actually interpreted images for the study. There was a clear potential for bias because the highly motivated centers and readers may have been more likely to volunteer, and these centers and readers may not have been representative of the population. It is unclear how to overcome this type of bias.

BIAS IN CHOOSING AND APPLYING THE REFERENCE-STANDARD TEST

This section is focused on studies in which the objective is to measure the accuracy (ie, sensitivity, specificity, and receiver operating characteristic curve) or comparative accuracy of tests. For these studies, a reference-standard, or "gold standard," procedure is needed to determine the true disease status of the patients. A reference-standard procedure is a test or procedure whose results tell us, with nearly 100% accuracy, the true disease status of patients. Choosing a reference-standard procedure and applying it equitably is often the most challenging part of designing a study.

Imperfect standard bias occurs when the reference-standard procedure yields results that are not nearly 100% accurate. An example would be a study of the accuracy of head CT for the diagnosis of multiple sclerosis. If MR imaging were used as the reference-standard test, then the measures of the accuracy of CT would be biased (ie, probably too low in value) (7) because the accuracy of MR imaging in the diagnosis of multiple sclerosis is not near 100%.

Some might argue that there is no such thing as a "gold" standard. Even pathologic analysis results are not 100% accurate because, like radiology, pathology is an interpretative discipline. For all studies it is important to have operational standards (Table) that take into account the condition being studied, the objectives of the study, and the potential effects of any bias. Some common sense is needed as well (8).

There are various solutions to imperfect standard bias. First, we can choose a better reference-standard procedure, if one exists. For the multiple sclerosis study, we could follow up the patients for several months or years to establish a clinical diagnosis and use the follow-up findings as the reference standard for comparison with the results of CT. Sometimes, however, there is no reference-standard procedure. For example, suppose we want to estimate the accuracy of a new test for identifying the location in the brain that is responsible for epileptic seizures. There is no reference-standard test in this case. However, as an alternative to measuring the test's accuracy, we could frame the problem in terms of the clinical outcome (7): We could compare the test results with the patients' seizure status after nerve stimulation to various locations and report the strength of this relationship. Such analysis can yield useful clinical information, even when the test's accuracy cannot be adequately evaluated.

Another solution is to use an expert panel to establish a working diagnosis. Thornbury et al (9) formed an expert panel to determine the diagnoses for patients who underwent MR imaging and CT for acute low back pain. The panel was given the patients' medical histories, physical examination results, laboratory findings, treatment results, and follow-up information to decide whether a herniated disk was present. The determinations of the expert panel regarding the patients' true diagnoses were used as the reference standards with which the MR imaging and CT results were compared. Note that the expert panel was not given the results of MR imaging or CT. This was planned to avoid incorporation bias, which occurs when the results of the diagnostic test(s) under evaluation are incorporated—in full or in part—into the evidence used to establish the definitive diagnosis (4).

A fourth solution to imperfect standard bias is to apply one of several statistical corrections (10). To apply these corrections, one must make some assumptions about the imperfect reference standard (eg, that its sensitivity and specificity are known) and/or the relationship between the results of the test being assessed and the results of the reference-standard test (eg, that the test in question and the reference-standard test make errors independently of one another). There is continuing research of new statistical methods for addressing imperfect standard bias.

In some studies, a reference-standard procedure exists, but it cannot be performed in all of the study patients, usually owing to ethical reasons. An example of such bias is that which may be encountered in a study to assess the accuracy of lung cancer screening with CT. If a patient has negative CT results, then we cannot perform biopsy or surgery to determine his or her true disease status. Verification bias occurs when patients with positive or negative test results are preferentially referred for the reference-standard procedure and then the sensitivity and specificity are based only on those patients who underwent the reference-standard test (11). This bias is counterintuitive in that investigators usually believe that including only the patients for whom there was rigorous verification of the presence or absence of disease will make their study design ideal (12). The opposite is true, however: Studies in which the most stringent verification of disease status is required and the cases with less definitive confirmation are discarded often yield the most biased estimates of accuracy (11,13).

One solution to verification bias is to design the study so that the diagnostic test results will not be used to determine which patients will undergo disease status verification. Rather, the study patients can be selected to undergo the reference-standard procedure on the basis of their signs, symptoms, and other test results—not the results of the test(s) evaluated in the study. This is not always possible because the test(s) under evaluation may be the usual clinical test(s) used to make diagnoses and manage the treatment of these patients.

Another solution is to use different reference-standard procedure(s) for different patients. For example, in evaluating the accuracy of CT for lung cancer screening, some patients may undergo biopsy and surgery and others can be followed up clinically and radiologically for a specified period (eg, 2 years) to detect wrongly diagnosed cases (Table). We cannot simply assume that patients with negative test results are disease free; this assumption can lead to a serious overestimation of test specificity (11).

A third solution to verification bias is to apply a statistical correction to the estimates of accuracy. A number of correction methods exist (14). Most of these methods are based on the assumption that the decision to verify a patient's diagnosis—that is, to refer the patient for further diagnostic work-up, including the reference-standard test used in the study—is a conscious one and thus is based on visible factors, such as the test result and the patient's signs and symptoms. To apply any of the correction methods, it is essential that we record the results of all patients who undergo the test being assessed—not just those of patients who undergo the evaluated test and the reference-standard procedure.

BIAS IN PERFORMING AND INTERPRETING TESTS

Tests that are being evaluated in a study must be performed and interpreted without knowledge of the results of competing tests and, when applicable, without knowledge of the results of the reference-standard procedure. If a reference-standard procedure is used in a study, it must be performed and interpreted without knowledge of the results of the diagnostic test(s) being evaluated.

Review bias (4) occurs when a diagnostic test, or the reference-standard test, is performed or interpreted without proper blinding (Table). Consider as an example a study to compare the capability of CT and ultrasonography (US) to depict tumors. When performing US, the technician and radiologist should not be aware of the CT findings because the technician might search with more scrutiny in locations where a tumor was found at CT and the radiologist may have a tendency to "overread" a suspicious area when he or she knows that the CT reader interpreted it to be a tumor. The simplest way to avoid this type of bias is to "blind" both the technician and the reader to the results of the other tests.

In retrospective studies in which the tests have already been performed and interpreted, it is critical that we scrutinize the usual clinical practice in search of review bias. For example, suppose we are reviewing the test findings of all patients who underwent CT and pulmonary angiography for detection of pulmonary emboli. We may find that angiography was almost always performed after CT, and we may suspect that the angiogram was obtained and interpreted with knowledge of the CT findings. For such a study, it may be possible to reinterpret the angiogram while blinded to the CT results. However, one cannot perform the angiographic examination again while blinded to the CT results. In these situations we must be aware that the potential for bias exists and interpret the study findings with the appropriate level of caution.

When two tests—for example, tests A and B—are performed in the same patient and the images are interpreted by the same reader, the images read last—for example, the test B images—will tend to be interpreted more accurately than the images read first—that is, the test A images—if the reader retains any information (15). This situation is called reading-order bias, and it can *(a)* negate a real difference (ie, if test A is really superior to test B), *(b)* inflate the true difference (ie, if test B is really superior to test A), or *(c)* create a difference when no true difference exists.

The simplest way to reduce or eliminate reading-order bias is to vary the order in which the test findings are interpreted (15). For example, suppose 50 patients underwent both test A and test B. The reader could first interpret the results of test A for half of the patients—let us call them group 1. Next, the reader would interpret the results of test B for the second half of the patients—let us call them group 2. After a sufficient time lag, the reader would interpret the test B results for group 1 and then the test A results for group 2. This way, the effect of reading-order bias would be cancelled out, because although the test A results would be read first for half of the patients, the test B results also would be read first for half of the patients.

Note that patients would have to be randomized (Table) to the two groups and the images obtained in the two groups would need to be presented to the readers in random order. The rationale for this protocol is that readers sometimes remember the first (and even second and last) case in a reading session, so by randomizing patients we reduce the effect of any retained information.

An additional way to reduce the effect of retained information is to allow a sufficient time lag between the first and subsequent readings of images in the same case. No standard time is appropriate for all studies. Rather, the duration of the time lag should depend on the complexity of the readings and the volume of the study cases and similar clinical cases that the reader is expected to interpret. For example, if the study cases are those from screening examinations and the reader in his or her typical clinical practice interprets the results of many screening examinations, then a short time lag (ie, a few days) is probably sufficient. In contrast, if the study cases are difficult and complex to interpret and thus a great deal of time is required to determine the diagnosis, and/or if the reader does not typically interpret the types of cases included in the study, then a long time lag (ie, several months) is needed to minimize the retained information.

One last bias that I will discuss in this section occurs when tests are interpreted in an artificial environment. Intuitively, in an experimental setting, we might expect readers to interpret cases with more care because they know that their performance is being measured. Egglin and Feinstein (16) addressed another issue that affects reader performance. They performed a study to assess the effect that disease prevalence has on test interpretation. They assembled a test set of pulmonary arteriograms with a depicted pulmonary embolism prevalence of 33% and embedded this set into two larger groups of arteriograms such that group A had an overall prevalence rate of 60% and group B an overall prevalence rate of 20%. After blinded randomized reviews by six readers, they concluded that readers' accuracies differ depending on the context and often improve when the disease prevalence is higher. Egglin and Feinstein (16) defined context bias as the bias in accuracy measurements that occurs when the disease prevalence in the sample differs greatly from the prevalence in the clinical population. They suggested that investigators use a sample with a disease prevalence similar to that in the clinically relevant population.

BIAS IN ANALYZING TEST RESULTS

Some tests yield uninterpretable results (17). Causes of uninterpretable results include insufficient cell specimens from needle biopsy, abdominal gas interfering with pelvic US imaging, and dense breast tissue at mammography screening. In analyzing the results of a study it is critical not to omit these cases. Rather, we must report the frequency and causes of such cases. When comparing tests, the frequencies of uninterpretable results from the tests should be compared. Poynard et al (18) compared three tests for diagnosing extrahepatic cholestasis and found that the clinical usefulness of the three tests was strongly influenced by the frequencies of uninterpretable results of the different examinations.

Another common problem occurs when some study forms are missing or parts of the forms are incomplete or filled out incorrectly. Response bias occurs when we include just the complete data in our analysis and ignore the missing data. The problem is that there is often a pattern to the missing data—for example, patients who are found to be disease free tend to be followed up with less scrutiny compared with patients who have disease, so data on, for example, patient satisfaction are mostly from patients with disease. However, the results might be different for disease-free patients.

Although there are statistical methods to account for data that are missing not at random (19), it is best to minimize the frequency of missing data by properly training the staff who complete the forms and including mechanisms to collect the incomplete data (eg, multiple telephone and mail messages to nonresponders, cross checks in other databases for information on medical utilization and major outcomes).

CONCLUSION

Sources of bias are everywhere, making it very challenging to design and interpret studies. Researchers should implement ways to avoid bias or to minimize its effect while still in the planning phase of their study. We all should be aware of common biases so that we are able to make informed judgments about the generalizability of study results to our clinical practice.

References

1. Altman DG, Bland JM. Generalisation and extrapolation. BMJ 1998; 317:409–410. Available at: *www.bmj.com/cgi/content/full/317/7155/409*. Accessed September 26, 2003.
2. Altman DG, Bland JM. How to randomise. BMJ 1999; 319:703–704. Available at: *www.bmj.com/cgi/content/full/319/7211/703*. Accessed September 26, 2003.
3. Sox H, Stern S, Owens D, Abrams HL. Assessment of diagnostic technology in health care: rationale, methods, problems, and directions. Washington, DC: National Academy Press, 1989.
4. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 299:926–930.
5. Beam CA, Baker ME, Paine SS, Sostman HD, Sullivan DC. Answering unanswered questions: proposal for a shared resource in clinical diagnostic radiology research. Radiology 1992; 183:619–620.
6. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. Arch Intern Med 1996; 156:209–213.
7. Valenstein PN. Evaluating diagnostic tests with imperfect standards. Am J Clin Pathol 1990; 93:252–258.
8. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–298.
9. Thornbury JR, Fryback DG, Turski PA, et al. Disk-caused nerve compression in patients with acute low-back pain: diagnosis with MR, CT myelography, and plain CT. Radiology 1993; 186:731–738.
10. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Stat Methods Med Res 1998; 7:354–370.
11. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987; 6:411–423.
12. Begg CB, McNeil BJ. Assessment of radiologic tests, control of bias and other design considerations. Radiology 1988; 167:565–569.
13. Black WC. How to evaluate the radiology literature. AJR Am J Roentgenol 1990; 154:17–22.
14. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. Stat Methods Med Res 1998; 7:337–353.
15. Metz CE. Some practical issues of experimental design and data analysis in radiologic ROC studies. Invest Radiol 1989; 24:234–245.
16. Egglin TK, Feinstein AR. Context bias: a problem in diagnostic radiology. JAMA 1996; 276:1752–1755.
17. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. J Chronic Dis 1986; 39:575–584.
18. Poynard T, Chaput JC, Etienne JP. Relations between effectiveness of a diagnostic test, prevalence of the disease, and percentages of uninterpretable results: an example in the diagnosis of jaundice. Med Decis Making 1982; 2:285–297.
19. Little RJA, Rubin DB. Statistical analysis with missing data. New York, NY: Wiley, 1987.

Proportions, Odds, and Risk[1]

Christopher L. Sistrom, MD, MPH
Cynthia W. Garvan, PhD

Index term:
Statistical analysis

Published online
10.1148/radiol.2301031028
Radiology 2004; 230:12–19

Abbreviations:
OR = odds ratio
RR = relative risk

Perhaps the most common and familiar way that the results of medical research and epidemiologic investigations are summarized is in a table of counts. Numbers of subjects with and without the outcome of interest are listed for each treatment or risk factor group. By using the study sample data thus tabulated, investigators quantify the association between treatment or risk factor and outcome. Three simple statistical calculations are used for this purpose: difference in proportions, relative risk, and odds ratio. The appropriate use of these statistics to estimate the association between treatment or risk factor and outcome in the relevant population depends on the design of the research. Herein, the enumeration of proportions, odds ratios, and risks and the relationships between them are demonstrated, along with guidelines for use and interpretation of these statistics appropriate to the type of study that gives rise to the data.

In a previous article in this series (1), the 2 × 2 contingency table was introduced as a way of organizing data from a study of diagnostic test performance. Applegate et al (2) have previously described analysis of nominal and ordinal data as counts and medians. Binary variables are a special case of nominal data where there are only two possible levels (eg, yes/no, true/false). Data in two binary variables arise from a variety of research methods that include cross-sectional, case-control, cohort, and experimental designs. In this article, we will describe three ways to quantify the strength of the relationship between two binary variables: difference of proportions, relative risk (RR), and odds ratio (OR). Appropriate use of these statistics depends on the type of data to be analyzed and the research study design.

Correct interpretation of the difference of proportions, the RR, and the OR is key to the understanding of published research results. Misuse or misinterpretation of them can lead to errors in medical decision making and may even have adverse public policy implications. An example can be found in an article by Schulman et al (3) published in the *New England Journal of Medicine* about the effects of race and sex on physician referrals for cardiac catheterization. Results of this study of Schulman et al received extensive media coverage about the findings that blacks and women were referred less often than white men for cardiac catheterization. In a follow-up article, Schwartz et al (4) showed how the magnitude of the findings of Schulman et al was overstated, chiefly because of confusion among OR, RR, and probability. The resulting controversy underscores the importance of understanding the nuances of these statistical measures.

