Fundamentals of Clinical Research for Radiologists

Craig A. Beam[1]
C. Craig Blackmore[2]
Steven Karlik[3]
Caroline Reinhold[4]

Editors' Introduction to the Series

Research is the "eccentric uncle" of radiology. The specialty acknowledges "his" presence, brings "him" out at appropriate times to be viewed and admired, and, when the mood strikes, pays homage to "his" importance. However, the specialty has always treated research at arm's length, outside the greater, clinical concerns of organized radiology [1].

The preceding telling quote was uttered by Charles Putman and comes from a special article reporting the findings of the 1991 Radiology Summit Meeting [1]. This meeting, one in a series of annual events sponsored by the Intersociety Commission of the American College of Radiology (ACR), was held in Asheville, NC. For this meeting, radiology leaders from the United States and Canada were invited to discuss the issue of how to improve the research performed by radiologists. Obviously, the point of the quotation is that leaders in radiology think it is time to assign research a role greater than that of just the too often ignored and impotent relative.

The group reached the consensus that research has important intrinsic values both to the specialty and to individual radiologists and made the following recommendation [1]:

> To improve understanding of the value and methods of research, all trainees and faculty should receive basic instruction in critically reading the medical literature, experimental design, and biostatistics. Those wishing to conduct research should receive more extensive training.

The Value of Research to Radiology

Examples of the value of research to the specialty of radiology are not hard to find. The intimate synergistic relationship with technology is obvious. Isn't it equally apparent that research is the means by which radiologists maintain leadership of technical innovation and utilization?

From a more pedestrian perspective, research can be seen as a means to protect and expand "turf." As an example, consider the fact that research by radiologists in minimally invasive therapies, and development of these techniques, has allowed radiologists to assume a dominant role in this area. However, many believe that this area of interventional radiology is currently at risk of being swallowed by the surgical specialties. Active research and continued leadership in innovation and technology improvement by members of our specialty will help radiology maintain a primary role and prevent the attrition of the many areas of radiology practice.

Finally, from a loftier perspective, research is essential for practicing good medicine. We all have anecdotes about how cautious we must be in drawing conclusions from limited and subjective experience. For example, because we have diagnosed a case of pericardial tamponade from CT findings does not mean that CT is the imaging modality of choice for this condition, or that all patients at risk for pericardial tamponade should undergo CT. Good medicine requires decision making based on evidence, and research is the method by which this evidence is acquired, synthesized, and put into action. Sometimes this pattern of research is codified into practice guidelines and disseminated for the benefit of other practitioners. Greater effort in conducting the research for developing practice guidelines in radiology is needed.

Received July 14, 2000; accepted without revision July 14, 2000.

Series editors: Craig A. Beam, C. Craig Blackmore, Steven Karlik , and Caroline Reinhold.

This is the introduction to the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1]Department of Radiology, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226. Address correspondence to C. A. Beam.

[2] Department of Radiology, University of Washington, 325 Ninth Ave., Box 359728, Seattle, WA 98104.

[3]Diagnostic Radiology and Nuclear Medicine, University of Western Ontario, London Health Sciences Center, University Campus, 339 Windermere Rd., London, Ontario N6A 5A5, Canada.

[4]Department of Radiology, Montreal General Hospital, 1650 Cedar Ave., Montreal P. Q. H3G 1A4, Canada.

AJR 2001;176:323–325

0361–803X/01/1762–323

Goals and Overview

Recently, the ACR and the Canadian Association of Radiologists (CAR) formed a joint executive panel to address the need for improved training in clinical research methods. The outgrowth of the work of this panel was the commitment of these two organizations and the *American Journal of Roentgenology* to publish a series of articles and to develop interactive software to meet this training need. It was decided that the goal was to establish a program that will allow the progressive education of trainees and junior faculty interested in clinical research so that they can proceed from a level of nearly total ignorance to one of methodologic sophistication capable of critically understanding the literature, intelligently applying the results of research to clinical practice, communicating with methodology experts, and directing independent research.

To meet this goal, the coeditors of the series designed 22 articles with associated software that form modules of self-instruction. Each journal article and its associated software are intended to be complementary, not repetitive, learning experiences. The software, which will be available on the ACR Web site, will help readers better understand, evaluate, and refine their mastery of the material, as well as allow them to practice what has been learned. As outlined in the Appendix, it is our plan to offer an initial series of six modules at a basic level, eight modules at an intermediate level, and eight advanced-level modules.

The editors are sensitive to the ease with which methodology articles can become user-unfriendly when discussing statistical aspects of research. To ensure that modules are applicable to all readers, articles about statistical methods will show relevance to clinical radiology research by providing examples of the methods from the radiology literature within the past 10 years. These articles will accentuate concepts, definitions, and rules for use. Pictures and diagrams will be encouraged. Formulas will be discouraged and, if absolutely necessary, will be limited to an appendix.

The goal for the more advanced statistical articles is to give readers a basic understanding of the research methods available and to evaluate their appropriateness when used in the literature. This will allow readers to critically review the statistical methods section of an article or proposal and the resultant interpretation of results. This is a critical skill, not only for the researcher but also for the clinical radiologist, who must continuously reassess his or her daily practice on the basis of new information available in the literature. Another goal is to provide the reader with a sufficient base with which to conduct independent radiology research. However, these modules are not designed to replace formal training in epidemiology and biostatistics. It is our goal that these modules provide enough background so that radiologists know when to seek statistical expertise and to facilitate communication with experts in methodology.

Discussion

Any radiologist wanting to conduct or lead research must be knowledgeable of the research tradition that has arisen in medicine over the past 100 years and of the methodologic advances that have been made in the past 10 years. In addition, every radiologist needs to be able to critically appraise the medical literature—within and outside the specialty—to make the best use of new information for their patients. Because much of the medical literature is the reporting of research, a fundamental knowledge of research is essential to every practicing radiologist. In brief, to understand the message of research, radiologists must understand the methods of research.

Training residents, fellows, and junior faculty in facets of research and critical inquiry by radiology departments in both the United States and Canada is recognized by leaders of our specialty to be a critical need. However, exposure to the discipline of research has been sporadic in distribution and nonuniform in content. The recent evaluation of the introduction to research program for second-year residents by the Radiological Society of North America, Association of Univeristy Radiologists, and American Roentgen Ray Society indicated that such a program encourages development of research careers in those individuals who are oriented to research independently of participating in the training program [2].

Although there is considerable background information available for teaching aspects of critical inquiry, these materials are tailored to the academic disciplines from which they arise and are sometimes too esoteric for the specific needs of radiologists and radiology residents. What is needed is a thorough introduction to the topics with radiology-specific examples cast in a professor- and student-friendly manner.

Stolberg et al. [3] have recently detailed aspects of a core curriculum in the evaluative sciences for diagnostic imaging. The list of desirable areas of interest include clinical epidemiology, scientific method and study design, evaluation of diagnostic tests and screening, biostatistics and health economics, and technology assessment. One of the significant issues they identified was the mechanism to teach the evaluative sciences to radiology residents. Specifically, their discussion focused on problem-based or lecture-based alternatives, with an argument being made that some combination of the two would likely be optimal, depending on the individual program. Our ACR–CAR program will be a resource for all radiology residency programs that could be presented by local experts. The software will provide an interactive, problem-based adjunct to this presentation.

Our goal is to commission 22 modules. This is a prodigious list of concepts that most practicing radiologists in the United States and Canada have likely not had the opportunity to study. These topics remain significant by their absence in many radiology training programs today. Without an appreciation of these issues and their vital role in producing research excellence, radiology publications will continue in their time-honored and out-of-date series descriptions.

However, the materials we are proposing in this series can be considered only as expert resources. They will give specific and comprehensive information at the junior level to form a basis for the teaching of the critical inquiry to radiology residents, fellows, and junior faculty. The materials are meant to support, not replace, institutional instruction in these disciplines. With such materials easily available throughout the radiology community, it will become a far easier task to ensure exposure of radiologists and residents to these very important topics.

These efforts are not meant to dilute any of the essential aspects of the radiology training program. On the contrary, this series will provide a specific, highly concentrated, and relevant primer in critical inquiry. It is time that radiology incorporates these effective and scientific aspects into the discipline. Otherwise, our research efforts might come to be regarded as the “eccentric uncle” of medicine.