Our purpose is to show how a 2 × 2 contingency table summarizes results of several common types of biomedical research. We will describe the four basic study designs that give rise to such data and provide an example of each one from literature related to radiology. The appropriate use of difference in proportion, RR, and OR depends on the study design used to generate the data. A key concept to be developed about using odds to estimate risk is that the relationship between OR and RR depends on outcome frequency. Both graphic and computational correction of OR to estimate RR will be shown. With rare diseases, even a corrected RR estimate may overstate the effect of a risk factor or treatment. Use of difference in proportions (expressed as attributable risk) may give a better picture of societal impact. Finally, we introduce the concept of confounding by factors extraneous to the research question that may lead to inaccurate or contradictory results.

2 × 2 CONTINGENCY TABLES

Let X and Y denote two binary variables that each have only two possible levels. Another term for binary is dichotomous. Results are most often presented as counts of observations at each level. The relationship between X and Y can be displayed in a 2 × 2 contingency table. Another name for a contingency table is a cross-classification table. A 2 × 2

[1] From the Departments of Radiology (C.L.S.) and Biostatistics (C.W.G.), University of Florida College of Medicine, PO Box 100374, Gainesville, FL 32610. Received July 2, 2003; revision requested July 30; revision received August 4; accepted August 13. **Address correspondence to** C.L.S. (e-mail: *sistrc@radiology.ufl.edu*).

contingency table consists of four cells: the cell in the first row and first column (cell 1–1), the cell in the first row and second column (cell 1–2), the cell in the second row and first column (cell 2–1), and the cell in the second row and second column (cell 2–2). Commonly used symbols for the cell contents include *n* with subscripts, *p* with subscripts, and the letters *a–d*. The n_{11}, n_{12}, n_{21}, and n_{22} notation refers to the number of subjects observed in the corresponding cells. In general, "n_{ij}" refers to the number of observations in the ith row (i = 1, 2) and jth column (j = 1, 2). The total number of observations will be denoted by *n* (ie, $n = n_{11} + n_{12} + n_{21} + n_{22}$). The p_{11}, p_{12}, p_{21}, and p_{22} notation refers to the proportion of subjects observed in each cell. In general, "p_{ij}" refers to the proportion of observations in the ith row (i = 1, 2) and jth column (j = 1, 2). Note that $p_{ij} = n_{ij}/n$. For simplicity, many authors use the letters *a–d* to label the four cells as follows: *a* = cell 1–1, *b* = cell 1–2, *c* = cell 2–1, and *d* = cell 2–2. We will use the *a–d* notation in equations that follow. Table 1 shows the general layout of a 2 × 2 contingency table with symbolic labels for each cell and common row and column assignments for data from medical studies.

TABLE 1
Notation for 2 × 2 Contingency Table

	*Y**	
X†	Yes‡	No§
Present‖	n_{11} p_{11} *a*	n_{12} p_{12} *b*
Absent#	n_{21} p_{21} *c*	n_{22} p_{22} *d*

Note.—*n* = number of subjects in the cell, *p* = proportion of entire sample in the cell, *a–d* = commonly used cell labels.

* Response, outcome, or disease status variable.

† Explanatory, risk factor, or exposure variable.

‡ Adverse outcome or disease-positive response.

§ No adverse outcome or disease-negative response.

‖ Exposed or risk-positive group.

# Unexposed or risk-negative group.

In many contingency tables, one variable is a response (outcome or dependent variable) and the other is an explanatory (independent) variable. In medical studies, the explanatory variable (*X* in the rows) is often a risk or a protective factor and the response (*Y* in the columns) is a disease state. The distribution of observed data in a 2 × 2 table indicates the strength of relationship between the explanatory and the response variables. Figure 1 illustrates possible patterns of observed data. The solid circle represents a cell containing numerous observations. Intuitively, we would expect that if the *X* and *Y* variables are associated, then pattern A or B would be observed. Patterns C, D, and E suggest that *X* and *Y* are independent of each other (ie, there is no relationship between them).

Pattern A	Y	
X	Yes	No
Present	●	
Absent		●

Pattern B	Y	
X	Yes	No
Present		●
Absent	●	

Pattern C	Y	
X	Yes	No
Present	●	
Absent	●	

Pattern D	Y	
X	Yes	No
Present		●
Absent		●

Pattern E	Y	
X	Yes	No
Present	●	●
Absent	●	●

Figure 1. Diagram shows possible patterns of observed data in a 2 × 2 table. Cells with black circles contain relatively large numbers of counts. With pattern A, n_{11} and n_{22} are large, suggesting that when *X* is present, *Y* is "yes." With pattern B, n_{12} and n_{21} are large, suggesting that when *X* is absent, *Y* is "yes." With pattern C, n_{11} and n_{21} are large, suggesting that *Y* is "yes" regardless of *X*. With pattern D, n_{12} and n_{22} are large, suggesting that *Y* is "no" regardless of *X*. With pattern E, n_{11}, n_{12}, n_{21}, and n_{22} are all about the same, suggesting that *Y* is "yes" and "no" in equal proportion regardless of *X*.

STUDY DESIGNS THAT YIELD 2 × 2 TABLES

The statistical methods used to analyze research data depend on how the study was conducted. There are four types of designs in which two-by-two tables may be used to organize study data: case-control, cohort, cross sectional, and experimental. The first three designs are often called observational to distinguish them from experimental studies; of the experimental studies, the controlled clinical trial is the most familiar. The 2 × 2 tables that result from the four designs may look similar to each other. The outcome is typically recorded in columns, and the explanatory variable is listed in the rows. The χ^2 statistic may be calculated and used to test the null hypotheses of independence between row and column variables for all four types of studies. These methods are described in a previous article in this series (2). Table 2 summarizes the features of the four types of designs in which 2 × 2 tables are used to organize study data. Each is briefly described next, with a radiology-related example provided for illustration. The ordering of the study designs in the following paragraphs reflects, in general, the strength (ie, weaker to stronger) of evidence for causation obtained from each one. The advantages and disadvantages of the different designs and situations where each is most appropriate are beyond the scope of this article, and the reader is encouraged to seek additional information in biostatistics, research design, or clinical epidemiology reference books (5–7).

The statistics used to quantify the relationship between variables (ie, the effect size) are detailed and the examples of study designs are summarized in Table 3. These will be briefly defined now. The difference in proportion is the difference in the fraction of subjects who have the outcome between the two levels of the explanatory variable. The RR is the ratio of proportion (ie, risk) of subjects who have the outcome between different levels of the explanatory variable. The OR is the ratio of the odds that a subject will have the outcome between the two levels of the explanatory variable. Each of these statistics has an associated standard error that can be calculated and used to form CIs around the estimate at any chosen level of precision. The calculation of CIs and their use for inference testing are described in a previous article in this series (8).

Cross-sectional Studies

A cross-sectional study does not involve the passage of time. A single sample is selected without regard to disease state or exposure status. Information on disease state and exposure status is determined with data collected at a single time point. Data about exposure status and

TABLE 2
Comparison of Four Study Designs

Attribute	Cross-sectional Study	Case-Control Study	Cohort Study	Experimental Study
Sample selection	One sample selected without regard to disease or exposure status	Two samples selected: one from disease-positive population, one from disease-negative population	Two samples selected: one from exposed population, one from unexposed population	One sample selected that is disease negative; sample is randomly assigned to treatment or control group
Proportions that can be estimated	Prevalence of disease in the exposed and unexposed groups	Proportion of cases and controls that have been exposed to a risk factor	Incidence of disease in exposed and unexposed groups	Incidence of disease in treated and untreated (control) groups
Time reference	Present look at time	Backward look in time	Forward look in time	Forward look in time
Effect measure	OR, difference in proportions*	OR	RR, difference in proportions*	RR, difference in proportions*

* Difference in proportions may be used as an alternate measure of effect.

TABLE 3
Quantification of Effect Size for Two Binary Variables

Definition	Difference of Proportions*	RR†	OR‡
Calculation for estimate based on sample data	$n_{11}/n_{11} + n_{12} - n_{21}/n_{21} + n_{22}$	$\frac{n_{11}/n_{11} + n_{12}}{n_{21}/n_{21} + n_{22}}$	$n_{11}n_{22}/n_{12}n_{21}$
Calculation in terms of *a*–*d*-cell labels	$[a/(a + b)] - [c/(c + d)]$	$\frac{[a/(a + b)]}{[c/(c + d)]}$	ad/bc
Cross-sectional example (Table 4)	0.22	1.69	2.52
Case-control example (Table 5)	NA§	NA§	2.22
Cohort example (Table 6)	0.046	1.07	1.25‖
Experimental example (Table 7)	0.029	1.56	1.61‖

* Proportion with outcome in exposed group minus proportion with outcome in unexposed group.
† Risk of outcome in exposed group divided by risk of outcome in unexposed group.
‡ Odds of outcome in exposed group divided by odds of outcome in unexposed group.
§ NA = not typically calculated or reported, since the statistic is not meaningful for this study design.
‖ ORs may be calculated for cohort and experimental studies, but RR is preferred.

TABLE 4
Example of Cross-sectional Study Data

Major Risk Factor	DVT Positive*	DVT Negative*	Total
Present	81	67	148
Absent	90	188	278

Source.—Reference 9.
* DVT = deep venous thrombosis.

disease state can be organized into a 2 × 2 contingency table, and the prevalence (ie, the proportion of a group that currently has a disease) can be compared for the exposed and unexposed groups. Effect size from cross-sectional studies may be assessed with difference in proportions, RR, or OR. For example, Cogo et al (9) studied the association between having a major risk factor (eg, immobilization, trauma, and/or recent surgery) and deep vein thrombosis. This study was performed in an outpatient setting by using a cross-sectional design. A total of 426 subjects who were referred by general practitioners underwent contrast material–enhanced venography to determine deep vein thrombosis status (positive or negative). Concurrently, information on major risk factors was recorded as being present or absent. They found that deep vein thrombosis was more likely to occur when a major risk factor was present (81 [55%] of 148) than when none was present (90 [32%] of 278). The data are shown in Table 4, and effect measures are included in Table 3.

Case-Control Studies

In a case-control study, the investigator compares instances of a certain disease or condition (ie, the cases) with individuals who do not have the disease or condition (ie, the "controls" or control subjects). Control subjects are usually selected to match the patients with cases of disease in characteristics that might be related to the disease or condition of interest. Matching by age and sex is commonly used. Investigators look backward in time (ie, retrospectively) to collect information about risk or protective factors for both cases and controls. This is achieved by examining past records, interviewing the subject, or in some other way. The only correct measure of effect size for a case-control study is the OR. However, the calculated OR may be used to estimate RR after appropriate correction for disease frequency in the population of interest, which will be explained later. For example, Vachon et al (10) studied the association between type of hormone replacement therapy and increased mammographic breast density by using a case-control study design. They identified 172 women who were undergoing hormone replacement therapy who had increased breast density (cases) and 172 women who were undergoing hormone replacement therapy who did not have increased breast density (controls). The type of hormone replacement therapy used by all subjects was then determined. They found that combined hormone replacement therapy was associated with increased breast density more often than was therapy with estrogen alone (OR = 2.22). The data are presented in Table 5.

Cohort Studies

A cohort is simply a group of individuals. The term is derived from Roman military tradition; according to this tradition, legions of the army were divided into 10 cohorts. This term now means any specified subdivision or group of people marching together through time. In other words, cohort studies are about the life histories of sections of populations and the individuals who are in-

TABLE 5
Example of Case-Control Study Data

Condition	Cases*	Controls†
Exposed‡	111	79
Nonexposed§	50	79

Source.—Reference 10.
* Increased breast density.
† No increased breast density.
‡ Combined therapy.
§ Estrogen alone.

TABLE 6
Example of Cohort Study Data

	No. Undergoing Mammography within 2 Years*		
Result at Last Mammography	Returned	Did Not Return	Total
False-positive	602	211	813
True-negative	3,098	1,359	4,457

Source.—Reference 11.
* Within 2 years after last mammography.

TABLE 7
Example of Experimental Study Data

	Underwent Mammography within 2 Years		
Treatment	No. Who Did	No. Who Did Not	Total
Intervention*	100	1,129	1,229
Control†	64	1,165	1,229

Source.—Reference 12.
* Intervention indicates that subjects received a mailed reminder.
† Control indicates that subjects did not receive a mailed reminder.

cluded in them. In a prospective study, investigators follow up subjects after study inception to collect information about development of disease. In a retrospective study, disease status is determined from medical records produced prior to the beginning of the study but after articulation of the cohorts. In both types of studies, initially disease-free subjects are classified into groups (ie, cohorts) on the basis of exposure status with respect to risk factors. Cumulative incidence (ie, the proportion of subjects who develop disease in a specified length of time) can be computed and compared for the exposed and unexposed cohorts. The main difference between prospective and retrospective cohort studies is whether the time period in question is before (retrospective) or after (prospective) the study begins. Effect size from a cohort study is typically quantified with RR and/or difference in proportions. For example, Burman et al (11) studied the association between false-positive mammograms and interval breast cancer screening by using a prospective cohort study design. Women in whom a false-positive mammogram was obtained at the most recent screening formed one cohort, and women in whom a previously negative mammogram was obtained formed the other cohort. All of the women were followed up to determine if they obtained a subsequent screening mammogram within the recommended interval (up to 2 years, depending on age). Burman et al found no significant difference in the likelihood that a woman would obtain a mammogram between the two cohorts (RR = 1.07). The data are included in Table 6.

Experimental Studies

The characteristic that distinguishes any experiment is that the investigator directly manipulates one or more variables (not the outcome!). A clinical trial is the most common type of experimental study used in medical research. Here, the investigator selects a sample of subjects and assigns each to a treatment. In many cases, one treatment may be standard therapy or an inactive (ie, placebo) treatment. These subjects are the controls, and they are compared with the subjects who are receiving a new or alternative treatment. Treatment assignment is almost always achieved randomly so that subjects have an equal chance of receiving one of the treatments. Subjects are followed up in time, and the cumulative incidence of the outcome or disease is compared between the treatment groups. RR and/or difference in proportions is typically used to quantify treatment effect on the outcome. An example of such a study can be found in an article by Harrison et al (12) in which they describe their trial of direct mailings to encourage attendance for mammographic screening. At the start of the study, half of the women were randomly assigned to receive a personally addressed informational letter encouraging them to attend screening. The other half (ie, controls) received no intervention. The number of women in each group who underwent mammography during the subsequent 2 years was enumerated. The women to whom a letter was sent were more likely to obtain a mammogram (RR = 1.56). The data are listed in Table 7.

CALCULATION OF PROPORTIONS FROM A 2 × 2 TABLE

Various proportions can be calculated from the data represented in a 2 × 2 contingency table. The cell, row, and column proportions each give different information about the data. Proportions may be represented by percentages or fractions, with the former having the advantage of being familiar to most people. Cell proportions are simply the observed number in each of the four cells divided by the total sample size. Each cell also has a row proportion. This is the number in the cell divided by the total in the row containing it. Likewise, there are four column proportions that are calculated by dividing the number in each cell by the total in the column that contains it.

In a 2 × 2 table organized with outcome in the columns and exposure in the rows, the various proportions have commonly understood meanings. Cell proportions are the fractions of the whole sample found in each of the four combinations of exposure and outcome status. Row proportions are the fractions with and without the outcome. It may seem counterintuitive that row proportions give information about the outcome. However, remembering that cells in a given row have the same exposure status helps one to clarify the issue. Similarly, column proportions are simply the fraction of exposed and unexposed subjects.

DIFFERENCE IN PROPORTIONS

The difference in proportions is used to compare the response *Y* (eg, disease: yes or no) according to the value of the explanatory variable *X* (eg, risk factor: exposed or unexposed). The difference is defined as the proportion with the outcome in the exposed group minus the proportion with the outcome in the unexposed group. By using the *a*–*d* letters for cell labels (Table 1), the calculation is as follows:

$$[a/(a + b)] - [c/(c + d)].$$

For the cross-sectional study data (Table 4) we would calculate the following equation:

$$[81/(81 + 67)] - [90/(90 + 188)] = 0.547 - 0.324 = 0.223.$$

The difference in proportions always is between −1.0 and 1.0. It equals zero when the response *Y* is statistically independent of the explanatory variable *X*. When *X* and *Y* are independent, then there is no association between them. It is appropriate to calculate the difference in proportions for the cohort, cross-sectional, and experimental study designs. In a case-control study, there is no information about the proportions that are outcome (or disease) positive in the population. This is because the investigator actually selects subjects to get a fixed number at each level of the outcome (ie, cases and controls). Therefore, the difference in proportions statistic is inappropriate for estimating the association between exposure and outcome in a case-control study. Table 3 lists the difference in proportions estimate for cross-sectional, cohort, and experimental study examples (9,11,12).

RISK AND RR

Risk is a term often used in medicine for the probability that an adverse outcome, such as a side effect, development of a disease, or death, will occur during a specific period of time. Risk is a parameter that is completely known only in the rare situation when data are available for an entire population. Most often, an investigator estimates risk in a particular population by taking a representative random sample, counting those that experience the adverse outcome during a specified time interval, and forming a proportion by dividing the number of adverse outcomes by the sample size. For example, the estimate of risk is equal to the number of subjects who experience an event or outcome divided by the sample size.

The epidemiologic term for the resulting rate is the cumulative incidence of the outcome. The incidence of an event or outcome must be distinguished from the prevalence of a disease or condition. Incidence refers to events (eg, the acquisition of a disease), while prevalence refers to states (eg, the state of having a disease). RR is a measure of association between exposure to a particular factor and risk of a certain outcome. The RR is defined to be the ratio of risk in the exposed and unexposed groups. An equivalent term for RR that is sometimes used in epidemiology is the cumulative incidence ratio, which may be calculated as follows: RR is equal to the risk among exposed subjects divided by the risk among unexposed subjects.

In terms of the letter labels for cells (Table 1), RR is calculated as follows:

$$\mathrm{RR} = [a/(a + b)]/[c/(c + d)].$$

The RR for our cohort study example (Table 6) would be calculated by dividing the fraction with a mammogram in the false-positive cohort (602/813 = 0.74) by the fraction with a mammogram in the true-negative cohort (3,098/4,457 = 0.695). This yields 1.065. The value of RR can be any nonnegative number. An RR of 1.0 corresponds to independence of (or no association between) exposure status and adverse outcome. When RR is greater than 1.0, the risk of disease is increased when the risk factor is present. When RR is less than 1.0, the risk of disease is decreased when the risk factor is present. In the latter case, the factor is more properly described as a protective factor. The interpretation of RR is quite natural. For example, an RR equal to 2.0 means that an exposed person is twice as likely to have an adverse outcome as one who is not exposed, and an RR of 0.5 means that an exposed person is half as likely to have the outcome. Table 3 illustrates the calculation of the RR estimates from the various examples.

Any estimate of relative risk must be considered in the context of the absolute risk. Motulsky (5) gives an example to show how looking only at RR can be misleading. Consider a vaccine that halves the risk of a particular infection. In other words, the vaccinated subjects have an RR of 0.5 of getting infected compared with their unvaccinated peers. The public health impact depends not only on the RR but also on the absolute risk of infection. If the risk of infection is two per million unvaccinated people in a year, then halving the risk to one per million is not so important. However, if the risk of infection is two in 10 unvaccinated people in a year, then halving the risk is of immense consequence by preventing 100,000 cases per million. Therefore, it is more informative to compare vaccinated to unvaccinated cohorts by using the difference in proportions. With the rare disease, the difference is 0.0000001, while for the common disease it is 0.1. This logic underlies the concept of number needed to treat and number needed to harm. These popular measures developed for evidence-based medicine allow direct comparison of effects of interventions (ie, number needed to treat) or risk factors (ie, number needed to harm). In our vaccination scenario, the number needed to harm (ie, to prevent one infection) for the rare disease is 1 million and for the common disease is 10.

ODDS AND THE OR

The OR provides a third way of comparing proportions in a 2 × 2 contingency table. An OR is computed from odds. Odds and probabilities are different ways of expressing the chance that an outcome may occur. They are defined as follows: The probability of outcome is equal to the number of times the outcome is observed divided by the total observations. The odds of outcome is equal to the probability that the outcome does occur divided by the probability that the outcome does not occur.

We are familiar with the concept of odds through gambling. Suppose that the odds a horse named Lulu will win a race are 3:2 (ie, read as "three to two"). The 3:2 is equivalent to 3/2 or 1.5. The probability that Lulu will be the first to cross the finish line can be calculated from the odds, since there is a deterministic relationship between odds and probability (ie, if you know the value of one, then you can find the value of the other). We know that:

$$Pr = \mathrm{Odds}/(1 + \mathrm{Odds})$$

and

$$\mathrm{Odds} = Pr/(1 - Pr),$$

where *Pr* is probability.

The probability that Lulu will win the race is (1.5)/1 + (1.5) = 0.60, or 60%.

Probabilities always range from 0 to 1.0, while odds can be any nonnegative number. The odds of a medical outcome in exposed and unexposed groups are defined as follows: Odds of disease in the exposed group is equal to the probability that the disease occurs in the exposed group divided by the probability that the disease does not occur in the exposed group. Odds of disease in the unexposed group is equal to the probability that the disease occurs in the unexposed group divided by the probability that the disease does not occur in the unexposed group.

It is helpful to define odds in terms of the *a–d* notation shown in Table 1. Useful mathematical simplifications (where $\mathrm{Odds}_{\mathrm{exp}}$ is the odds of outcome in the exposed group and $\mathrm{Odds}_{\mathrm{unex}}$ is the odds of outcome in the unexposed group) that arise from this definition are as follows:

$$\mathrm{Odds}_{\mathrm{exp}} = [a/(a + b)]/[b/(a + b)] = a/b$$

and

$$\mathrm{Odds}_{\mathrm{unex}} = [c/(c + d)]/[d/(c + d)] = c/d.$$

Note that these simplifications mean that the outcome probabilities do not have to be known in order to calculate odds. This is especially relevant in the analysis of case-control studies, as will be illustrated. The OR is defined as the ratio of the odds and may be calculated as follows: OR is equal to the odds of disease in the exposed group divided by the odds of disease in the unexposed group. By using the *a–d* labeling,

$$\mathrm{OR} = (a/b)/(c/d) = ad/bc.$$

The OR for our case-control example (Table 5) would be calculated as follows:

$$\mathrm{OR} = [(111)(79)]/[(79)(50)] = 2.22.$$

The OR has another property that is particularly useful for analyzing case-control studies. The OR we calculate from a case-control study is actually the ratio of odds of exposure, not outcome. This is because the numbers of subjects with and without the outcome are always fixed in a case-control study. However, the calculation for exposure OR and that for outcome OR are mathematically equivalent, as shown here:

$$(a/c)/(b/d) = ad/bc = (a/b)/(c/d).$$

Therefore, in our example, we can correctly state that the odds of increased breast density is 2.2 times greater in those receiving combined hormone replacement therapy than it is in those receiving estrogen alone. Authors, and readers, must be very circumspect about continuing with the chain of inference to state that the risk of increased breast density in those receiving combined hormone replacement therapy is 2.2 times higher than is the risk of increased breast density in those receiving estrogen alone. In doing this, one commits the logical error of assuming causation from association. Furthermore, the OR is not a good estimator of RR when the outcome is common in the population being studied.

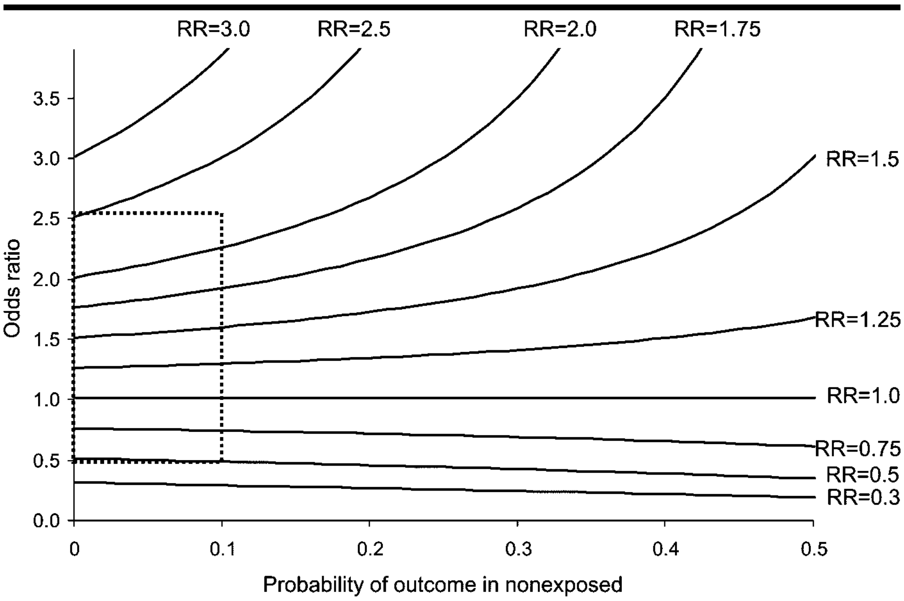


Figure 2. Graph shows relationship between OR and probability of outcome in unexposed group. Curves represent values of underlying RR as labeled. Rectangle formed by dotted lines represents suggested bounds on OR and probability of outcome in the unexposed group within which no correction from OR to RR is needed. When OR is more than 2.5 or less than 0.5 or probability of outcome in the unexposed group is more than 0.1, a correction (calculation in text) should be applied.

RELATIONSHIP BETWEEN RR AND OR

There is a strict and invariant mathematic relationship between RR and OR when they both are from the same population, as may be observed with the following equation:

$$\mathrm{RR} = \mathrm{OR}/[(1 - Pr_{\mathrm{o}}) + (Pr_{\mathrm{o}})(\mathrm{OR})],$$

where Pr_{o} is the probability of the outcome in the unexposed group.

The relationship implies that the magnitude of OR and that of RR are similar only when Pr_{o} is low (13). In epidemiology, Pr_{o} is referred to as the incidence of outcome in the unexposed group. Thus, the OR obtained in a case-control study accurately estimates RR only when the outcome in the population being studied is rare. Figure 2 shows the relationship between RR and OR for various values of Pr_{o}. Note that the values of RR and OR are similar only when the Pr_{o} is small (eg, 10/100 = 10% or less). At increasing Pr_{o}, ORs that are less than 1.0 underestimate the RR, and ORs that are greater than 1.0 overestimate the RR. A rule of thumb is that the OR should be corrected when incidence of the outcome being studied is greater than 10% if the OR is greater than 2.5 or the OR is less than 0.5 (4). Note that with a case-control study, the probability of outcome in the unexposed must be obtained separately because it cannot be estimated from the sample.

This distinction was at the heart of the critique of Schwartz et al (4) regarding the article by Schulman et al (3). Schulman et al had reported an OR of 0.6 for referral to cardiac catheterization (outcome) between blacks and whites (explanatory). However, referral for catheterization occurred up to 90% of the time, so the corresponding RR should have been 0.93. This information would have created a rather unspectacular news story (ie, blacks referred 7% less often than whites) compared with the initial, and incorrect, headlines stating that blacks were referred 40% less often than whites.

In our examples, this relationship is also apparent. In the studies where both RR and OR were calculated, the OR is

TABLE 8
Death Penalty Verdicts Following Murder Convictions in Florida, 1976–1987

		Death Penalty*		
Victim's Race	Defendant's Race	Yes	No	RR, White/Black
White	White	53 (11)	414 (89)	0.495
White	Black	11 (23)	37 (77)	. . .
Black	White	0 (0)	16 (100)	0
Black	Black	4 (3)	139 (97)	. . .
Combined	White	53 (11)	430 (89)	UNC, 1.40† MH, 0.48‡
Combined	Black	15 (8)	176 (92)	. . .

Source.—Reference 15.

* Data are numbers of verdicts. Data in parentheses are percentages of the totals.

† UNC = uncorrected.

‡ MH = Mantel-Haenszel estimate.

larger than the RR, as we now expect. In the experimental study example, the OR of 1.61 is only slightly larger than the RR of 1.56. This is because the probability of mammography (the outcome) is rare at 6.7 per 100. In contrast, the cohort study yields an OR of 1.25, which is considerably larger than the RR of 1.07, with the high overall probability of mammography of 73 per 100 explaining the larger difference.

CONFOUNDING VARIABLES AND THE SIMPSON PARADOX

An important consideration in any study design is the possibility of confounding by additional variables that mask or alter the nature of the relationship between an explanatory variable and the outcome. Consider data from three binary variables, the now familiar X and Y, as well as a new variable Z. These can be represented in two separate 2 × 2 contingency tables: one for X and Y at level 1 of variable Z and one for X and Y at level 2 of variable Z. Alternatively, the X and Y data can be represented in a single table that ignores the values of Z (ie, pools the data across Z).

The problem occurs when the magnitude of association between X (explanatory) and Y (outcome) is different at each level of Z (confounder). Estimation of the association between X and Y from the pooled 2 × 2 table (ignoring Z) is inappropriate and often incorrect. In such cases, it is misleading to even list the data in aggregate form. There are statistical methods to correct for confounding variables, with the assumption that they are known and measurable. A family of techniques attributed to Cochran (14) and Mantel and Haenszel (15) are commonly used to produce summary estimates of association between X and Y corrected for the third variable Z.

The entire rationale for the use of randomized clinical trials is to eliminate the problem of unknown and/or immeasurable confounding variables. In a clinical trial, subjects are randomly assigned to levels of X (the treatment or independent variable). The outcome (Y) is then measured during the course of the study. The beauty of this scheme is that all potentially confounding variables (Z) are equally distributed among the treatment groups. Therefore, they cannot affect the estimate of association between treatment and outcome. For randomization to be effective in elimination of the potential for confounding, it must be successful. This requires scrupulous attention to detail by investigators in conducting, verifying, and documenting the randomization used in any study.

It is possible for the relationship between X and Y to actually change direction if the Z data are ignored. For instance, OR calculated from pooled data may be less than 1.0, while OR calculated with the separate (so-called stratified) tables is greater than 1.0. This phenomenon is referred to as the Simpson paradox, as Simpson is credited with an article in which he explains mathematically how this contradiction can occur (16). Such paradoxic results are not only numerically possible but they actually arise in real-world situations and can have profound social implications.