References

1. Hillman BJ, Putman CE. Fostering research by radiologists: recommendations of the 1991 summit meeting. *Radiology* **1992**;182:315–318
2. Hillman BJ, Nash KD, Witzke DB, Fajardo LL, Davis D. The RSNA-AUR-ARRS introduction to research program for 2nd year radiology residents: effect on career choice and early academic performance. *Radiology* **1998**;209:323–326
3. Stolberg HO, Norman GR, Moran LA, Gafni A. A core curriculum in the evaluative sciences for diagnostic imaging. *Can Assoc Radiol J* **1998**;49:295–306

Fundamentals of Clinical Research for Radiologists

APPENDIX: Series Modules

Basic Modules

- Introduction to clinical research for radiologists
- The research framework
- How to develop and critique a research protocol—meeting the "so what?" challenge
- Selecting a study population
- Collecting data
- Statistically engineering the study for success

Intermediate Modules

- Critical literature review
- Screening
- Exploring, presenting, and summarizing data
- Probability and samples
- Clinical evaluation of diagnostic technology
- Observational studies
- Decision analysis and simulation modeling
- Outcomes studies

Advanced Modules

- Inference on means and medians
- Estimating and comparing proportions
- Reader agreement studies
- Correlation and regression
- Multivariate statistical methods
- Receiver operating characteristic curve analysis
- Survival analysis
- Assessing the evidence: methods for combining published data

Fundamental of Clinical Research Series
(July 2000 – February 2006)

Contents (page numbers in PDF file)

Fundamentals of Clinical Research for Radiologists

Craig A. Beam[1]
C. Craig Blackmore[2]
Steven Karlik[3]
Caroline Reinhold[4]

Editors' Introduction to the Series

Research is the "eccentric uncle" of radiology. The specialty acknowledges "his" presence, brings "him" out at appropriate times to be viewed and admired, and, when the mood strikes, pays homage to "his" importance. However, the specialty has always treated research at arm's length, outside the greater, clinical concerns of organized radiology [1].

The preceding telling quote was uttered by Charles Putman and comes from a special article reporting the findings of the 1991 Radiology Summit Meeting [1]. This meeting, one in a series of annual events sponsored by the Intersociety Commission of the American College of Radiology (ACR), was held in Asheville, NC. For this meeting, radiology leaders from the United States and Canada were invited to discuss the issue of how to improve the research performed by radiologists. Obviously, the point of the quotation is that leaders in radiology think it is time to assign research a role greater than that of just the too often ignored and impotent relative.

The group reached the consensus that research has important intrinsic values both to the specialty and to individual radiologists and made the following recommendation [1]:

> To improve understanding of the value and methods of research, all trainees and faculty should receive basic instruction in critically reading the medical literature, experimental design, and biostatistics. Those wishing to conduct research should receive more extensive training.

The Value of Research to Radiology

Examples of the value of research to the specialty of radiology are not hard to find. The intimate synergistic relationship with technology is obvious. Isn't it equally apparent that research is the means by which radiologists maintain leadership of technical innovation and utilization?

From a more pedestrian perspective, research can be seen as a means to protect and expand "turf." As an example, consider the fact that research by radiologists in minimally invasive therapies, and development of these techniques, has allowed radiologists to assume a dominant role in this area. However, many believe that this area of interventional radiology is currently at risk of being swallowed by the surgical specialties. Active research and continued leadership in innovation and technology improvement by members of our specialty will help radiology maintain a primary role and prevent the attrition of the many areas of radiology practice.

Finally, from a loftier perspective, research is essential for practicing good medicine. We all have anecdotes about how cautious we must be in drawing conclusions from limited and subjective experience. For example, because we have diagnosed a case of pericardial tamponade from CT findings does not mean that CT is the imaging modality of choice for this condition, or that all patients at risk for pericardial tamponade should undergo CT. Good medicine requires decision making based on evidence, and research is the method by which this evidence is acquired, synthesized, and put into action. Sometimes this pattern of research is codified into practice guidelines and disseminated for the benefit of other practitioners. Greater effort in conducting the research for developing practice guidelines in radiology is needed.

Received July 14, 2000; accepted without revision July 14, 2000.

Series editors: Craig A. Beam, C. Craig Blackmore, Steven Karlik , and Caroline Reinhold.

This is the introduction to the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1]Department of Radiology, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226. Address correspondence to C. A. Beam.

[2] Department of Radiology, University of Washington, 325 Ninth Ave., Box 359728, Seattle, WA 98104.

[3]Diagnostic Radiology and Nuclear Medicine, University of Western Ontario, London Health Sciences Center, University Campus, 339 Windermere Rd., London, Ontario N6A 5A5, Canada.

[4]Department of Radiology, Montreal General Hospital, 1650 Cedar Ave., Montreal P. Q. H3G 1A4, Canada.

AJR 2001;176:323–325

0361–803X/01/1762–323

Goals and Overview

Recently, the ACR and the Canadian Association of Radiologists (CAR) formed a joint executive panel to address the need for improved training in clinical research methods. The outgrowth of the work of this panel was the commitment of these two organizations and the *American Journal of Roentgenology* to publish a series of articles and to develop interactive software to meet this training need. It was decided that the goal was to establish a program that will allow the progressive education of trainees and junior faculty interested in clinical research so that they can proceed from a level of nearly total ignorance to one of methodologic sophistication capable of critically understanding the literature, intelligently applying the results of research to clinical practice, communicating with methodology experts, and directing independent research.

To meet this goal, the coeditors of the series designed 22 articles with associated software that form modules of self-instruction. Each journal article and its associated software are intended to be complementary, not repetitive, learning experiences. The software, which will be available on the ACR Web site, will help readers better understand, evaluate, and refine their mastery of the material, as well as allow them to practice what has been learned. As outlined in the Appendix, it is our plan to offer an initial series of six modules at a basic level, eight modules at an intermediate level, and eight advanced-level modules.

The editors are sensitive to the ease with which methodology articles can become user-unfriendly when discussing statistical aspects of research. To ensure that modules are applicable to all readers, articles about statistical methods will show relevance to clinical radiology research by providing examples of the methods from the radiology literature within the past 10 years. These articles will accentuate concepts, definitions, and rules for use. Pictures and diagrams will be encouraged. Formulas will be discouraged and, if absolutely necessary, will be limited to an appendix.

The goal for the more advanced statistical articles is to give readers a basic understanding of the research methods available and to evaluate their appropriateness when used in the literature. This will allow readers to critically review the statistical methods section of an article or proposal and the resultant interpretation of results. This is a critical skill, not only for the researcher but also for the clinical radiologist, who must continuously reassess his or her daily practice on the basis of new information available in the literature. Another goal is to provide the reader with a sufficient base with which to conduct independent radiology research. However, these modules are not designed to replace formal training in epidemiology and biostatistics. It is our goal that these modules provide enough background so that radiologists know when to seek statistical expertise and to facilitate communication with experts in methodology.

Discussion

Any radiologist wanting to conduct or lead research must be knowledgeable of the research tradition that has arisen in medicine over the past 100 years and of the methodologic advances that have been made in the past 10 years. In addition, every radiologist needs to be able to critically appraise the medical literature—within and outside the specialty—to make the best use of new information for their patients. Because much of the medical literature is the reporting of research, a fundamental knowledge of research is essential to every practicing radiologist. In brief, to understand the message of research, radiologists must understand the methods of research.

Training residents, fellows, and junior faculty in facets of research and critical inquiry by radiology departments in both the United States and Canada is recognized by leaders of our specialty to be a critical need. However, exposure to the discipline of research has been sporadic in distribution and nonuniform in content. The recent evaluation of the introduction to research program for second-year residents by the Radiological Society of North America, Association of Univeristy Radiologists, and American Roentgen Ray Society indicated that such a program encourages development of research careers in those individuals who are oriented to research independently of participating in the training program [2].

Although there is considerable background information available for teaching aspects of critical inquiry, these materials are tailored to the academic disciplines from which they arise and are sometimes too esoteric for the specific needs of radiologists and radiology residents. What is needed is a thorough introduction to the topics with radiology-specific examples cast in a professor- and student-friendly manner.