Agresti (13) illustrated the Simpson paradox by using death penalty data for black and white defendants charged with murder in Florida between 1976 and 1987 (17). He showed that when the victim's race is ignored, the percentage of death penalty sentences was higher for white defendants. However, after controlling for the victim's race, the percentage of death penalty sentences was higher for black defendants. The paradox arose because juries applied the death penalty more frequently when the victim was white, and defendants in such cases were mostly white. The victim's race (Z) dominated the relationship between the defendant's race (X) and the death penalty verdict (Y). Table 8 lists the data stratified by the victim's race and combined across the victim's race. As indicated in the table, unadjusted RR of receiving the death penalty (white/black) with white victims is 0.495; with black victims, 0.0; and with combined victims, 1.40. The Mantel-Haenszel estimate of the common RR is 0.48, thus solving the paradox. The method for calculating the Mantel-Haenszel summary estimates is beyond the scope of this article. However, confounding variables, the Simpson paradox, and how to handle them are discussed in any comprehensive text about biostatistics or medical research design (5,6).

CONCLUSION

This article has focused on statistical analysis of count data that arise from the four basic designs used in medical research (ie, cross-sectional, case-control, cohort, and experimental study designs). Each of these designs often yields data that are best summarized in a 2 × 2 contingency table. Statistics calculated from such tables include cell, row, and column proportions; differences in proportion; RRs; and ORs. These statistics are used to estimate associated population parameters and are selected to suit the specific aims and design of the study. For inference concerning association between study variables, one must use the correct statistic, allow for variability, and account for any confounding variables.

References

1. Langlotz CP. Fundamental measures of diagnostic examination performance: usefulness for clinical decision making and research. Radiology 2003; 228: 3–9.
2. Applegate KE, Tello R, Ying J. Hypothesis testing III: counts and medians. Radiology 2003; 228:603–608.
3. Schulman KA, Berlin JA, Harless W, et al. The effect of race and sex on physicians' recommendations for cardiac cath-

eterization. N Engl J Med 1999; 340:618–626.

4. Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. N Engl J Med 1999; 341:279–283.
5. Motulsky H. Intuitive biostatistics. Oxford, England: Oxford University Press, 1995.
6. Riegelman RK. Studying a study and testing a test. 4th ed. Philadelphia, Pa: Lippincott Williams & Wilkins, 2000.
7. Sackett DL. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. Boston, Mass: Little, Brown, 1991.
8. Medina LS, Zurakowski D. Measurement variability and confidence intervals in medicine: why should radiologists care? Radiology 2003; 226:297–301.
9. Cogo A, Bernardi E, Prandoni P, et al. Acquired risk factors for deep-vein thrombosis in symptomatic outpatients. Arch Intern Med 1994; 154:164–168.
10. Vachon CM, Sellers TA, Vierkant RA, Wu FF, Brandt KR. Case-control study of increased mammographic breast density response to hormone replacement therapy. Cancer Epidemiol Biomarkers Prev 2002; 11:1382–1388.
11. Burman ML, Taplin SH, Herta DF, Elmore JG. Effect of false-positive mammograms on interval breast cancer screening in a health maintenance organization. Ann Intern Med 1999; 131:1–6.
12. Harrison RV, Janz NK, Wolfe RA, et al. Personalized targeted mailing increases mammography among long-term noncompliant medicare beneficiaries: a randomized trial. Med Care 2003; 41:375–385.
13. Agresti A. Categorical data analysis. Hoboken, NJ: Wiley, 2003.
14. Cochran WG. Some methods for strengthening the common chi-squared tests. Biometrics 1954; 10:417–451.
15. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959; 22:719–748.
16. Simpson EH. The interpretation of interaction in contingency tables. J R Stat Soc B 1951; 13:238–241.
17. Radelet ML, Pierce GL. Choosing those who will die: race and the death penalty in Florida. Florida Law Rev 1991; 43:1–34.

Jonathan H. Sunshine, PhD
Kimberly E. Applegate, MD, MS

Index terms:
Cancer screening
Efficacy study
Radiology and radiologists, outcomes studies
Technology assessment

Published online
10.1148/radiol.2302031277
Radiology 2004; 230:309–314

[1] From the Department of Research, American College of Radiology, 1891 Preston White Dr, Reston, VA 20191 (J.H.S.); Riley Hospital for Children, Indiana University Medical Center, Indianapolis (K.E.A.); and Department of Diagnostic Radiology, Yale University, New Haven, Conn (J.H.S.). Received August 10, 2003; revision requested August 19; revision received and accepted August 21. **Address correspondence to** J.H.S. (e-mail: *jonathans@acr.org*).

Technology Assessment for Radiologists[1]

Health technology assessment is the systematic and quantitative evaluation of the safety, efficacy, and cost of health care interventions. This article outlines aspects of technology assessment of diagnostic imaging. First, it presents a conceptual framework of a hierarchy of levels of efficacy that should guide thinking about imaging test evaluation. In particular, the framework shows how the question answered by most evaluations of imaging tests, "How well does this test distinguish disease from the nondiseased state?" relates to the fundamental questions for all health technology assessment, "How much does this intervention improve the health of people?" and "What is the cost of that improvement?" Second, it describes decision analysis and cost-effectiveness analysis, which are quantitative modeling techniques usually used to answer the two core questions for imaging. Third, it outlines design and operational considerations that are vital if researchers who are conducting an experimental study are to make a quality contribution to technology assessment, either directly through their findings or as an input into decision analyses. Finally, it includes a separate discussion of screening—that is, the application of diagnostic tests to nonsymptomatic populations—because the requirements for good screening tests are different from those for diagnostic tests of symptomatic patients and because the appropriate evaluation methods also differ.

Technologic innovation and diffusion of technology into daily practice in radiology have been nothing short of remarkable in the past several decades. Health technology assessment is the careful evaluation of a medical technology for evidence of its safety, efficacy, cost, cost-effectiveness, and ethical and legal implications (1). Interest and research in health technology assessment are growing in response to the wider application of new technology and the increasing costs of health care today (2).

The goal of this article is to describe some of the rationale and the methods of technology assessment as applied to radiology. For any health care intervention, including diagnostic imaging tests, the ultimate questions are, "How much does this do to improve the health of people?" and "How much does it cost for that gain in health?" We need such an understanding of the radiology services we provide to advocate for our patients and to use our resources efficiently and effectively.

OUTCOMES

Measures of diagnostic accuracy, which are the metrics most commonly used for evaluation of diagnostic tests, answer the question, "How well does this test distinguish disease from the nondiseased state?" The answer to that question often does not provide an answer to the questions about improvement of health and the cost of that improvement, which are the core outcome questions about health care interventions (3,4).

The most productive way to think about this gap between diagnostic accuracy on the one hand and outcomes on the other hand and to think about the inclusion of relevant outcomes in the evaluation of diagnostic tests is to use the conceptual scheme of a six-level "hierarchy of efficacy" developed by Fryback and Thornbury (5,6) (Table). They point out that efficacy at any level in their hierarchy is necessary for efficacy at the level with the next highest number but is not sufficient. In their scheme, diagnostic accuracy is at level 2, and patient and societal outcomes are at levels 5 and 6, respectively. Thus, there may be "many a slip between cup and lip"—that is, between diagnostic accuracy of an imaging test on the one hand and improved health and adequate cost-effectiveness on the other.

Let us trace partway through the schema, starting at the lowest level, to understand the principle that efficacy at one level is necessary but not sufficient for efficacy at the next level. Technical efficacy (level 1), such as a certain minimum spatial resolution, is necessary for diagnostic accuracy (level 2), but it does not guarantee it. Similarly, diagnostic accuracy is necessary if a test is to affect the clinician's diagnosis (level 3), but it is not sufficient. Rather, other sources of information, such as patient history, may dominate, so that even a highly accurate test may have little or no effect on the diagnosis. In such an instance, fairly obviously, the test does not contribute to the level 5 goal of improving patient health.

Hierarchy of Efficacy for Diagnostic Tests

Level	Typical Measures
1, Technical efficacy	Resolution of line pairs
	Pixels per millimeter
	Section thickness
	Noise level
2, Diagnostic accuracy	Sensitivity
	Specificity
	Area under the receiver operating characteristic curve
3, Diagnosis	Percentage of cases in which image is judged helpful in making the diagnosis
	Percentage of cases in which diagnosis made without the test is altered—or altered substantially—when information from the test is received
4, Treatment	Percentage of cases in which image is judged helpful in planning patient treatment
	Percentage of cases in which treatment planned without the test is changed after information from the test is received
5, Patient health outcomes	Percentage of patients improved with test conducted compared with that improved without test conducted
	Percentage difference in specific morbidities with test compared with those without
	Mean increase in quality-adjusted life years with test compared with that without
6, Societal value	Cost-effectiveness from a societal perspective
	Cost per life saved, calculated from a societal perspective

Source.—Adapted and reprinted, with permission, from reference 6.

As the Table shows, there are multiple measures that can be used to quantify the efficacy of a diagnostic imaging test at any of the six levels. Hence, evaluations of imaging tests can involve a variety of measures. Thinking in terms of the hierarchy is also helpful for identification of the level(s) at which information should be obtained in an evaluation of a diagnostic imaging test. Experience, as well as reflection, has taught some lessons. The most important of these include:

1. Because higher-level efficacy is possible only if lower-level efficacy exists, it is often useful to measure efficacy at relatively low-numbered levels.

2. In particular, in the development of a test, it is helpful to measure aspects of technical efficacy (level 1), such as sharpness, noise level, and ability to visualize the anatomic structures of interest. An important aspect of test development consists of finding the technical parameters (voltage, section thickness, etc) that give the best diagnostic accuracy; these measures of technical efficacy are often key results in that process.

3. Diagnostic accuracy (level 2) is the highest level of efficacy that is characteristic of the test alone. For example, the sensitivity and specificity of a test are not dependent on what other diagnostic information is available, unlike level 3 (diagnosis). Also, the methodology and statistics used in measurement of diagnostic accuracy are relatively fully developed. Therefore, measurement of diagnostic accuracy is usually worthwhile.

4. Above diagnostic accuracy, effect on treatment (level 4), an "intermediate outcome," is relatively attractive to measure. It can be measured fairly easily and reliably in a prospective study, and it is closer in the hierarchy to the ultimate criteria, effect on patient health (level 5) and cost-effectiveness (level 6).

5. Effect on patient health (level 5) is usually observable only after a substantial delay, especially for chronic illnesses, such as cardiovascular disease and cancer, which are currently the predominant causes of mortality in the United States. Also, it is the end result of a multistep process of health care. Because diagnostic tests occur near the beginning of the process, and some random variation enters into the results at every step, the effect of a diagnostic test on final outcomes is usually difficult to observe without an inordinate number of patients. For example, the current principal randomized controlled trial of computed tomographic (CT) screening for lung cancer requires some 50,000 patients and is expected to take 8 years and cost $200 million (7). Thus, effects on patient health (level 5) and cost-effectiveness (level 6) are uncommon as end points in experimental studies on the evaluation of diagnostic tests.

CLINICAL DECISION ANALYSIS AND COST-EFFECTIVENESS ANALYSIS

Instead, assessments of imaging technologies at levels 5 and 6 of the efficacy hierarchy are generally conducted by using decision analysis rather than direct experimental studies. Decision analysis (8–11) is an objective and systematic technique for combining the results of experimental studies that cover different health care steps to estimate effects of care processes more extensive than those directly studied in any single experimental research project. Cost-effectiveness analysis is a form of decision analysis that involves evaluation of the costs of health care, as well as the outcomes (12,13). What follows is a brief explanation of clinical decision analysis and cost-effectiveness analysis and the role they may play in technology assessment in radiology. Although we concentrate on cost-effectiveness analysis, the same methods and applications apply to decision analysis.

Cost-effectiveness analysis recognizes that the results of care are rarely 0% and 100% outcomes but rather are probabilistic (14). It involves the creation of algorithms, usually displayed as decision trees, as shown in Figure 1, which incorporate probabilities of events and, often, the valuations (usually called "utilities") of possible outcomes of these events. Individual or population-based preferences for certain outcomes and treatments are factored into these utilities.

Cost-effectiveness analysis can be divided into three basic steps: defining the problem, building the decision model, and analyzing the model.

Defining the Problem

For any cost-effectiveness analysis, one of the most difficult tasks is defining the appropriate research question. The issues to address in defining the problem are the

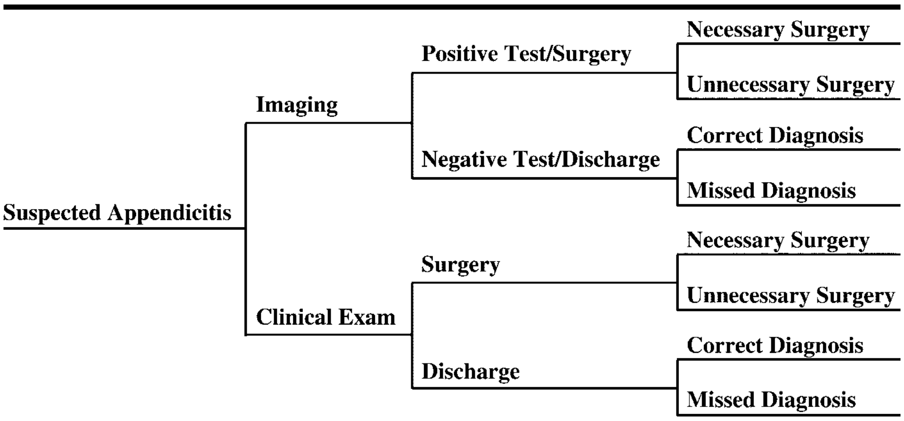


Figure 1. Example of a typical imaging decision analysis tree. In this example, an imaging test is compared with clinical examination for the correct diagnosis of acute appendicitis.

population reference case, strategies, time horizon, perspective, and efficacy (outcome) measures. The reference case is a description of the patient population the cost-effectiveness analysis is intended to cover. For example, the reference case for the cost-effectiveness analysis in Figure 1 consists of persons with acute abdominal pain seen in the emergency department.

The issue of strategies is, what are the care strategies that we should compare? Too many strategies may be confusing to compare. Too few may make an analysis suspect of missing possibly superior strategies. The decision tree in Figure 1 compares costs and outcomes of a clinical examination versus an imaging test for the diagnosis of acute appendicitis; in a fuller model, ultrasonography (US) and CT might be considered separate imaging strategies. In general, cost-effectiveness analysis and decision analysis address whether a new diagnostic test or treatment strategy should replace the current standard of care, in which case the current standard and the proposed new approach are the strategies to include. Alternatively, often the issue is which of a series of tests or treatments is best, and these then become the strategies to include.

The time horizon for which the cost-effectiveness analysis model is used to evaluate costs, benefits, and risks of each strategy must be stated and explained. Sometimes, the time horizon may be limited because of incomplete data, but this creates a bias against strategies with long-term benefits.

Finally, cost-effectiveness analysis allows costs to be counted from different perspectives. The perspective might be that of a third-party payer, in which case only insurance payments count as costs, or that of society, in which case all monetary costs, including those paid by the patient, count, and so—at least in some analyses—do nonmonetary costs, such as travel and waiting time involved in obtaining care.