Stolberg et al. [3] have recently detailed aspects of a core curriculum in the evaluative sciences for diagnostic imaging. The list of desirable areas of interest include clinical epidemiology, scientific method and study design, evaluation of diagnostic tests and screening, biostatistics and health economics, and technology assessment. One of the significant issues they identified was the mechanism to teach the evaluative sciences to radiology residents. Specifically, their discussion focused on problem-based or lecture-based alternatives, with an argument being made that some combination of the two would likely be optimal, depending on the individual program. Our ACR–CAR program will be a resource for all radiology residency programs that could be presented by local experts. The software will provide an interactive, problem-based adjunct to this presentation.

Our goal is to commission 22 modules. This is a prodigious list of concepts that most practicing radiologists in the United States and Canada have likely not had the opportunity to study. These topics remain significant by their absence in many radiology training programs today. Without an appreciation of these issues and their vital role in producing research excellence, radiology publications will continue in their time-honored and out-of-date series descriptions.

However, the materials we are proposing in this series can be considered only as expert resources. They will give specific and comprehensive information at the junior level to form a basis for the teaching of the critical inquiry to radiology residents, fellows, and junior faculty. The materials are meant to support, not replace, institutional instruction in these disciplines. With such materials easily available throughout the radiology community, it will become a far easier task to ensure exposure of radiologists and residents to these very important topics.

These efforts are not meant to dilute any of the essential aspects of the radiology training program. On the contrary, this series will provide a specific, highly concentrated, and relevant primer in critical inquiry. It is time that radiology incorporates these effective and scientific aspects into the discipline. Otherwise, our research efforts might come to be regarded as the “eccentric uncle” of medicine.

References

1. Hillman BJ, Putman CE. Fostering research by radiologists: recommendations of the 1991 summit meeting. *Radiology* **1992**;182:315–318
2. Hillman BJ, Nash KD, Witzke DB, Fajardo LL, Davis D. The RSNA-AUR-ARRS introduction to research program for 2nd year radiology residents: effect on career choice and early academic performance. *Radiology* **1998**;209:323–326
3. Stolberg HO, Norman GR, Moran LA, Gafni A. A core curriculum in the evaluative sciences for diagnostic imaging. *Can Assoc Radiol J* **1998**;49:295–306

Fundamentals of Clinical Research for Radiologists

APPENDIX: Series Modules

Basic Modules

- Introduction to clinical research for radiologists
- The research framework
- How to develop and critique a research protocol—meeting the "so what?" challenge
- Selecting a study population
- Collecting data
- Statistically engineering the study for success

Intermediate Modules

- Critical literature review
- Screening
- Exploring, presenting, and summarizing data
- Probability and samples
- Clinical evaluation of diagnostic technology
- Observational studies
- Decision analysis and simulation modeling
- Outcomes studies

Advanced Modules

- Inference on means and medians
- Estimating and comparing proportions
- Reader agreement studies
- Correlation and regression
- Multivariate statistical methods
- Receiver operating characteristic curve analysis
- Survival analysis
- Assessing the evidence: methods for combining published data

C. Craig Blackmore[1]

The Challenge of Clinical Radiology Research

The development of new technology traditionally has been the lifeblood of radiology. Many of the spectacular advances in medicine over the past few decades have centered around radiology. One does not have to go far into the past to predate the development of CT, MR imaging, and sonography, technologies that now are omnipresent, critical components of medical care. Yet for all the advances in the development of imaging technology, radiology research has come under deserved criticism in its efforts to assess the effectiveness and appropriate use of such imaging technology [1–5]. Production of a technologically adequate image is a starting point, but it is only the first step in determining whether such a technology should be used in clinical care. To be useful, an imaging study must also be accurate and provide information that has the potential to change the medical care, and ultimately the health, of the patient [6, 7].

This article is the first of an ongoing series that, taken together, will form a comprehensive teaching primer on basic and advanced concepts in technology assessment and outcomes research as described in the introductory article in this month's issue of the *American Journal of Roentgenology* (*AJR*) [8]. This series of articles published in the *AJR* will form one component of the research course cosponsored by the American College of Radiology and Canadian Association of Radiologists on the fundamentals of clinical research for radiologists. Tightly linked with these articles will be Web-based interactive teaching modules. The intent of this integrated series is to be progressive, starting with basic introductory concepts and gradually adding complexity through intermediate and more advanced modules. The objective is to provide a pathway for the novice researcher to learn to critically appraise the literature and to conduct evidence-based radiology, to communicate effectively with methodology experts, and finally, to perform or direct independent, scientifically valid, and clinically useful research.

The concepts introduced in this first article will be by design simplistic. The intent of this first module is to introduce the scope of the material that is to be covered in much greater detail in the sessions to come. Many of the major concepts of rigorous technology assessment will be introduced, with detailed discussions to follow in future modules. This introduction describes the problems of research in radiology and attempts to provide the radiology investigator with an understanding of some of the potential pitfalls to be avoided.

Evidence-Based Radiology

Every day in the clinical practice of radiology, we make observations and adjust our practice accordingly. Many of the great advances in science have arisen from just such observations. The fortuitous observation that bacteria colonies did not grow around bread mold led Alexander Fleming to discover penicillin. In radiology, we constantly observe the imaging appearances of diseases and healthy states and subtly adjust our thresholds for interpretation. However, at the same time, this simple anecdote and experience is, by definition, limited to what we personally have seen and is most strongly influenced by what we have seen recently. We have all observed the phenomenon that after a patient

Received July 14, 2000; accepted without revision July 14, 2000.

C. C. Blackmore received salary support as a General Electric–Association of University Radiologists Academic Fellow.

Series editors: Craig A. Beam, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the first in a series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1] Department of Radiology, University of Washington, Harborview Medical Center, 325 Ninth Ave., Box 359728, Seattle, WA 98104. Address correspondence to C. A. Beam, Department of Radiology, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226.

AJR 2001;176:327–331

0361–803X/01/1762–327

presents with a rare and difficult-to-diagnose disease, the next group of patients that appear at all similar will be examined for that same malady. Our belief that a disease is rare is shaken by the fact that we have seen it, and have seen it recently. The same is true for the use of diagnostic technology. For example, because we have diagnosed a case of testicular seminoma from CT findings does not mean that CT is the imaging modality of choice for this condition, or that all patients at risk for testicular seminoma should undergo CT.

To supersede this practice based on anecdote, the field of evidence-based medicine has evolved and has become the standard for medical practice [9, 10]. Although less established in radiology than in other areas of medicine, this evidence-based paradigm is no less relevant for radiology [11]. The construct underlying evidence-based medicine is that one individual's experience is limited. Decisions should be based on the best evidence from the medical literature rather than one's own limited experience [9, 11]. As a corollary, as physicians we tend to cling to what we were taught in residency or fellowship, often by acknowledged experts in the field. However, the evidence-based paradigm suggests that the experts are also individuals, and we should trust their anecdotal experience only somewhat more than we trust our own. Instead, practice should be guided by rigorous scientific investigation [9, 11, 12].

The major source for the evidence on which to base practice is the medical literature. With the rapid proliferation in radiology technology has come a parallel increase in the volume of the radiology literature. There are now more than 40 radiology journals and more than 4000 articles published each year [13]. However, the published literature has its own perils and should be interpreted with a critical eye. First, case reports, even if published, are essentially anecdotes that are codified in print. Although they are often interesting, may be provocative, and can invoke questions for scientific study, they should not form the basis for practice. Second, and more insidious, are published reports that, although well intended, contain biases or flaws in the methodology that attenuate the applicability of the results into practice. A central tenet of evidence-based medicine is that the literature must be analyzed critically, and only those studies that are robust should be used as the basis for practice [11, 14]. A useful framework for evaluating the value of a literature article is promoted by Kent et al. [2], who propose a four-grade scale (Appendix). At the top level (grade A) are methodologically rigorous studies with broad generalizability, including large randomized clinical trials and prospective comparisons of diagnostic test results to an appropriate gold standard. At the bottom level of this hierarchy are grade D studies, which include multiple methodologic flaws, biases in study design, or unsubstantiated opinion [2, 15]. Most of the radiology literature relates to development of new techniques and descriptive work. Actual assessment of these new technologies and determination of any impact on patient outcome is relatively uncommon [4]. Few grade A or B studies exist. New radiology technologies have been rapidly developed and disseminated, often without adequate proof of efficacy [1, 16, 17]. Although radiologists may not have paid great attention to the shortcomings in their research efforts, these limitations may have been more apparent to the remainder of the medical community.