Building the Cost-Effectiveness Analysis Model

Cost-effectiveness analysis is usually based on a decision tree, a visual representation of the research question (Fig 1). These decision trees are created and analyzed with readily available computer software, such as DATA (TreeAge Software, Williamstown, Mass). The tree incorporates the choices, probabilities of events occurring, outcomes, and utilities for each strategy being considered. Each branch of the tree must have a probability assigned to it, and each path in the tree must have a cost and outcome assigned. Data typically come from direct studies of varying quality, from expert opinion (which is usually unavoidable because some needed data values can not be obtained in any other way), and from some less directly relevant literature. For example, in Figure 1, the probability of a positive test result may be selected from published literature and added to the decision tree under the branch labeled "Positive Test/Surgery." Costs are frequently not ascertained directly, but rather are estimated by using proxies such as Medicare reimbursement rates or the charge and/or cost data of a hospital. Building the decision tree requires experience and judgment.

The complexity of cost-effectiveness analysis sometimes makes it difficult to understand and therefore undervalued (14,15). One way to improve understanding and allow readers to judge for themselves the value of a cost-effectiveness analysis model is to be explicit about the assumptions of the model. Many assumptions are needed simply because of limited data available to answer the research question.

Analyzing the Cost-Effectiveness Analysis Model

Once the model has been created, analysis should then include baseline analysis of cost and effectiveness and sensitivity analysis. The average cost and effectiveness for each strategy, considering all the outcomes to which it might lead, are computed simultaneously. We calculate averages by weighting the end probabilities of each branch and by summing for each strategy by moving from right to left in the tree. In cost-effectiveness analysis decision trees such as that in Figure 1, the costs and utilities for each outcome would be placed in the decision tree at the right end of each branch.

Possible results when comparing two strategies include the following: One strategy is less expensive and more effective than another, one strategy is more expensive and less effective, one strategy is less expensive but less effective, and one strategy is more expensive but more effective. The choice in the first two situations is clear, and the better strategy is called "dominant." The final two situations involve trade-offs in cost versus effectiveness, however. In these situations, one compares strategies by using the incremental cost-effectiveness ratio, which allows evaluation of the ratio of increase in cost to increase in effectiveness. What maximal incremental cost-effectiveness ratio is acceptable is open to debate, but for the United States, $50,000–$100,000 per year of life in perfect health (usually called a "quality-adjusted life-year") is commonly recommended as a maximum.

Almost all payers in the United States state that they consider only effectiveness, not cost. Implicitly, then, they accept an indefinitely high incremental cost-effectiveness ratio—it does not matter how much more expensive a strategy is, as long as it is the least bit more effective or the public demands it intensely.

The final task in cost-effectiveness analysis is sensitivity analysis. Sensitivity analysis consists of changing "parameter values" (numerical values, such as probabilities, costs, and valuation of outcomes) in the model to find out what effect they have on the conclusions. A model should be tested in this way for "robustness," or strength of its conclusions with regard to changes in its assumptions and uncertainty in the parameters taken from the literature or expert opinion. If a small change in the value of a parameter leads to a change in the preferred strategy of the model, then the conclusion is said to be sensitive to that parameter, and the conclusion is weak. Sensitivity analysis may persuade doubtful readers of the soundness of the conclusions of the model by showing that the researchers were thorough and unbiased and the conclusions are not sensitive to the assumptions or parameters the readers question. Often, however, sensitivity analysis will show that conclusions are not robust. Alternatively, another cost-effectiveness analysis, conducted by different researchers by using different assumptions and parameters (which is really a form of sensitivity analysis), will reach different conclusions. While discouraging, a similar situation is not uncommon with experimental studies (such as clinical research), with one study having findings different from another. Also, identification of the parameters and assumptions to which the results are sensitive can be very helpful, because it tells researchers what needs to be investigated further through experimental studies to reach reliable conclusions.

CHARACTERISTICS OF HIGH-QUALITY EXPERIMENTAL STUDIES

Whether an experimental study is intended to provide direct findings (principally, as we have seen, at efficacy levels 1 through 4) or to provide findings to be used as input into decision analysis and/or cost-effectiveness analysis (which are then used to assess level 5 and 6 efficacy), several design and operational considerations are important for the study to be of high quality and substantial value (2,16–19). Regrettably, the quality of studies on the evaluation of diagnostic imaging is very often poor (20–23). Therefore, radiologists should be aware of these considerations so that they may read the literature critically and also improve the quality of the technology assessment studies they conduct.

The most important considerations follow. We focus on studies of diagnostic accuracy, since these are most common and constitute the principal focus of radiologists, but most of what is said applies to experimental studies of other levels of the hierarchy of efficacy.

Patient Characteristics

Patients in a study should be like those in whom a test will be applied in practice. Often, in initial studies, a test is applied predominantly to very sick patients or completely healthy individuals. This "spectrum bias" exaggerates the real-world ability of the test to distinguish disease from health because intermediate cases that are less than totally clear cut are eliminated. As a result, initial reports on a new test are often overly optimistic. On the other hand, such spectrum bias can be useful in initial studies to ascertain if a test has any possible promise and to help establish the operating parameters at which the test works best.

Number of Cases

The number of cases included in studies should be adequate. Almost always, the smaller the number of cases, the larger the minimum difference that can reliably be observed. Before a study is begun, a statistician should be asked to perform a power calculation to ascertain the number of cases required to detect, with desired reliability, the minimum difference regarded as clinically important. Often, the number of cases included in actual studies is inadequate (22). Such studies are referred to as "underpowered" and can lead to errors.

Design Considerations

Prospective studies are almost always preferable to retrospective studies. "Well begun is half done" carries a corollary that "poorly begun is hard to salvage." In a retrospective study, one has to work from someone else's design and data collection, and these are typically far from optimal from the standpoint of your purposes.

The temptation to include in the research everything that might be studied should be resisted, lest the study collapse from its own complexity.

Often, the purpose of a study is to compare two diagnostic tests—for example, to compare a proposed new test with an established one. In this situation, unless data on patient health outcomes and cost must be directly obtained, an optimal design consists of applying both tests to all study patients, with interpretation of each test performed while blinded to the results of the other. In contrast, the common practice of using "historical controls" to represent the performance of the established test is usually a poor choice. The patient population in the historical control may be different, and the execution of the historical series may not meet standards of current best practice.

Reference Standard

The reference standard (sometimes less formally called the "gold standard") needs to be chosen carefully. While a perfect reference standard—one with 100% accuracy—often cannot be attained, it is important to do as well as possible. Methodologists routinely warn (4,22,24) that a reference standard that is dependent, even in part, on the test(s) being evaluated involves circular reasoning, and they say it is therefore seriously deficient, but they note that such standards are nonetheless not infrequently used.

Timing

Timing is important because diagnostic imaging is a field that is changing relatively rapidly. There is little point in undertaking a large-scale study when a new technique is in the initial developmental stage and is changing particularly rapidly; results will be obsolete before they are published. On the other hand, it is not wise to wait until a technique is fully mature because, by then, it will often be widely disseminated, making the study too late for its results to readily influence general clinical practice. Use of techniques that lead to rapid completion of a study, such as gathering data from multiple sites, is highly desirable because imaging evolves relatively rapidly.

Efficacy and Effectiveness

Most evaluations of diagnostic tests—and of any other medical care—are studies of efficacy, which is defined as results obtained under ideal conditions, such as those of a careful research project. Initially, efficacy is important to ascertain, but ultimately, one would want to know effectiveness, which is defined as results obtained in ordinary practice. Effectiveness is usually poorer than efficacy. For example, studies in individual academic institutions—that is, efficacy studies—showed that abdominal CT for patients suspected of having appendicitis significantly reduced the perforation rate and unnecessary surgery rate (25,26), but a study of essentially all hospital discharges in Washington state—that is, an effectiveness study—showed no improvement in either rate between 1987 and 1998, a period when laparoscopy and cross-sectional imaging techniques, including CT, became widely available (27). The systematization necessary for an organized study tends to preclude observation of effectiveness—the study protocol ensures uniform application of the test with its parameters set at optimal levels, and people are generally more careful and consistent and do better

Project Selection and Definition Phase

- Evaluate measures of diagnostic accuracy that are highly clinically relevant
- Use an experienced statistician, involving him or her from the beginning
- Involve the treating physicians from the beginning

Study Design and Start-Up Phase

- Specify the protocol carefully and in detail
- Use a sophisticated statistical analysis, including multivariate techniques
- Hold face-to-face meetings of the full range of study participants at critical points
- Conduct extensive pretesting

Project Operation Phase

- Use existing experienced data management and statistical analysis centers
- Include multiple sites, preferably involving some nonacademic participants
- Hold periodic telephone conference calls
- Have "fill-ins" available in case initial participants drop out and in case—as almost always happens—participants obtain fewer patients per month than they anticipate
- Require participants to have a data manager on site
- Send participants periodic reminders about overdue forms

Figure 2. Additional procedures for enhancement of study quality and rapidity, with particular reference to a study of substantial scale.

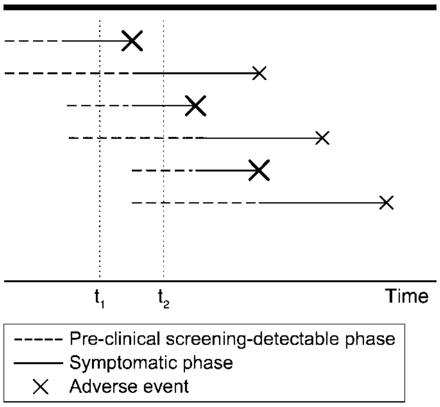


Figure 3. Example of length bias. Half of the cases are the more indolent form (longer preclinical phase, longer symptomatic phase, and less severe adverse events, as shown by a smaller x). At any point in time (*t1* and *t2* are randomly chosen points in time), however, two-thirds of the cases detectable only with screening are indolent.

when they know their activity is being observed (this is called the Hawthorne effect).

Figure 2 lists some additional important considerations for high-quality studies. Sunshine and McNeil (16) discuss the above considerations and those in Figure 2 in more detail.

SCREENING

Screening (28,29) is the performance of a diagnostic test in an asymptomatic population with the aim of reducing morbidity and/or mortality from disease. The requirements of efficacious screening are somewhat different from those of “conventional” diagnostic testing—that is, testing applied to symptomatic patients. These differences apply to the diagnostic test, available treatment, and evaluation of the test.

The Test

Because the prevalence of disease in a screening population is very low—for example, approximately one-half percent in screening mammography—a screening test must be highly specific. Otherwise, false-positive findings will greatly outnumber true-positive findings (even at the relatively high 90%–95% specificity rate for mammography—ie, 5%–10% recall rate—false-positive findings outnumber true-positive findings by 10–20 to 1), and the cost and morbidity of working up patients with false-positive findings will outweigh the gains from early detection in those with true-positive findings. Similarly, the cost and morbidity of the screening test itself (which apply to every patient screened) must be relatively low; otherwise, they will outweigh the gains of screening, which can occur only for the very small percentage of patients with true-positive findings.

In contrast, sensitivity can be modest. For example, screening mammography has an approximate 75% sensitivity, yet it allows us to identify three of every four possible breast cancers that could be detected if the test were perfectly (100%) sensitive. These requirements for a screening test can be somewhat eased if a high-risk population is identified, because the proportion of true-positive findings will increase. Note that while a screening test optimally has high specificity and may only need modest sensitivity, an optimal diagnostic test for symptomatic patients should have a high sensitivity, but the specificity may be modest.

Treatment

Oddly, the available treatment must be intermediate in efficacy. If treatment is fully efficacious—more specifically, if treatment of symptomatic patients is as efficacious and no more costly than the presymptomatic treatment made possible by screening—then nothing is to be gained by identifying disease before it becomes symptomatic. Conversely, if treatment is completely inefficacious—that is, there is no useful treatment for even presymptomatic disease—there is also no possible gain from screening. Screening can only be beneficial if treatment of presymptomatic disease is more efficacious than treatment of symptomatic disease (29–31). (However, some hold that screening for untreatable genetic diseases and other untreatable diseases can be reasonable because parents can alter reproductive behavior and patients can gain more time to prepare for the consequences of disease.) Given these requirements regarding treatment effectiveness for screening to be sensible, new developments in treatment—for example, the introduction of pharmaceuticals such as donepezil hydrochloride (Aricept; Eisai America, New York, NY) that slows the previously unalterable rate of progression of Alzheimer disease—can completely alter the relevance of screening.

Evaluation of Screening

In general, the efficacy of treatment of presymptomatic disease relative to that of symptomatic disease is not known, although this is a critical issue for screening, as indicated in the previous paragraph. The reason for the lack of knowledge is as follows: if screening has not been done previously, relative efficacy simply is not known because presymptomatic cases have not been identified and treated. On the other hand, if the issue is introduction of a more sensitive screening test, one does not know the efficacy of treating the additional, presumably less advanced cases the new test detects. Partly for this reason, evaluation of screening generally has to consist of a randomized controlled trial in which *(a)* the intervention consists of the test and the treatment in combination and *(b)* the end point studied is the death rate, morbidity, or other adverse outcome(s) from the disease being screening for in the intervention population compared with the rates in the control population.

Biases

Three well-known biases (30,32,33) also generally necessitate this randomized controlled trial study design for evaluation of screening tests and generally preclude the use of other end points, such as 5-year survival from time of diagnosis. These three biases should be understood by all radiologists.

“Lead-time bias” refers to the fact that screening will allow detection of disease earlier in its natural history than will waiting for symptoms, so any measurement from time of diagnosis will be biased in favor of screening, regardless of the effectiveness of treatment. Consider an oversimplified example: For lung cancer, 5-year survival from diagnosis is currently 10%–20%. Assume that CT screening advances diagnosis by 5½ years, but treatment has absolutely no value. Then 5-year survival would nonetheless increase to essentially 100% with screening. In short, survival time in a screened group will incorrectly appear to be better than that in a nonscreened group.

"Overdiagnosis bias" or "pseudodisease" (29,31) refers to the fact that applying a diagnostic test to asymptomatic individuals will identify "positive cases" that will never become clinically manifest in a person's lifetime. Prostate cancer provides a striking example. It is the most common nonskin malignancy in men in the United States, affecting 10% of them, but careful histopathologic examination at autopsy shows microscopic prostate cancers in nearly 50% of men over the age of 75 years (34). If an imaging test as sensitive as histologic examination at autopsy were developed, but early detection had absolutely no effect on outcomes, the percentage of "cases" showing adverse outcomes would nonetheless decrease by four-fifths—but only because four-fifths of the "cases" never would have shown any effects of the disease in the absence of screening and treatment. The general point is that, because of overdiagnosis bias, any study of the outcome of cases identified with a screening test will be biased toward screening, for many of the cases identified with screening would never have had any adverse outcomes, even in the absence of treatment. Incidentally, the morbidity and cost of treating such cases is one of the negative consequences of screening.

"Length bias" can be thought of as an attenuated form of pseudodisease. It arises because cases of a disease vary in aggressiveness, with the faster-progressing cases typically also having a natural history with greater morbidity and mortality. Cases detected with screening are typically disproportionately indolent. This is because slow-progressing cases remain longer in the presymptomatic phase in which they are detectable only with screening and do not manifest symptoms. Thus, a test that helps identify asymptomatic cases disproportionately uncovers indolent cases, as Figure 3 shows. Hence, cases detected with screening disproportionately have a relatively favorable prognosis, regardless of the effectiveness of treatment. Thus, any study of outcomes in cases detected with screening (vs those detected when symptoms occur) will be biased toward screening.