Early studies of MR imaging represent an illustrative example of how radiology research has come under external criticism, particularly for methodologic deficiencies. Developed in the 1970s and early 1980s, MR imaging was initially greeted with a variety of investigations and reports in the radiology literature in particular, describing the exciting potential of this new modality. However, most of this early research was merely descriptive. Those studies that attempted to assess even accuracy were limited in size and generally suffered from important design flaws [2, 16, 18, 19]. A 1988 article by Cooper et al. [16] noted that none of the initial 54 research reports on the efficacy of MR imaging met accepted contemporary standards for research design. The article concluded that "health care professionals paying for expensive innovative technology should demand better research on diagnostic efficacy." In 1994, Kent et al. [2] found that of 142 studies of MR neuroimaging published through 1993, only one provided grade A information, 28 provided grade B or C, and most (113) provided only grade D information. Kent et al. concluded that despite the fact that more than 2000 MR imaging scanners had been installed, the evidence supporting the use of MR imaging in clinical practice was weak.

The credibility of the radiology research community was shaken by these criticisms, with some nonradiologists questioning whether conflicts of interest would influence radiologists and organized radiology [17]. Similar methodologic deficiencies have also been reported for radiology economic analyses [3, 20]. Today, more sophisticated and dependable research methods have been applied to MR imaging and assessment of efficacy with this modality for a number of indications. However, most of the research literature on the use of radiology techniques remains descriptive, with little published work on the influence of radiology on patient treatment or outcome [4]. One of the reasons for these deficiencies is the lack of research training of the individual radiology investigators. Unfortunately, training in research methodology has been underemphasized in radiology residency training in the United States [21]. Many radiologists, although highly skilled clinicians, have only a rudimentary background in research methodology and lack many of the basic tools required to perform a critical review of the medical literature. The objective of this discussion is to introduce some major concepts in research design and in critical literature review. More detailed discussion will be included in subsequent modules.

Anatomy of a Research Project

It is useful to review the anatomy of a research project. This standard framework is the foundation of the scientific literature. In brief, a research question is formulated, methods are derived to answer the question, data are collected and analyzed, and conclusions are drawn. Within this framework are several key concepts that are discussed in the following text, including formulation of the research question, use of efficient study design, avoidance of error and bias, and appropriate data analysis.

The Research Question

The first step in any research endeavor is to frame an appropriate research question. This question must be important (or it is not worth our efforts), but it also must be precise [22, 23]. As an example, we can start with a common and vexing clinical problem that has been the cause of considerable interest in the radiology literature, "Which test is better in patients with possible appendicitis, CT or sonography?" This question is certainly important and clinically relevant, but as framed above it cannot be answered. The question must be defined more precisely with respect to the type of patients in whom the question is being raised, the target population, and what is actually being asked. The imaging accuracy and usefulness of sonography and CT will likely vary on the basis of a number of patient-specific variables. Are the patients we are interested in adults or children? Are they thin or fat? Are they cooperative or uncooperative? Are they men or women? Disease-specific factors may also affect the imaging. Has the patient been symptomatic for a few hours and we suspect simple unperforated appendicitis, or has the patient been symptomatic for 4 days and appears septic, leading us to suspect an abscess? These factors also might affect the performance of sonography and CT.

Finally, how we are using the findings of an imaging study might affect the determination of optimal imaging modality. Are we using imaging to confirm appendicitis en route to the operating room, or are we using imaging to look for other abnormalities that might mimic appendicitis, such as ureteral calculi, diverticulitis, or even abdominal aortic aneurysm? A better defined research question might be, "In nonpregnant women younger than 40 years with symptoms suggestive of appendicitis but no peritoneal signs, what is the preferred imaging modality to exclude the presence of an abdominal condition that might require surgical intervention?" This reformulated research question is perhaps less "sexy" than "Which test is better?" but it is also much more useful. The reformulated question is no longer an issue of comparing radiology tests. Instead, we are asking a clinical question about a specific group of patients that can potentially affect the health of those patients [22–25]. Some experienced researchers believe that formulating and framing the research question is the most challenging aspect of doing research [22].

Study Design

Having determined the question to be answered, the next issue is the research methodology itself. To produce evidence that will appropriately drive decision making, experimental design is of critical importance and will be the focus of much of this article series. The goal of study design is to achieve the most with the least (i.e., to achieve efficiency). Fortunately, we have the experience of clinical epidemiologists and biostatisticians with decades of experience from which to draw to determine the most efficient way of designing studies and the most appropriate way to productively critique research. Prospective comparisons of diagnostic test results with a well-defined reference test and randomized double-blinded clinical trials are the study designs that provide the best information to guide clinical practice [2, 26]. However, other study designs, including cohort and case-control investigations and modeling studies can also provide useful information [4, 26]. These study designs will be discussed in detail in future modules.

Error

The research design is intended to arrive at the truth for the question under study. One of the major driving factors of research design is the effort to avoid or control error. Error can be divided into two general categories: random error, and systematic error, also known as bias. Random error, as the name implies, is due to chance events that have the potential to lead to false conclusions. The field of statistics has evolved in large part to deal with the random and therefore unpredictable error that can occur in any study design. Statistics is a methodology for drawing inference about populations from data collected on samples [27]. In medicine, we generally accept events as being true (not related to random chance) if the probability of their random occurrence is less than 5%, expressed as the common statistical p value of 0.05. Of course, unlikely events do occur. Type I (also known as alpha) error occurs when we conclude that a difference exists when in fact two groups are the same. At a significance threshold of p less than 0.05, we will make such type I errors in 5% of comparisons. However, if a study involves multiple comparisons (i.e., comparing six different MR imaging pulse sequences), then the probability of a type I error also increases [28].

The opposite of type I error, known as type II error, is when we conclude that two populations are the same when in fact they are not. Unfortunately, the commonly reported p value gives no information about the potential for this type II, or beta, error. There is a common misconception that a p value greater than 0.05 indicates that two groups are the same. However, this is only true if the study sample has sufficient size to have the power to detect a difference if it is present [27]. Sufficient sample size is determined by the size of difference we are interested in detecting, usually the amount of difference that would be clinically significant, and by the desired power of the study [27, 29]. Power is the chance the study will reveal the clinically significant difference when it exists and equals one minus the type II error probability. As an example, a study might report 90% power to detect a difference of 5%.

Bias

The opposite of random error is systematic error that is introduced through inadequacy in the study design, subject selection, or analysis. Statistics are for the most part unable to compensate for systematic error. Avoidance of such systematic error, or bias, is one of the major challenges of research design. Unfortunately, many of the apparently simple research designs that are common in the radiology literature succumb to bias. As an example, one could imagine a study designed to compare CT and MR imaging for detection of liver metastases in patients with known adenocarcinoma of another organ. To identify patients for such a study, one might review all the patients who underwent both tests, and using some external gold standard, make a comparison. However, would this study design be free of bias? Likely, there would be significant bias in the selection of the subjects. For example, if at a given center CT is generally used as the initial imaging modality for the evaluation of possible liver metastases, then the patients who undergo both imaging studies would be the ones in whom the initial CT was equivocal. The comparison would not be CT versus MR imaging, but rather, CT versus MR imaging in patients in whom the CT was equivocal. Of course, the results of such a study would underestimate the accuracy of CT, because only those cases that are difficult to diagnose with CT were included. This is a simple but unfortunately common example of selection bias in recruiting patients for a study. Selection bias occurs when the subjects studied are not representative of the target population. In the previous example, the target population is all patients with known adenocarcinoma of another organ. However, the study group is only those patients with known adenocarcinoma who underwent both CT and MR imaging. To avoid this bias, subject selection should be based on clinical criteria (i.e., all subjects with a new diagnosis of adenocarcinoma) rather than availability of imaging studies [14, 22].

When using a test to screen a population, selection bias can be more subtle but equally problematic. Intuitively, one would expect that if a cohort of subjects is randomly selected to undergo a radiologic screening test, we could compare the subjects who actually undergo screening with those who elect not to undergo screening and make reasonable conclusions. However, convincing evidence from previous screening studies indicates that differences exist between subjects who elect screening and those who refuse. Subjects who elect to undergo screening may be more health conscious, or more optimistic, or there may be some other factor that is not understood [4, 30]. Thus, in a research study designed to investigate patient outcome for a new screening study, comparison of those who undergo screening with those who elect against screening could show improved outcomes in the screened group even if the test has no benefit, or is even harmful. Therefore, to investigate the effectiveness of a screening study, it is essential to compare patients who are randomized to be invited for screening to those who are randomized not to be invited. In the analysis, all subjects are included, regardless of whether they actually undergo the screening study. This is known as an intention-to-treat analysis and avoids the subtle bias I have described [4].

Other bias can develop from the way in which data are collected. All humans have preconceived notions, both conscious and unconscious. These preconceptions alter the way in which we observe our surroundings and can unintentionally affect data that we collect, which is referred to as review bias. To remove any review bias, it is necessary to ensure that the individual who collects the data is unaware of the outcome under study. For example, the individual who determines if a test is positive should not know whether the subject truly has the disease in question. Also, when comparing two tests, the results of the first test should not be known before interpretation of the second. A recent analysis of research on diagnostic tests performed by Reid et al. [1] included some radiology studies that reported that 62% of research studies did not document that appropriate steps had been taken to avoid such review bias.

Similarly, if different gold standards are used for patients with disease than for those without, then results of accuracy studies may be overestimated. Lijmer et al. [31] found that the reported accuracy of diagnostic studies was significantly greater if different verification standards were applied to patients with and without disease than if the same gold standard was applied to all. The term “verification bias” has been applied to this problem [31, 32].

Additional potential biases in diagnostic test evaluation include spectrum bias, in which only patients with overt disease are used in assessment of a diagnostic test. Not including subtle or indeterminate cases can also lead to overestimation of disease accuracy [31, 32]. Prospective data collection is generally less subject to bias than retrospective collection and is therefore preferred when designing a study. However, retrospective data collection may be preferred in a few circumstances, such as when prospective data collection would remove the ability to blind the observers and would therefore potentially introduce greater bias.

The effect of these various biases has been documented. In general, studies with bias tend to report more encouraging results than those without bias [31]. In addition, preliminary studies of a diagnostic technology, performed with small sample size and vulnerable to bias, often will be highly optimistic about the capabilities of that technology. Subsequent reports may present a more realistic appraisal [32].

Data Analysis

Research is conducted on samples. We measure outcome or accuracy on a relatively small number of subjects. Yet the intent of research is (eventually) to influence clinical care. To achieve this, the research results must be valid on subjects other than those included in the study. Statistics is the science that allows us to make inferences about populations from measurements made on samples. A vast array of tools is available to the biostatistician to enable such inference. These tools must be familiar to the research radiologist and will be discussed in future modules. In this discussion I will limit myself to introduction of the concepts of validity and reliability.

Validity can be divided into internal validity and external validity, which is also known as generalizability. Internal validity refers to the extent to which the results and conclusions of a study actually relate to true events in the sample under study. Some of the biases and study design considerations described previously relate to validity. For example, an observer who is aware of the results of the reference test might unintentionally overestimate the accuracy of the diagnostic test under study. Thus, the recorded results might not be an internally valid representation of the actual sample. The method of data analysis and the statistical tests used are also critical to the internal validity of the study, because use of inappropriate analysis can lead to false conclusions.

Similarly, the external validity of a study is dependent on both the research design and the analytic methods. The extent to which the sample selected truly reflects the target population is a strong determinate of the generalizability of a study [22]. Also, the use of appropriate statistics allows determination of what inferences can be drawn about the target population on the basis of the sample data.

A final consideration is study reliability. Reliability refers to the extent to which the study is reproducible [1, 24]. The opposite of reliability is variability. Interpretation of some diagnostic tests can be quite subjective. If different observers cannot agree on the test result on the same subject, then interobserver variability is high. Similarly, if the same observer determines the results of the same test to be different at different times, then intraobserver variability is high. If a test has low reliability, then the test cannot achieve high accuracy in general practice [1].

Conclusion

Performing methodologically rigorous scientific research is not a trivial task. The optimal research study will be directed at an important, precisely defined clinical question, with a specified target population matched by the subject selection. The most efficient study design will be used and the sample size will be sufficient to limit type II error to an acceptable level. Further, bias will be avoided, and the results will be reliable, internally valid, and generalizable to the target population and possibly beyond. Success at such demanding research endeavors is certainly within the reach of radiologists and radiology researchers. However, training—the goal of this series of articles—is necessary.

In this article, I have attempted to introduce the problem—the need for improved research methodology in radiology research. I have also begun to outline the solution through briefly introducing the concept of evidence-based radiology and discussing the basics of research methodology: posing the research question, and study design, error, bias, and data analysis. I am certain that this discussion has been too basic for some and too sophisticated for others. However, in the modules that follow, increasing depth, clarity, and detail will be added to the rough outline that has been described in this article. By the conclusion of this project, the radiology investigator will have a comprehensive resource to aid the transition from relative novice to skilled researcher.

References

1. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA* **1995**;274:645–651
2. Kent DL, Haynor DR, Longstreth WT Jr, Larson EB. The clinical efficacy of magnetic resonance imaging in neuroimaging. *Ann Intern Med* **1994**;120:856–871
3. Blackmore CC, Magid DJ. Methodologic evaluation of the radiology cost-effectiveness literature. *Radiology* **1997**;203:87–91

4. Blackmore CC, Black WB, Jarvik JG, Langlotz CP. A critical synopsis of the diagnostic and screening radiology outcomes literature. *Acad Radiol* **1999**;6[supp 1]:S8–S18
5. Hillman BJ. Outcomes research and cost-effectiveness analysis for diagnostic imaging. *Radiology* **1994**;193:307–310
6. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* **1991**;11:88–94
7. Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. (Eugene W. Caldwell lecture) *AJR* **1994**;162:1–8
8. Beam CA, Blackmore CC, Karlik S, Reinhold, C. Fundamentals of clinical research for radiologists: editors' introduction to the series. *AJR* **2001**;176:323–325
9. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine.* New York: Churchill Livingstone, **1997**:2–3
10. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA* **1992**; 268: 2420–2425
11. Wood BP. What's the evidence? *Radiology* **1999**;213:635–637
12. Eisenberg JM. Ten lessons for evidence-based technology assessment. *JAMA* **1999**;282:1865–1869
13. Index to imaging literature. *Radiology* **1999**;210 [suppl]:iv–v
14. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA* **1994**;271:389–391
15. Kent DL, Larson EB. Disease, level of impact, and quality of research methods: three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. *Invest Radiol* **1992**;27:245–254
16. Cooper LS, Chalmers TC, McCally M, Berrier J, Sacks HS. The poor quality of early evaluations of magnetic resonance imaging. *JAMA* **1988**; 259:3277–3280
17. Kent DL, Larson EB. Diagnostic technology assessments: problems and prospects. *Ann Intern Med* **1988**;108:759–761
18. Beam CA, Sostman HD, Zheng J. Status of clinical MR evaluations 1985-1988: baseline and design for future assessments. *Radiology* **1991**; 180:265–270
19. Kent DL, Larson EB. Magnetic resonance imaging of the brain and spine: is clinical efficacy established after the first decade? *Ann Intern Med* **1988**;108:402–424 [Erratum in *Ann Intern Med* **1988**;109:438]
20. Blackmore CC, Smith WJ. Economic analyses of radiological procedures: a methodological evaluation of the medical literature. *Eur J Radiol* **1998**;27:123–130
21. Hillman BJ, Putman CE. Fostering research by radiologists: recommendations of the 1991 summit meeting. *Radiology* **1992**;182:315–318
22. Eng J, Siegelman SS. Improving radiology research methods: what is being asked and who is being studied? *Radiology* **1997**;205:651–655
23. Hulley SB, Cummings SR. *Designing clinical research.* Baltimore: Williams & Wilkins, **1988**:12–18
24. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA* **1994**;271:703–707
25. Black WC. How to evaluate the radiology literature. *AJR* **1990**;154:17–22
26. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little, Brown, **1991**:51–68
27. Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, **1991**
28. Fleiss JL. *Statistical methods for rates and proportions*, 2nd ed. New York: Wiley, **1981**:121
29. Obuchowski NA. Testing for equivalence of diagnostic tests. *AJR* **1997**;168:13–17
30. Black WC, Welch HG. Screening for disease. *AJR* **1997**;168:3–11
31. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* **1999**;282:1061–1066
32. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* **1978**;299:926–930