Other Considerations

While change in morbidity or mortality from the disease being screened for is the prime measure of the effect of screening, changes in other morbidity and mortality possibly caused by screening and/or treatment should also be considered. Concerns of this type include surgical complications, chemotherapy toxicity, radiation treatment–induced secondary cancers, radiation dose from screening, patient anxiety, and changes in patient satisfaction.

The percentage reduction in the risk of an adverse effect from the disease being screened for, called "relative risk reduction," is a common measure of the benefit of screening, but this measure needs to be set in context (35). For example, if screening reduces an individual's risk of dying of a particular disease over the next decade from 1.0% to 0.4%, that is a 60% decrease in relative risk, but only 0.6 of a percentage point increase in the probability of surviving the decade.

In conclusion, for any health care intervention, including diagnostic imaging tests, the ultimate questions are, "How much does this do to improve the health of people?" and "How much does it cost for that gain in health?" By using the methods described in this article, we have the ability to answer these questions as we assess the remarkable imaging technologies available today.

References

1. Perry S, Thamer M. Medical innovation and the critical role of health technology assessment. JAMA 1999; 282:1869–1872.
2. Eisenberg J. Ten lessons for evidence-based technology assessment. JAMA 1999; 282: 1865–1869.
3. Clancy CM, Eisenberg JM. Outcomes research: measuring the end results of health care. Science 1998; 282:245–246.
4. Hillman BJ. Outcomes research and cost-effectiveness analysis for diagnostic imaging. Radiology 1994; 193:307–310.
5. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11:88–94.
6. Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. AJR Am J Roentgenol 1994; 162:1–8.
7. American College of Radiology Imaging Network. Contemporary screening for the detection of lung cancer. Available at: *www.acrin-nlst.org/6654factsheet.html*. Accessed August 20, 2003.
8. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia, Pa: Saunders, 1980.
9. Hunink MG, Glasziou PP, Siegel J, et al. Decision making in health and medicine: integrating evidence and values. Cambridge, England: Cambridge University Press, 2001.
10. Chapman GB, Sonnenberg FA. Decision making in health care: theory, psychology, and applications. Cambridge, England: Cambridge University Press, 2000.
11. Janne D'Othee B, Black WC, Pirard S, Zhuang Z, Bettman MA. Decision analysis in radiology. J Radiol 2001; 82:1693–1698.
12. Singer ME, Applegate KE. Cost-effectiveness analysis in radiology. Radiology 2001; 219:611–620.
13. Soto J. Health economic evaluations using decision analytic modeling: principles and practices—utilization of a checklist to their development and appraisal. Int J Technol Assess Health Care 2002; 18:94–111.
14. Kleinmuntz B. Clinical and actuarial judgment. Science 1990; 247:146–147.
15. Tsevat J. SMDM presidential address: hearsay or heresy—are health decision scientists too left-brained? Med Decis Making 2003; 23:83–87.
16. Sunshine JH, McNeil BJ. Rapid method for rigorous assessment of radiologic imaging technologies. Radiology 1997; 202:549–557.
17. Baum RA, Rutter CM, Sunshine JH, et al. Multicenter trial to evaluate vascular magnetic resonance angiography of the lower extremity. JAMA 1995; 274:875–880.
18. Lilford RJ, Pauker SG, Braunholtz DA, Chard J. Decision analysis and the implementation of research findings. BMJ 1998; 317:405–409.
19. Blackmore CC. The challenge of clinical radiology research. AJR Am J Roentgenol 2001; 176:327–331.
20. Kent DL, Haynor DR, Longstreth WT, Larson EB. The clinical efficacy of magnetic resonance imaging in neuroimaging. Ann Intern Med 1994; 120:856–871.
21. Holman BL. The research that radiologists do: perspective based on a survey of the literature. Radiology 1990; 176:329–332.
22. Blackmore CC, Black WC, Jarvik JG, Langlotz CP. A critical synopsis of the diagnostic and screening radiology outcomes literature. Acad Radiol 1999; 6:S8–S18.
23. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Radiology 2003; 226:24–28.
24. Sox H, Stern S, Owens D, Abrams HL. Monograph of the council on health care technology: assessment of diagnostic technology in health care-rationale, methods, problems, and directions. Washington, DC: National Academy Press, 1989.
25. Rao PM, Rhea JT, Novelline RA, Mostafavi AA, Lawrason JN, McCabe CJ. Helical CT combined with contrast material administered only through the colon for imaging of suspected appendicitis. AJR Am J Roentgenol 1997; 169:1275–1280.
26. Sivit CJ, Siegel MJ, Applegate KE, Newman KD. When appendicitis is suspected in children. RadioGraphics 2001; 21:247–262.
27. Flum DR, Morris A, Koepsell T, Dellinger EP. Has misdiagnosis of appendicitis decreased over time? a population based analysis. JAMA 2001; 286:1748–1753.
28. Herman CR, Gill HK, Eng J, Fajardo LL. Screening for preclinical disease: test and disease characteristics. AJR Am J Roentgenol 2002; 179:825–831.
29. Black WC, Welch HG. Screening for disease. AJR Am J Roentgenol 1997; 168:3–11.
30. Black WC, Ling A. Is earlier diagnosis really better? the misleading effects of lead time and length biases. AJR Am J Roentgenol 1990; 155:625–630.
31. Morrison AS. Screening in chronic disease. New York, NY: Oxford University Press, 1992; 125–127.
32. Black WC, Welch HG. Advances in diagnostic imaging and overestimation of disease prevalence and the benefits of therapy. N Engl J Med 1993; 328:1237–1243.
33. Morrison AS. The effects of early treatment, lead time and length bias on the mortality experienced by cases detected by screening. Int J Epidemiol 1982; 11:261–267.
34. Brant WE, Helms CA, eds. Fundamentals of diagnostic radiology. 2nd ed. Philadelphia, Pa: Lippincott Williams & Wilkins, 1999; 825.
35. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988; 318:1728–1733.

Sample Size Estimation: A Glimpse beyond Simple Formulas[1]

John Eng, MD

Index terms:
Radiology and radiologists, research
Receiver operating characteristic (ROC) curve
Statistical analysis

Published online
10.1148/radiol.2303030297
Radiology 2004; 230:606–612

Abbreviations:
A_z = area under ROC curve
CNR = contrast-to-noise ratio
ROC = receiver operating characteristic

Small increments in the complexity of clinical studies can readily take sample size estimation and statistical power analysis beyond the capabilities of simple mathematic formulas. In this article, the method of simulation is presented as a general technique with which sample size may be calculated for complex study designs. Applications of simulation for determining sample size requirements in studies involving correlated data and comparisons of receiver operating characteristic curves are discussed.

In a previous article in this series (1), I discussed the fundamental concepts involved in determining the appropriate number of subjects that should be included in a clinical investigation. This number is known as the sample size. In the earlier article (1), the necessity for considering sample size, how certain study design characteristics affect sample size (Fig 1), and how to calculate sample size for several simple study designs were discussed. Also discussed was how sample size is related to statistical power, which is the sensitivity of detecting a statistically significant difference in a comparative study when a difference is truly present.

In this article, I will first discuss some important consequences of sample size and power calculations, then go on to discuss issues that arise when these basic principles are applied to real clinical investigations, which are often more complex than the simple situations covered in the previous article and in introductory biostatistics textbooks. My intent is to provide an overview and appreciation of some of the advanced statistical methods for handling some of the complex situations that arise. Since advanced statistical methods for sample size or power calculations cannot receive comprehensive treatment in the setting of an overview article, an investigator needing such methods is advised to seek help from a statistician early in the research project. However, I hope the material herein will at least help bridge the knowledge gap between investigator and statistician so that their interaction can be more productive.

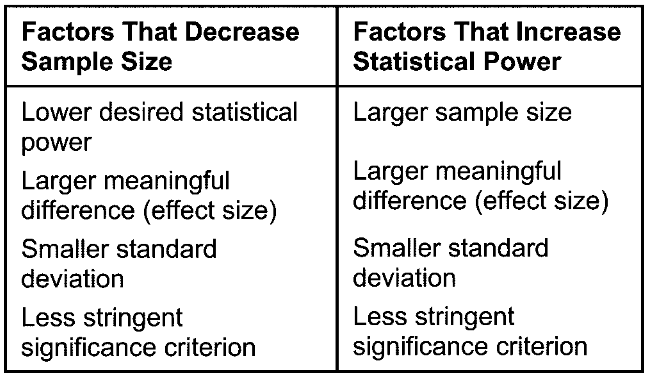
CONSEQUENCES OF SAMPLE SIZE CALCULATIONS

Academic and Ethical Importance

In conjunction with a well-defined research question (2), an adequate sample size can help ensure an academically interesting result, whether or not a statistically significant difference is eventually found in the study. The investigator does not have to be overly concerned that the study will only be interesting (and worth the expenditure of resources) if its results are “positive.” For example, suppose a study is conducted to see if a new imaging technique is better than the conventional one. Obviously, the study would be interesting if a statistically significant difference was found between the two techniques. But if no statistically significant difference is found, an adequate sample size allows the investigator to conclude that no clinically important difference was found rather than wonder whether an important difference is being hidden by an inadequate sample size.

An inadequate sample size also has ethical implications. If a study is not designed to include enough individuals to adequately test the research hypothesis, then the study unethically exposes individuals to the risks and discomfort of the research even though there is no potential for scientific gain. Although the connection between research ethics

[1] From the Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University, Central Radiology Viewing Area, Room 117, 600 N Wolfe St, Baltimore, MD 21287. Received February 21, 2003; revision requested April 10; revision received July 18; accepted July 21. **Address correspondence to** the author (e-mail: *jeng@jhmi.edu*).



Factors That Decrease Sample Size	Factors That Increase Statistical Power
Lower desired statistical power	Larger sample size
Larger meaningful difference (effect size)	Larger meaningful difference (effect size)
Smaller standard deviation	Smaller standard deviation
Less stringent significance criterion	Less stringent significance criterion

Figure 1. Study design characteristics that affect sample size and statistical power.

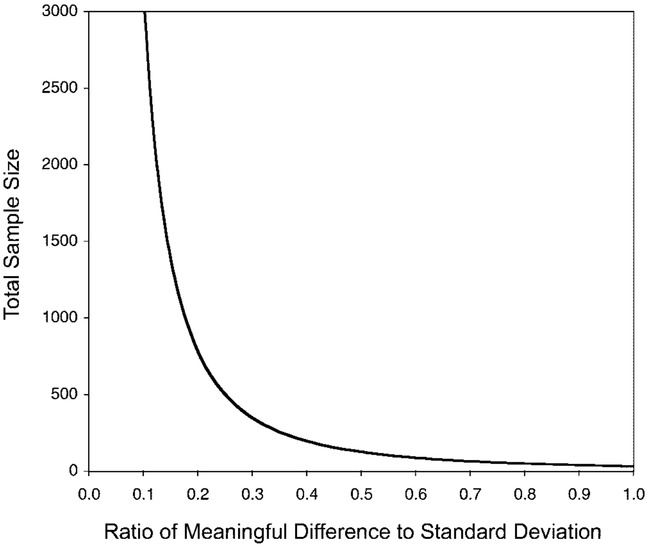


Figure 2. Graph shows the relationship between sample size and the ratio of meaningful difference to SD for studies in which means are compared. The graph was created by using Equation (1) and illustrates the fact that sample size increases exponentially as the ratio decreases.

and adequate sample size has been recognized for at least 25 years (3), the performance of clinical trials with inadequate sample sizes remains widespread (4).

Practical Consequences of Mathematic Properties

A more intuitive understanding of the determinants of sample size can be obtained through closer inspection of the formulas for sample size. We saw in the previous article (1) that when the outcome variable of a comparative study is a continuous value for which means are compared, the appropriate sample size (5) is given by

$$N = \frac{4\sigma^2(z_{\text{crit}} + z_{\text{pwr}})^2}{D^2}, \qquad (1)$$

where N is the total sample size (ie, the total of the two comparison groups), D is the smallest meaningful difference between the two means being compared, σ is the SD of each group, and z_{crit} and z_{pwr} are constants determined by the specified significance criterion (eg, .05) and desired statistical power (eg, .8), respectively. Since z_{crit} and z_{pwr} are independent of the properties of the data, sample size depends only on the ratio between the smallest meaningful difference and the SD (Fig 2).

Furthermore, because the ratio is in an inverse exponential relationship to sample size, anything that can be done to decrease the SD or increase the meaningful difference can substantially reduce the required sample size. The SD could be decreased by reducing measurement variability (eg, by using more precise instruments or procedures) and/or by selecting a more homogeneous study population. The meaningful difference could be increased by employing more sensitive instruments or procedures.

Another property of the comparison of means is that for a given SD, only the arithmetic difference between the comparison groups affects the sample size. For example, the sample size would be the same for detecting a systolic blood pressure difference of 10 mm Hg whether it is to be measured in normotensive individuals (eg, 110 vs 120 mm Hg) or hypertensive individuals (eg, 170 vs 180 mm Hg).

When proportions are being compared—a common task in clinical imaging research—the sample size depends on both the smallest meaningful difference between the proportions and the size of the proportions themselves. That is, when proportions are being compared, in contrast to when means are being compared, the sample size depends not just on the difference alone. The sample size increases dramatically as the meaningful difference between proportions is made smaller (Fig 3). The sample size also increases if the two proportions being compared (ie, the mean of the two proportions) are close to 0.5.

Retrospective Power Analysis

In sample size calculations, appropriate values for the smallest meaningful difference and the estimated SD are often difficult to obtain. Therefore, the formulas are sometimes applied after the study is completed, when the difference and SD actually observed in the study can be substituted in the appropriate sample size formula. Since sample size is also known after the study is completed, the formula will yield statistical power. In this case, power refers to the sensitivity of the study to enable detection of a statistically significant difference of the magnitude observed in the study. This activity, known as retrospective power analysis, is sometimes performed to aid in the interpretation of the statistical results of a study. If the results were not statistically significant, the investigator might explain the result as being due to a low power.