APPENDIX: Quality of Research Methods

Grade A: Studies with broad generalizability

- No significant flaws
- Prospective comparison of a diagnostic test with a well-defined diagnosis
- Large randomized, blinded clinical trial assessing therapeutic efficacy or patient outcome

Grade B: Studies with narrower spectrum of generalizability

- Few well-described flaws with definable impact on the results
- Prospective study of diagnostic tests
- Randomized trial of therapeutic effects and patient outcomes

Grade C: Studies with limited generalizability

- Multiple flaws in research methods, small sample size, incomplete reporting
- Retrospective studies of diagnostic accuracy

Grade D: Studies with multiple flaws in research methods

- Obvious selection bias
- Opinions without substantiating data

(Modified from Kent et al. [2, 15])

Fundamentals of Clinical Research for Radiologists

Jeffrey G. Jarvik[1]

The Research Framework

In recent years, the evaluation of diagnostic technologies has become more demanding. It is no longer sufficient to show that a new diagnostic technology can better depict anatomy or function. From the perspective of either a single hospital or society as a whole, the purchase of new technology, such as an upgrade for an MR scanner, competes directly with resources that could be spent on other aspects of health care, such as childhood immunizations. A key question in an environment of scarce resources is always, “What is the most cost-effective expenditure of our dollars?” or put another way, “Where can we get the biggest bang for our buck?” The most comprehensive evaluations try to answer this question.

In 1977, Fineberg [1] described a hierarchal scheme for evaluating diagnostic tests that consisted of four levels of efficacy. Fryback and Thornbury [2] and Thornbury [3] later revised this scheme into a model consisting of six tiers of diagnostic efficacy (Table 1).

In addition to a hierarchy for what to evaluate, there is also a hierarchy for how to evaluate it. The randomized clinical trial is the “gold standard” in the realm of clinical trials, although few have actually been performed for diagnostic tests. This is in part because of the expense and difficulty conducting randomized clinical trials. Although the randomized clinical trial is the best scientific method to combat bias, other strategies exist for evaluating diagnostic tests. These strategies include case series, case-control studies, cohort studies, and modeling.

In this article, I review the hierarchal scheme for assessing the efficacy of diagnostic technologies and the various study designs that can be used to evaluate the different levels of efficacy. I end with a brief introduction to some of the issues involved in diagnostic screening.

Levels of Diagnostic Efficacy

The six-tiered model of Fryback and Thornbury [2] is based on efficacy, which has been defined as the benefit from technology applied under ideal circumstances [4]. This is in distinction to effectiveness, which refers to the use of a technology in everyday, usual circumstances. Efficacy must be shown before effectiveness, because a test that cannot perform well under ideal circumstances has no chance of succeeding under less-than-ideal conditions.

Once the decision has been made to concentrate on efficacy, the next question is on which aspect of efficacy to focus. Guyatt et al. [5] made the observation that “…we must go beyond accuracy and try to determine if our patients are better off as a result of new technologies.” However, the link between patient outcomes and a diagnostic test is frequently tenuous. One may fail to observe a beneficial effect on patient outcome because a test is truly worthless, meaning that it is not accurate. However, there are other possibilities. The information from an accurate test may be used incorrectly by the clinician. Or there may be no effective therapy. Or the patient does not comply with effective therapy. Or the patient may not have adequate access to effective therapy. The six-tiered model disaggregates the overall effect of a diagnostic test in an attempt to discern and account for these various possibilities.

Technical Efficacy

Technical efficacy refers to the ability to produce an image and is generally measured through the physical characteristics of the image (e.g., signal-to-noise ratio, resolution).

Received September 21, 2000; accepted after revision October 5, 2000.

Supported in part by grant HS-094990 from the Agency for Healthcare Research and Quality and by a Veterans Administration ERIC (Epidemiology Research and Information Center) grant.

Series editors: Craig A. Beam, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the second in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology.* The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1] Departments of Radiology, Neurosurgery and Health Services, and the Center for Cost and Outcomes Research, University of Washington, Seattle, WA. Address correspondence to J. G. Jarvik, Department of Radiology, University of Washington, Box 357115, 1959 N.E. Pacific St., Seattle, WA 98195.

AJR 2001;176:873–877

0361–803X/01/1764–873

TABLE 1 **Six-Tiered Model of Diagnostic Efficacy**

Stage of Efficacy	Definition
Technical capacity	Resolution, sharpness, reliability
Diagnostic accuracy	Sensitivity, specificity, predictive values, ROC curves
Diagnostic impact	Ability of a diagnostic test to affect the diagnostic workup
Therapeutic impact	Ability of a diagnostic test to affect therapeutic choices
Patient outcomes	Ability of a diagnostic test to increase the length or quality of life
Societal outcomes	Cost-effectiveness and cost-utility

Note.—Data adapted from [3]. ROC = receiver operating characteristic.

This phase of investigation should be exploratory to determine the possible uses for a diagnostic test. One should explore a wide range of conditions and patients. At this stage, blinded interpretations should be avoided to allow the discovery of unexpected correlations and to refine interpretations. The danger of being too stringent at this stage of evaluation is that the development of promising technologies might actually be delayed if a rigorous but inappropriately early evaluation is negative. This phase can also be thought of as the laboratory phase of investigation, at which time technical parameters are optimized for clinical use.

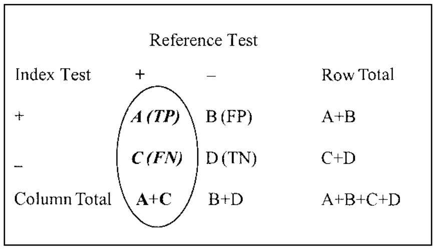
Diagnostic Accuracy Efficacy

To be useful, not only must an image be produced, it also must be interpreted. The ability to differentiate normal from abnormal in the interpretation of a test is diagnostic accuracy. Diagnostic tests are ideally compared with a gold standard to determine accuracy. The two-by-two table is the standard way to display the comparison of a new diagnostic test— usually called the index test—with that of a gold standard test, called the reference test (Table 2). The results of the reference test determine the presence or absence of disease. The parameters of sensitivity, specificity, positive predictive value, and negative predictive value can all be derived from a two-by-two table. The cells of the two-by-two table define four possible test results: true-positives, false-positives, false-negatives, and true-negatives. A case is a true-positive (TP) result when the diagnostic test is positive and the subject has the disease. Similarly, a true-negative (TN) result is when the diagnostic test is negative and the subject does not have the disease. False-positive (FP) results occur when a patient without the disease has positive test findings, and false-negative (FN) results occur when a patient with the disease has negative test findings.

The sensitivity of a diagnostic test is defined as the number of true-positive cases divided by all cases with the disease (TP / TP + FN) (Fig. 1). Specificity is the number of true-negative cases divided by all cases without the disease (TN / TN + FP) (Fig. 2). Sensitivity and specificity are related to the columns of the two-by-two table and are stable characteristics of a diagnostic test. This means that they do not change with varying disease prevalence. Positive predictive value refers to the number of patients with the disease with a positive test divided by all those with a positive test (TP / TP + FP) (Fig. 3). Negative predictive value is the number of patients without the disease with a negative test divided by all those with negative findings (TN / TN + FN) (Fig. 4).

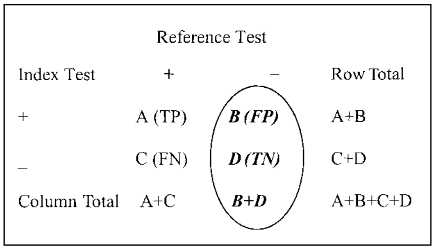
TABLE 2 **Typical Two-by-Two Table Comparing a New Test (Index Test) with a Reference Test**

Index Test	Reference Text		Row Total
	Positive	Negative	
Positive	A (True-positive)	B (False-positive)	A + B
Negative	C (False-negative)	D (True-negative)	C + D
Column total	A + C	B + D	A + B + C + D



	Reference Test		
Index Test	+	–	Row Total
+	***A (TP)***	B (FP)	A+B
_	***C (FN)***	D (TN)	C+D
Column Total	**A+C**	B+D	A+B+C+D

Fig. 1.—Diagram shows that test sensitivity focuses on first column of two-by-two table. Sensitivity equals A / (A + C), or number of patients with true-positive (TP) findings divided by all patients with positive reference test findings. + = positive test result, – = negative test result, FP = false-positive, FN = false-negative, TN = true-negative.

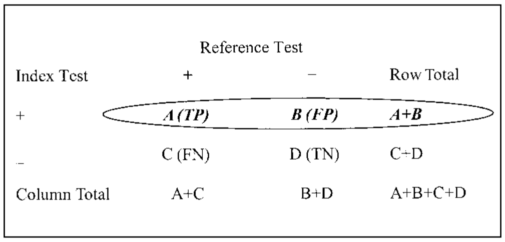


	Reference Test		
Index Test	+	–	Row Total
+	A (TP)	***B (FP)***	A+B
_	C (FN)	***D (TN)***	C+D
Column Total	A+C	***B+D***	A+B+C+D

Fig. 2.—Diagram shows how test specificity focuses on second column of two-by-two table. Test specificity equals D / (B + D), or number of patients with true-negative (TN) findings divided by all patients with negative findings on reference test. + = positive test result, – = negative test result, TP = true-positive, FP = false-positive, FN = false-negative.

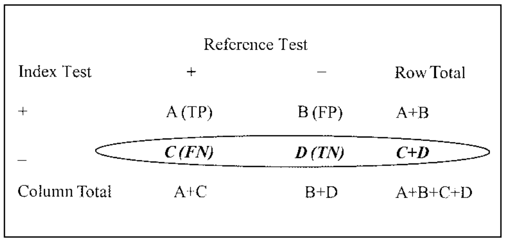
Predictive values are in one respect more clinically relevant than sensitivity and specificity because they answer the question, “If a test is positive or negative, what is the likelihood of a patient having the disease?” In contrast, sensitivity and specificity address the question, “Given that the patient does or doesn’t have the disease, what is the probability that the test will be positive or negative?” One important characteristic of predictive values is that, unlike sensitivity and specificity, they vary with disease prevalence. Tables 3 and 4 illustrate this point. Table 3 is a two-by-two table for a diagnostic test with 90% sensitivity and specificity that is applied to a population with a high (50%) prevalence of disease. In this setting, the predictive values are also quite high (90%). However, take the same diagnostic test and apply it to a population with a much lower disease prevalence (1%), and the positive predictive value decreases precipitously.

The two-by-two table assumes that a test result is dichotomous (either positive or negative). However, there are frequently many cut points to define a positive or negative test. This situation can be summarized using a receiver operator characteristic (ROC) curve. The ROC curve is a plot of sensitivity versus 1–specificity for a family of cut points that define positive and negative for a test. For example, a degenerated disk loses signal on T2-weighted MR images. One can create a scale of 1–5 to describe this signal loss, with 1 being no signal loss and 5 being complete signal loss. Now assume that we have a direct line to a divine, omniscient being who



	Reference Test		
Index Test	+	−	Row Total
+	*A (TP)*	*B (FP)*	*A+B*
_	C (FN)	D (TN)	C+D
Column Total	A+C	B+D	A+B+C+D

Fig. 3.—Predictive values are calculated from table rows rather than table columns. Positive predictive value equals A / (A + B), or number of true-positive (TP) findings divided by number of all patients with positive findings on index test. + = positive test result, – = negative test result, FP = false-positive, FN = false-negative, TN = true-negative.



	Reference Test		
Index Test	+	−	Row Total
+	A (TP)	B (FP)	A+B
_	*C (FN)*	*D (TN)*	*C+D*
Column Total	A+C	B+D	A+B+C+D

Fig. 4.—Negative predictive value is calculated from second row of table and equals D / (C + D), or number of true-negative (TN) findings divided by number of all patients with negative index test results. + = positive test result, – = negative test result, TP = true-positive, FP = false-positive, FN = false-negative.

TABLE 3 Disease Prevalence 50%

Index Test	Reference Text		Row Total
	Positive	Negative	
Positive	90	10	100
Negative	10	90	100
Column total	100	100	200

Note.—Diagnostic test with 90% sensitivity, specificity, and positive and negative predictive values. Prevalence of disease is a relatively high 50%.

TABLE 4 Disease Prevalence 1%

Index Test	Reference Text		Row Total
	Positive	Negative	
Positive	9	99	108
Negative	1	891	892
Column total	10	990	1000

Note.—Decreasing the disease prevalence to 1% leaves the sensitivity and specificity at 90%; however, the positive predictive value has decreased to 8% and the negative predictive value has increased to 99.9%.

tells us gold standard truth as to whether a disk is desiccated. We could then construct an ROC curve using each level of signal abnormality as a cutoff for normal versus abnormal. In the first instance, 1 represents normal and 2–5 represent abnormal. The second cutoff would be 1 or 2 are normal and 3–5 are abnormal, and so forth. An advantage of ROC curves is that diagnostic accuracy can be quantified for the complete range of cut points by calculating the area under the curve (A_z). A perfect diagnostic test would have an A_z of 1. A diagnostic test that conveyed no useful information would have an A_z of 0.5. Such quantification facilitates the comparison of diagnostic tests.

Diagnostic Impact Efficacy

A diagnostic test can be quite accurate and yet still not provide clinically useful information. Measures of diagnostic impact attempt to quantify the importance of a diagnostic test to diagnostic thinking. This is usually assessed using questionnaires that clinicians complete before and after receiving the results of the diagnostic test. Clinicians can be asked to rank diagnostic possibilities or even to assign probabilities to given diagnoses. If the probabilities converge on a given diagnosis, or important diagnoses are excluded, then the test has diagnostic merit. Diagnostic entropy is a concept that stems from the work of Shannon and Weaver [6] in the 1940s, based on engineering information theory. The probability for a given diagnosis is compared with the spread of probabilities over all diagnoses. Diagnostic entropy increases as the probabilities become more evenly spread across the diagnoses. Entropy decreases as probabilities concentrate around a single or a few possibilities. The problem with assessing diagnostic entropy, as well as other schemes to quantify diagnostic impact, is that it requires clinicians to make reliable and valid estimates of disease probabilities, something in which few physicians have training.

Therapeutic Impact Efficacy

Just as diagnostic impact assesses the ability of a diagnostic test to affect a diagnosis, therapeutic impact assesses the degree to which a diagnostic test influences subsequent therapeutic choices. This is also generally measured with questionnaires to physicians; but with appropriate study design, subsequent therapies can be measured, and differences in therapies can be attributed to diagnostic tests.

Fineberg [1] examined the impact that CT of the head had on diagnostic and therapeutic plans. All physicians requesting a head CT were asked to list the probabilities of the diagnoses being considered. They were also asked, if no CT were available, what diagnostic tests they would definitely and probably require and what their treatment plan would be. Medical records were then reviewed at discharge to determine which diagnostic tests were actually performed and what therapies were instituted. Fineberg found that between 41% and 73% fewer diagnostic tests were performed than were projected by the physician before CT. The therapeutic plan changed in 19% of patients. This study was one of the first published examples measuring the diagnostic and therapeutic impact of a radiologic intervention, and it helped to define the paradigm later adopted by Fryback and Thornbury [2].