However, it can be shown that the retrospective power—essentially an observed quantity—is inversely related to the observed *P* value (6). The retrospective power tends to be large in any study with a small (statistically significant) observed *P* value. Conversely, the retrospective power tends to be small in any study with a large (statistically insignificant) observed *P* value. Therefore, the observed retrospective power cannot provide any information in addition to the observed *P* value (7,8). The important point is that the smallest meaningful difference is not the same as the observed difference: The former must be set before the study is

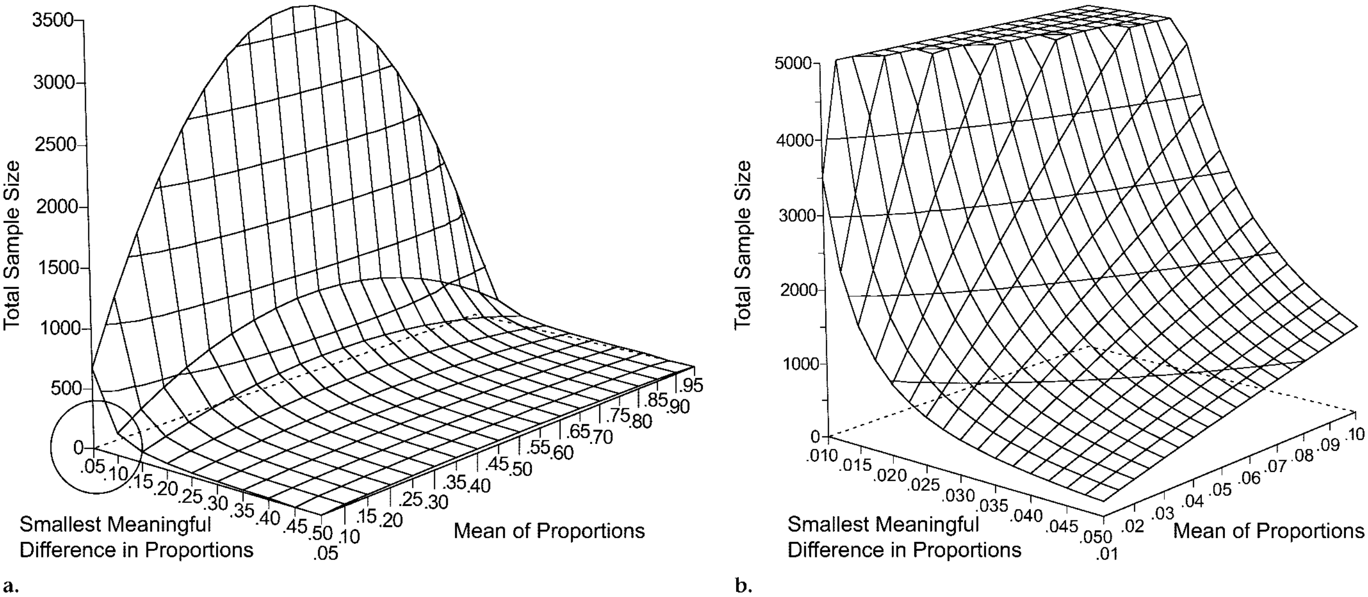


Figure 3. (**a**) Graph shows relationship between sample size and proportions being compared in a study involving comparison of proportions. Sample size increases dramatically as the meaningful difference decreases. Sample size also increases if the proportions being compared (ie, the mean of the two proportions) are near 0.50. (**b**) Extension of the circled corner of the graph in **a**, with x and y axes magnified; this corner corresponds to a region that is of particular interest to the design of clinical investigations.

conducted and is not determined after the study is completed.

Even though calculating the retrospective power is problematic, it remains important to consider the issue of adequate sample size when one is faced with a study whose results indicate there is no difference between comparison groups. Fortunately, several statistical approaches are available to guide the reader in terms of whether or not to "believe" a study that yields negative results (9). These approaches involve calculating CIs or performing χ^2 tests.

USE OF SIMULATION TO DETERMINE SAMPLE SIZE FOR COMPLEX STUDY DESIGNS

In contrast to the importance of considering sample size and statistical power, relatively few formulas exist for calculating them (10). The most simple formulas, such as those discussed previously (1) and in many introductory biostatistics textbooks, concern the estimation and comparison of means and proportions and are, fortunately, applicable to many situations in clinical radiology research. Beyond these simple formulas, methods have been established to determine sample size for general fixed-effect linear statistical models (of which the *t* test, ordinary linear regression, and analysis of variance are special cases), two-way contingency tables (of which the analysis of a 2 × 2 table with the χ^2 test is a special case), correlation coefficient analysis, and simple survival analysis (10). Approximations exist for some other statistical models, most notably logistic regression, but the accuracy of these approximations may be difficult to establish in all situations.

Thus, the list of all statistical tests for which exact sample size calculation methods exist is much smaller than the list of all statistical tests. When no formula exists, as often happens for moderately complex statistical designs, the investigator may try to perform a sample size analysis for a simplified version of the study design and hope that the sample size can be extrapolated to the actual (more complex) study design being planned.

For situations without corresponding formulas, it is becoming more common to estimate sample size by using the technique of simulation (11). The simulation approach is powerful because it can be applied to almost any statistical model, regardless of the model's complexity. In simulation, a mathematic model is used to generate a synthetic data set simulating one that might be collected in the study being planned. The mathematic model contains the dependent and independent variables being measured, along with estimates of each variable's SD. The synthetic data set contains the same number of subjects as the planned sample size.

The planned statistical analysis is performed with this synthetic data set, and a *P* value is determined. As usual, the null hypothesis is rejected if the *P* value is less than a certain criterion value (eg, $P <$ .05). This process is repeated a large number of times (perhaps hundreds or thousands of times) by using the mathematic model to generate a different synthetic data set for each iteration. The statistical power is equal to the percentage of these data sets in which the null hypothesis is rejected. In effect, simulation employs a mathematic model to virtually repeat the study an arbitrarily large number of times, allowing the statistical power to be determined essentially by direct measurement.

Since a real data set would contain random statistical error, random statistical error must be modeled in the synthetic data sets. To accomplish this in simulation, a random-number generator is used to add random error ("noise") to each synthetic data set. Because of their heavy reliance on random-number generators, simulation methods are also known as Monte Carlo methods, after the city in which random numbers also play an important role.

Let us consider a simple example. Suppose we are planning a clinical study to compare the contrast-to-noise ratio (CNR) between two magnetic resonance imaging pulse sequences used to examine each subject in a group of subjects. We would like to know the statistical power of the study to detect a smallest meaningful CNR difference of 2. We would like to plan a study with a power of .8 for detecting this smallest meaningful difference. We have resources to comfortably recruit and evaluate approximately 12 subjects. Suppose that from our previous experience with the pulse sequences, we estimate the SD of the CNR difference to be 4.

The statistical model for this study is

$$\mathrm{CNR}_i = D + \epsilon_i, \qquad (2)$$

where CNR_i is the observed CNR difference for subject *i* (of the 12 subjects), *D* is the true CNR difference (in this example, 2), and ϵ_i is the random error associated with the observation of subject *i*. To run the simulation, we use a normally distributed random-number generator for ϵ_i that generates a different normally distributed random number for each of the 12 observations. The mean of the numbers generated by the random number generator is 0 and the SD is 4, which we estimated on the basis of previous experience. With these 12 random numbers, we can generate a synthetic data set of 12 observations by using Equation (2). The simulated data set is then subjected to a *t* test, and the resulting *P* value is recorded.

The entire simulation process is then repeated, say, 1,000 times. The *P* value is recorded after each iteration. After completing the iterations, the *P* values are examined to determine what proportion of the iterations resulted in the detection of a statistically significant difference (indicated by $P < .05$); this proportion is equal to the power. The simulation for this example was performed with Stata version 7.0 (Stata, College Station, Tex), and the results are shown in the first line of the Table. In this example, the null hypothesis is rejected in 343 of the 1,000 iterations. Therefore, the statistical power of the *t* test, given the conditions of this example, is .34 (Table).

Obviously, it would have been easier to use the formula for comparison of means. But the advantage of simulation is the ability to consider more complex statistical models for which there are no simple sample size formulas. This ability is especially important because seemingly small changes in study design can cause the simple sample size formulas to become invalid.

Results of Simulations of Hypothetical Study in Which Difference between Two Imaging Techniques Is Being Compared within Each Subject

No. of Subjects	No. of Observations per Subject	Correlation between Observations within Each Subject	Statistical Test	No. of Simulated Data Sets	No. of Statistically Significant Results	Power of Statistical Test
12	1	0.0	*t* Test	1,000	343	.34
12	4	0.5	*t* Test*	1,000	846	.85
12	4	0.5	Regression†	1,000	528	.53
22	4	0.5	Regression†	1,000	829	.83

* Not the appropriate statistical test for these data, but done for purposes of illustration.
† Linear regression with adjustment for clustering by subjects.

Returning to the example, we note that the estimated power of our study is lower than desired. The only way to improve the power, given our assumptions, is to increase the number of observations. (For the moment, we only have resources to study 12 subjects.) So, we decide to make four measurements of CNR difference per subject. This strategy will increase the number of observations by a factor of four and will result in an increase in power. However, it is important to realize that this data collection strategy is not the same as increasing the number of subjects by a factor of four, because the four observations within each subject are not independent of one another. Within each subject, the observations are likely to be more similar to each other than to the observations in the other subjects. In statistical terms, this lack of independence is called correlation.

Because of correlation, an additional observation in the same subject does not provide as much additional information as an additional observation in a different subject. The more similar the observations within each subject are, the less additional information will be provided by the repeated observation. If the observations within each subject are identical (100% correlated), then the study would have the same results (and sample size) as it would without the repeated observations, so there would be no benefit from repeating the measurement for the same subjects. Conversely, if the repeated observations within each subject were completely uncorrelated (0% correlation), then the results (and sample size) would be identical to those of a study with the same total number of observations but with enough additional subjects that only one observation per subject is used.

Simulation can easily account for the correlation of the four observations within each subject. The statistical model used is a slight variation of Equation (2):

$$\mathrm{CNR}_{ij} = D + \epsilon_{ij}, \qquad (3)$$

where CNR_{ij} is the observed CNR for subject *i* (of the 12 subjects) and repetition *j* (of the four repetitions), *D* is the true CNR difference, and ϵ_{ij} is the random error associated with each of the 48 observations. As in Equation (2), ϵ_{ij} is generated by a normally distributed random-number generator having a mean of 0 and an SD of 4. In Equation (3), however, ϵ_{ij} is calculated in such a way that each error term ϵ_{ij} is correlated with the other error terms within the same subject. Correlation of the error terms is the mathematic mechanism for generating correlation in the observations. The amount of correlation is indicated by the correlation coefficient ρ. In this example, we assume a moderate amount of correlation ($\rho = 0.5$) between observations made within each subject. The results of the simulation are shown in the Table.

With an ordinary *t* test, there appears to be enough power in the proposed study design (Table). But an ordinary *t* test is inappropriate in this case because it treats each of the 48 observations as independent, ignoring the correlation between the four observations within each subject. An appropriate method that accounts for correlation is linear regression with an adjustment for clustering. When this type of linear regression is applied instead of the *t* test, the simulation reveals that the power is actually .5 (Table), which is lower than the desired power of .8. Results of further simulations indicate that increasing the number of subjects from 12 to 22 would result in adequate power (Table).

A discussion of statistical tests that adjust for correlation within subjects is beyond the scope of this article. However, without a simple formula for sample size, and even without extensive knowledge of the statistical test to be used, simulation still enabled the accurate determina-

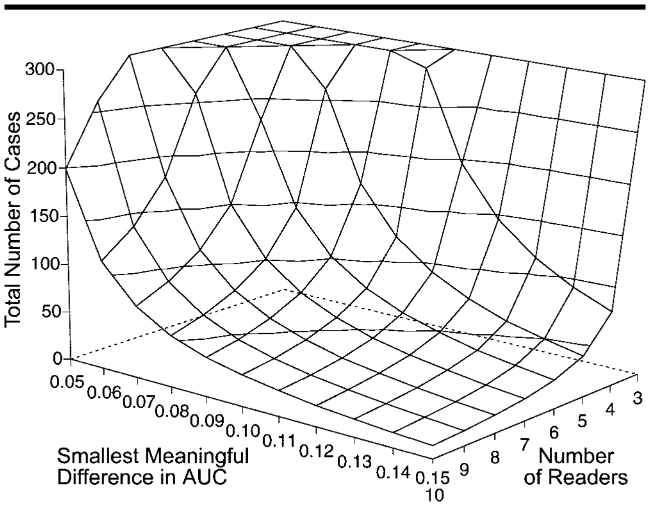


Figure 4. Graph shows the relationship between sample size, smallest meaningful difference, and number of readers in an ROC study in which A_z is used as the index of accuracy. Sample sizes were calculated by using the method described in the Appendix. The values used for all variables except J and Δ are the same as those in the example in Table A1.

tion of power in the preceding example; this demonstrates the utility and generalizability of simulation. In addition, the effects of the use of potentially inappropriate statistical analyses were also able to be examined.

One of the barriers to performing simulation is the requirement of iterative computation, which in turn requires fast computers. This barrier is becoming much less important as the speed of commonly available computers continues to increase. Even when the barrier of computational speed is overcome, simulation is successful only if the assumed statistical model accurately describes the study design being evaluated. Therefore, appropriate attention must be paid to establishing the model's validity. Fortunately, it is often easier to develop a mathematic model for a statistical situation (from which it is a straightforward process to determine power and sample size with simulation) than to search for a specific method or formula, if one even exists, for calculating sample size. In the preceding example, the introduction of correlation substantially increased the complexity of the analysis from a statistical point of view but caused only a minor change in the mathematic model and the subsequent simulation.

SAMPLE SIZE CALCULATIONS FOR READER STUDIES

A reader study involving the calculation of receiver operating characteristic (ROC) curves is another kind of study design that is fairly common in radiology and for which sample size and statistical power are difficult to determine. The area under the ROC curve *(A_z)* is commonly used as an indicator of the accuracy of a diagnostic test. A typical ROC study involves multiple readers interpreting images obtained in the same group of subjects who have all undergone imaging with two or more imaging techniques or conditions. The purpose of the study is to compare the imaging techniques. The difficulty in determining sample size and statistical power is a result of the fairly complicated computational process required to calculate A_z and the complicated correlations among the observations. In such a study, each reader generates multiple observations, and, likewise, each subject has a part in multiple observations. Therefore, correlation can occur simultaneously among the observations (readings) within the same reader and among the observations within the same subject.

One approach to sample size analysis in complex ROC studies involves an approximation performed by using the *F* distribution (12,13). Sample size tables created by using this method have been published (14); this method can also be used to calculate sample sizes for situations not addressed by such tables (Appendix). The method may be used to examine the trade-off between sample size, smallest meaningful difference, and number of readers (Fig 4). For most clinical investigations, it is likely to be difficult to include more than 10 readers or 100 cases. Given these constraints, we see that any ROC study will require at least four readers, even with a large meaningful difference of 0.15 in A_z. At the other extreme, the smallest meaningful difference in A_z that can be detected with 10 readers and 100 cases is 0.07. These two generalizations are based on many assumptions (Fig 4). More cases or readers are required if the interobserver variability or intraobserver variability is higher than assumed. Fewer cases or readers are required if the average A_z (ie, accuracy of the readers) is higher than assumed.

Another major method for analyzing data from an ROC study with multiple readers and multiple cases is the jackknifed pseudovalue method (15). In this method, the data are mathematically transformed into pseudovalues that can be analyzed by using a relatively straightforward analysis of variance. Reducing the problem of ROC analysis to a more manageable analysis of variance is a strength of the pseudovalue method. A disadvantage is the lack of exact, or even approximate, formulas for determining sample size and statistical power. The performance and validity of the pseudovalue method have been examined with simulations (16,17), so simulation could also provide a viable method for determining sample size and power for the pseudovalue method.

CONCLUSION

In contrast to the wide variety of statistical tools available to the clinical investigator, relatively few formulas exist for the exact calculation of the sample size and statistical power of a given study design. As demonstrated by the example of correlated data, a frequent occurrence in clinical research, it is relatively easy to construct a study design for which no simple formula exists. The availability of fast computers makes the iterative process of simulation a viable general method for performing sample size and power analysis for complex study designs.

At first glance, simulation may appear artificial and therefore suspicious because it relies on an equation and many assumptions about the terms in the equation, particularly the terms related to the variability of the components of the model. It should be noted, however, that similar (although perhaps less complex) mathematic models are the foundation

TABLE A1
Definition of Variables in Calculation of Sample Size for ROC Study in Which a Number of Readers Interpret Same Set of Cases Obtained by Using Two Different Imaging Techniques

Variable	Definition	Example*
Main design variables		
J	Number of readers	4
Δ	Smallest meaningful difference between A_z values associated with the two imaging techniques	0.15
w_b	Interobserver variability expressed as expected difference between A_z of most accurate observer in study and A_z of least accurate observer	0.05
K	Number of times each case is read by each reader for each imaging technique, typically equal to 1	1
θ	Expected average A_z for the two imaging techniques	0.75
R	Ratio of number of negative cases to number of positive cases in study, often equal to 1	1
Variables with suggested values		
w_w	Intraobserver variability expressed as expected difference between A_z values of observer who interprets the same images obtained with the same imaging technique on two different occasions; suggested value is $0.5 \cdot w_b$ (14)	0.025
r_1	Correlation between A_z values estimated for the same subjects by the same observer with different imaging techniques; an average value of 0.47 is suggested (14)	0.47
r_2	Correlation between A_z values estimated for the same subjects by different observers with the same imaging technique; a value of 0 is suggested for $r_2 - r_3$ (14)	0
r_3	Correlation between A_z values estimated for the same subjects by different observers with different imaging techniques; a value of 0 is suggested for $r_2 - r_3$ (14)	0
r_b	Correlation between A_z values estimated for the same subjects by the group of observers with different imaging techniques; a value of 0.8 is suggested (14)	0.8
Calculated or assigned variables		
λ	Noncentrality parameter of a noncentral F distribution with 1 degree of freedom associated with the numerator and $J - 1$ degrees of freedom associated with the denominator; the values are given in Table A2	18.11
σ_b^2	Interobserver variability expressed as a variance; $\sigma_b^2 = w_b \cdot m$, where m is a multiplier that converts range into variance and is obtained from Table A3 for $n = J$	$0.05 \cdot 0.486$
σ_w^2	Intraobserver variability expressed as a variance; $\sigma_w^2 = w_w \cdot m$, where m is a multiplier that converts range into variance and is obtained from Table A3 for $n = 2$	$0.025 \cdot 0.886$
σ_c^2	Variance between subjects, calculated from Equation (A1)	0.003540
$\Phi^{-1}()$	Inverse cumulative normal distribution function, equivalent to NORMSINV() function in Microsoft Excel (Microsoft, Redmond, Wash)	
N_{pos}	Number of positive cases in the study	38
N	Total number of cases (positive + negative) in the study—that is, the sample size of cases	76

* See Appendix for description of example values.

of most statistical analyses—even simple ones like comparing means with a *t* test. Furthermore, estimates of variance are also required in sample size and power analysis for simple analyses like the *t* test. The reason for the large number of assumptions in simulations has more to do with the complexity of the data set being simulated than the method of simulation itself.

In addition to the factors usually mentioned as affecting sample size, correlation among observations within groups due to nonindependent sampling can also increase sample size and decrease statistical power. Therefore, when planning the sample size, one should take care to account for potential correlation in the data set.

APPENDIX

Although tables for the determination of sample size in ROC studies are available (14), a practical presentation of the equations underlying these tables may be helpful for situations not addressed by the published tables. The equations necessary for calculating sample size in an ROC study that involves a number of readers interpreting images obtained in the same group of subjects who have all undergone imaging with two techniques are as follows (13,14):

$$\sigma_c^2 = \frac{\dfrac{J\Delta^2}{2\lambda} - \sigma_b^2(1 - r_b) - \dfrac{\sigma_w^2}{K}}{(1 - r_1) + (J - 1)(r_2 - r_3)}, \qquad \text{(A1)}$$

$$A = \Phi^{-1}(\theta) \cdot 1.414, \qquad \text{(A2)}$$

$$N_{\text{pos}} = \frac{0.0099 e^{-A^2/2}\left[(5A^2 + 8) + \dfrac{A^2 + 8}{R}\right]}{\sigma_c^2}, \qquad \text{(A3)}$$

and

$$N = N_{\text{pos}}(1 + R). \qquad \text{(A4)}$$

Note that Equations (A1) and (A3) have been algebraically rearranged from their published form to isolate the dependent variables for more convenient calculation. All symbols are defined in Table A1. To calculate sample size with these equations, first assign values to the variables in the first section of Table A1, then sequentially substitute the values into Equations (A1)–(A4), using the suggested values of the variables in the second section of Table A1 and values from Tables A2 and A3 where indicated.

For example, Table A1 shows all the values involved in the calculation of sample size for a study that includes four readers and is designed to examine the A_z difference between two imaging techniques. On the basis of preliminary study results, the expected average A_z (θ) of the two techniques is 0.75. The smallest meaningful difference (Δ) between the A_z values for the two techniques is set to 0.15. Each reader interprets each case once ($K = 1$), and the study involves an equal number of positive and negative cases ($R = 1$). The difference in A_z (w_b) between the most accurate and least accurate observers is estimated to be 0.05. The values for w_w, r_1, r_2, r_3, and r_b given in Table A1 are those suggested by published

TABLE A2
Noncentrality Parameter (λ) of the Noncentral *F* Distribution Corresponding to a Significance Criterion (α) of .05 and a Power of .8

Numerator Degrees of Freedom	Denominator Degrees of Freedom	Noncentrality Parameter (λ)*
1	1	266.80
1	2	31.96
1	3	18.11
1	4	14.15
1	5	12.35
1	6	11.34
1	7	10.69
1	8	10.25
1	9	9.92
1	10	9.67
1	11	9.48
1	12	9.32
1	13	9.19
1	14	9.08
1	15	8.99
1	16	8.91
1	17	8.84
1	18	8.78
1	19	8.72

Note.—The noncentral *F* distribution is a generalized form of the *F* distribution and contains a ratio of two χ^2 distributions (18). Each of the two component χ^2 distributions (in the numerator and denominator) are associated with a respective degrees of freedom parameter of the *F* distribution. The degrees of freedom signify the number of independent units of information relevant to the calculation of a statistic (19), in this case the component χ^2 distributions.

* Calculated by iteration, with Lenth (20) implementation of the noncentral *F* distribution function.

TABLE A3
Multiplier for Converting the Range of a Set of Observations into an Estimate of the Variance

No. of Observations (*n*)	Multiplier (*m*)
2	0.886
3	0.591
4	0.486
5	0.430
6	0.395
7	0.370
8	0.351
9	0.337
10	0.325
11	0.315
12	0.307
13	0.300
14	0.294
15	0.288
16	0.283
17	0.279
18	0.275
19	0.271
20	0.268

Note.—Revised and adapted, with permission, from reference 21.

reports (14). The calculated sample size (*N*) is 76.

References

1. Eng J. Sample size estimation: how many individuals should be studied? Radiology 2003; 227:309–313.
2. Eng J, Siegelman SS. Improving radiology research methods: what is being asked and who is being studied? Radiology 1997; 205:651–655.
3. Newell DJ. Type II errors and ethics (letter). BMJ 1978; 4:1789.
4. Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002; 288: 358–362.
5. Rosner B. Fundamentals of biostatistics. 5th ed. Pacific Grove, Calif: Duxbury, 2000; 308.
6. Lenth RV. Some practical guidelines for effective sample size determination. Am Stat 2001; 55:187–193.
7. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 2001; 55:19–24.
8. Thomas L. Retrospective power analysis. Conserv Biol 1997; 11:276–280.
9. Detsky AS, Sackett DL. When was a "negative" clinical trial big enough? How many patients you need depends on what you found. Arch Intern Med 1985; 145: 709–712.
10. Castelloe JM. Sample size computations and power analysis with the SAS system. Proceedings of the 25th Annual SAS Users Group International Conference, April 9–12, 2000, Indianapolis, Ind. Cary, NC: SAS Institute, 2000.
11. Feiveson AH. Power by simulation. Stata J 2002; 2:107–124.
12. Obuchowski NA. Multireader receiver operating characteristic studies: a comparison of study designs. Acad Radiol 1995; 2:709–716.
13. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York, NY: Wiley, 2002; 298–304.
14. Obuchowski NA. Sample size tables for receiver operating characteristic studies. AJR Am J Roentgenol 2000; 175:603–608.
15. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27:723–731.
16. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. Acad Radiol 1997; 4:298–303.
17. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. Acad Radiol 1998; 5:591–602.
18. Abramowitz M, Stegun IA. Handbook of mathematical functions with formulas, graphs, and mathematical tables. Washington, DC: U.S. Department of Commerce, National Bureau of Standards, 1964. Applied Mathematics Series No. 55.
19. Feinstein AR. Principles of medical statistics. Boca Raton, Fla: Chapman & Hall, 2002; 115–116.
20. Lenth RV. Java applets for power and sample size. Available at: *www.stat.uiowa.edu/~rlenth/Power/index.html*. Accessed January 5, 2004.
21. Snedecor GW, Cochran WG. Statistical methods. 8th ed. Ames, Iowa: Iowa State University Press, 1989; 469.

Kimberly E. Applegate, MD, MS
Philip E. Crewson, PhD

Statistical Literacy[1]

One should not go hunting for buried treasure, because buried treasure is found at random, and, by definition, one cannot go searching for something which is found at random.

Attributed to the Talmud;
cited by Salsburg (1)

With this issue of *Radiology*, the Statistical Concepts Series of articles reaches its conclusion. We take this opportunity to thank each of the talented authors for sharing with us their expertise and patience. It takes considerable skill to translate complex concepts into a format understandable to a wide audience. Without their efforts and considerable time commitment, this series would not have been possible. Thank you.

In the current issue of *Radiology,* the 17th article in the series, by Dr Eng (2), provides an example of how we can use advanced statistical modeling to understand and predict what may work in radiology research (2,3). While this final article is complex, its sophistication provides a window into a world we radiologists rarely visit. The articles that preceded this final article were designed to provide readers of *Radiology* with an understanding of the basic concepts of statistics, probability, and scientific methods that are used in the medical literature (4). Because the Accreditation Council for Graduate Medical Education now requires of residents a basic grasp of statistical concepts as a component of the six core competencies (5), one must have a sound understanding of the methods and results presented in today's medical literature, whether one is a resident or seasoned radiologist.

We are all consumers of information. Statistics allow us to organize and objectively evaluate empiric evidence that can ultimately lead to improved patient care.

Nearly all readers of the radiology literature know that understanding study results and determining their applicability to practice requires an understanding of statistical issues. The articles that compose the Statistical Concept Series in *Radiology* are meant to increase understanding of the statistics commonly used in radiology research.

Statistical methods revolutionized science in the 20th century (1). Statistical concepts have become an essential aspect of scientific inquiry and even of our common culture. The influence on society is evidenced by the use of statistics in the lay press and media; the use of terms such as probability and correlation in everyday language; and the willingness to collect data, accept scientific conclusions, and set public policy on the basis of averages and estimates (1).

In just over 1 century, statistics have altered our view of science. Together with the rapid evolution of computer capabilities, there are many new statistical methods on the horizon. Recent trends in medical statistics include the use of meta-analysis and clustered-data analysis (6). In addition, some statistical methods, formerly uncommon in medical research, are quickly becoming embedded in our literature. These include the bootstrap method, Gibbs sampler, generalized additive models, classification and regression trees, models for longitudinal data (general estimating equations), hierarchic models, and neural networks (6). Regardless of the sophistication of a technique, to take full advantage of their potential it is necessary to understand fundamental statistical methods. The challenge for physicians is to develop and maintain statistical "literacy," in addition to the scientific literacy of radiology and medicine.

We define a functional level of statistical literacy as that which includes an understanding of methods, the effect of statistics on research design and analysis, and a basic vocabulary of statistical terms (7).

As a profession, how can we encourage statistical literacy? First, we must educate ourselves by requiring the teaching of medical statistics in medical school and residency training; Second, we should encourage the development of consensus guidelines on the proper reporting of scientific research—for example, the CONSORT (Consolidated Standards of Reporting Trials) statement (8) for reporting results of randomized controlled trials, the QUOROM (Quality of Reporting of Meta-analyses) statement (9) for reporting results of meta-analyses and systematic reviews, and the STARD (Standards for Reporting of Diagnostic Accuracy) statement (10) for reporting results of diagnostic accuracy studies. Some journals have published statistical guidelines for contributors to medical journals (11), while others have statistical checklists for manuscript reviewers. In January 2001, *Radiology* became the first American radiology journal to provide statistical review of all published manuscripts that contain statistical content, to the benefit of both authors and readers (12). Third, we should continue to promote the learning of critical thinking skills and research methodology at our national meetings, such as the seminars held at the 2002 Radiological Society of North America and the 2003 Association of University

Index terms:
Education
Statistical analysis

Published online
10.1148/radiol.2303031661
Radiology 2004; 230:613–614

[1] From the Department of Radiology, Riley Hospital for Children, 702 Barnhill Rd, Indianapolis, IN 46202 (K.E.A.); and Health Services Research and Development Service, Department of Veterans Affairs, Washington, DC (P.E.C.). Received October 9, 2003; revision requested October 14; revision received October 17; accepted October 29. **Address correspondence to** K.E.A. (e-mail: *kiappleg@iupui.edu*).

Radiologists annual meetings. Fourth, we must continue to promote the value of scientifically rigorous reports, relative to that of less scientific ones, through our national organizations, the program content at our scientific meetings, and the support of these concepts through written and oral announcements by our leadership.

The goal of the Statistical Concept Series was to enhance the ability of radiologists to evaluate the literature competently and critically, not to make them statisticians. When contemplating the value of such a basic understanding of statistics, consider that Bland (13) argued that "bad statistics leads to bad research and bad research is unethical." We must beware of translating bad research into bad medicine and recognize that we have an essential role in increasing the evidence base of medical practice. Such an understanding is perhaps one of the most useful things that radiologists must learn.

References

1. Salsburg D. The lady tasting tea: how statistics revolutionized science in the twentieth century. New York, NY: Freeman, 2001.
2. Eng J. Simplified estimation: beyond simple formulas. Radiology 2004; 230:606–612.
3. Proto AV. Radiology 2002—Statistical Concepts Series. Radiology 2002; 225: 317.
4. Applegate KE, Crewson PE. An introduction to biostatistics. Radiology 2002; 225: 318–322.
5. ACGME outcome project. Available at: *www.acgme.org/outcome/comp/compMin.asp*. Accessed September 1, 2003.
6. Altman DG. Statistics in medical journals: some recent trends. Stat Med 2000; 19: 3275–3289.
7. Mossman KL. Nuclear literacy. Health Phys 1990; 58:639–643.
8. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA 1996; 276:637–639.
9. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement—Quality of Reporting of Meta-analyses. Lancet 1999; 354:1896–1900.
10. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Radiology 2003; 226: 24–28.
11. Altman DG, Gore SM, Gardener MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. BMJ 1983; 286: 1489–1493.
12. Proto AV. Radiology 2001—the upcoming year. Radiology 2001; 218:1–2.
13. Bland M. An introduction to medical statistics. 2nd ed. Oxford, England: Oxford University Press, 1995.