Patient Outcome Efficacy

Measures of patient outcome have traditionally been limited to mortality and morbidity. However, in recent years researchers have focused more attention on health-related quality of life, which refers to the patient's appraisal of and satisfaction with his current level of functioning as compared with what the patient perceives to be possible or ideal [7]. A physician's estimate of the success or failure of an intervention is no longer sufficient. The patient's perspective as well has become important in determining efficacy. This is seen in the study by Dixon et al. [8], in which the researchers compared quality-adjusted life years (QALYs), as well as diagnostic and therapeutic impact, before and after brain and spine MR imaging. A QALY indicates a patient's willingness to trade-off length of life for quality of life. There are a variety of methods to quality-adjust life years, including the standard gamble, time trade-off, and rating scales [9]. These methods will be described in detail in future articles. Dixon et al. [8] used a questionnaire (the QALY toolkit [10]) to estimate the adjusted quality of life for different health states. The key point is that quality adjustment is from the patient's and not the physician's perspective. Although Dixon et al. found important effects on the clinicians' diagnostic confidence and therapeutic plans, there was no change in the patients' quality of life.

Societal Efficacy

In the era of constrained resources, those who pay for health care demand value. This implies that a new technology not only must improve patient outcomes, but also must maximize the health that can be bought for a dollar. Cost-effectiveness analyses are now commonly incorporated into the evaluation of new technologies and in all likelihood will remain an important aspect of technology assessment. An excellent example of this sort of study was described by Colice et al. [11]. The researchers used decision analytic modeling to compare the cost-effectiveness of screening asymptomatic patients with lung cancer for brain metastases using head CT versus scanning patients only when they became symptomatic. They determined that the cost per QALY ($70,000) with the screening strategy would be substantially higher than that of many accepted medical interventions, and thus not justified given the assumptions used in their model.

Methods of Assessing Diagnostic Technologies

Randomized trials focusing on patient outcomes are the only way to investigate these issues with the absolute assurance that bias is being avoided, and such trials should be conducted when the stakes are high enough. However, other research tools are available that can be quite powerful in their own right, and because they are easier and cheaper, they should be the study design of choice for certain situations. In addition to the randomized controlled trial, we will consider three other study designs: the case-control study, the cross-sectional study, and the cohort study.

In choosing a study design, the first decision for researchers is whether they have a question that should be answered with a descriptive or an analytic study. Descriptive studies, which can also be regarded as hypothesis generating, include case reports, case series, and cross-sectional studies. They usually describe the epidemiologic characteristics of diseases, or in the case of radiology, how imaging findings relate to patient characteristics. Measuring all variables at a single time is the distinguishing characteristic of cross-sectional studies. The classic study by Jensen et al. [12] of MR imaging findings in patients without lower back pain is an example of a cross-sectional study. The researchers identified 98 subjects, performed MR imaging on them, and determined the lack of lower back pain at one time point. In fact, most imaging investigations are cross-sectional in nature. Although cross-sectional studies are relatively easy to perform, a disadvantage is that it is frequently impossible to determine if the exposure preceded the disease or the disease preceded the exposure. For example, it has been observed that individuals with spinal stenosis are more likely to have lower activity levels, but it is impossible to determine from cross-sectional data if it is the stenosis that leads to less activity or less activity that leads to spinal stenosis.

Unlike descriptive studies, analytic studies allow hypothesis testing to determine the association between an exposure (risk factor) and an outcome (disease). Analytic studies can be divided into observational and experimental. Observational studies can be further divided into case-control and cohort studies. Patients in case-control studies are selected on the basis of whether they have the disease (or outcome) of interest. The proportion of cases with the exposure of interest is then compared with controls. For example, if sciatica is the disease of interest and nerve root compression is the exposure, a case-control study would identify patients with sciatica and then a matched group of patients without sciatica.

In contrast, a cohort study chooses subjects on the basis of the exposure (or risk factor) and then examines the proportion of subjects in each exposure group with and without the outcome of interest. These studies are usually done prospectively, with the exposure identified and the subjects then followed up over time for the development of an outcome. However, cohort studies can also be retrospective. Risk factors can be identified in the past and then the cohort assembled on the basis of these past data. One can then look at the subjects' current disease status to determine if a relevant outcome has occurred. An example of a prospective cohort study in radiology is the study by Nevitt et al. [13], who assembled a cohort of subjects with and without new osteoporotic vertebral compression fractures (the risk factor) and looked at the proportion of patients in each group who developed subsequent back pain and functional limitation (the outcomes). They found that new vertebral fractures were strongly associated with increased pain and limitations in functional status.

Case-control studies are particularly useful for examining rare outcomes, because subjects are selected on the basis of their having the outcome of interest. Conversely, cohort studies are useful for rare risk factors, because subjects are chosen on the basis of their having a particular exposure.

Experimental or intervention studies are also prospective cohort studies, because participants are enrolled on the basis of risk factors. However, experimental studies differ from observational studies in that the exposure status is assigned by the investigator. We at the University of Washington are currently conducting a randomized trial comparing a rapid MR imaging with radiography as the initial imaging technique in patients with lower back pain. The exposure we are studying is the imaging study, to which patients are randomly assigned. We are measuring a variety of outcomes, but a back-pain-specific functional status measure, the modified Roland scale [14], is our primary outcome of interest. We will monitor patients for 1 year and determine if one exposure group has significantly different outcomes from the other.

Although observational studies can control for known risk factors, both at the design and the analysis stages, a researcher can never be confident that all important risk factors that influence outcome have been identified. The unique strength of a randomized trial is that, on average, all factors, known and unknown, are controlled. Deyo [15] provided the interesting example of comparing two batches of fruit and matching them on characteristics that would seem important, such as shape, source, edibility, size, and weight (Table 5). It might

TABLE 5 Why Not Find "Matching" Controls?

Characteristic	Apples	Oranges
Shape	Round	Round
Source	Tree	Tree
Edible	Yes	Yes
Size	Handled	Handled
Weight	.23 kg	.23 kg

Note.—Adapted from [15].

appear to some that the two groups were well matched, but ultimately you're still comparing apples with oranges.

Randomized trials are the most powerful study design for excluding bias, but because they are generally difficult to conduct and are quite expensive, it is neither practical nor desirable to do randomized trials for every diagnostic imaging question. An alternative study design that is potentially widely applicable is modeling. Modeling refers to the use of decision analytic techniques to model clinical situations. Frequently used for cost-effectiveness analysis, decision modeling usually refers to constructing a decision tree that incorporates, in a quantifiable manner, various aspects of clinical practice. The advantage of decision analysis is that it deals systematically with complex situations, although failure to account for all aspects of a complex situation is a potential weakness.

The first step in constructing a decision model is to identify the clinical starting point, which identifies the group of patients for whom the analysis is conducted. Second, the diagnostic and therapeutic choices that can be applied to that population are defined. Third, probabilities are assigned to the information derived from diagnostic tests and intermediate clinical states resulting from treatments. Fourth, patient outcomes are defined that form the end points for the analysis.

Screening refers to examining people who do not have signs or symptoms for the presence of disease. Black and Welch [16, 17] have highlighted three problems with screening: lead-time bias, length bias, and pseudodisease.

Lead time refers to the interval between detection of clinically occult disease by screening and the point when the disease would have manifested clinically. This lead time causes an apparent increase in survival, known as lead-time bias, in all screening programs. This increase in survival would be equal to the lead time if testing were continuous, but is one-half the lead time for single episodes of screening [18]. Adjusting for lead-time bias usually is not possible, because lead times for new tests are not known, and there is no guarantee that disease detected by screening progresses at the same rate as disease that appears clinically.

Disease that progresses more slowly will be more likely to be identified by a screening test than rapidly progressive disease simply because slower-growing cases are in the detectable preclinical stage for a longer time. Thus, screening preferentially detects disease with slower progression compared with disease that manifests clinically. Not surprisingly, this bias, termed length bias, may result in an apparent improvement in survival, when in fact the screening program has only increased the identification of slowly progressive cases relative to the clinically more important rapidly progressive ones.

Perhaps the ultimate example of length bias is when a screening test detects "disease" that would never manifest itself clinically. Some subjects may have disease that progresses so slowly that the individual would have died from other causes before the disease became clinically apparent. This effect is termed pseudodisease, and it causes an apparent improvement in survival attributable to screening.

I have reviewed a variety of research methods that can be applied to evaluating diagnostic tests. Each has relative advantages and disadvantages that must be weighed before deciding which to use. In addition, a test can be evaluated at several possible levels ranging from diagnostic accuracy to cost-effectiveness. Without a doubt, demand will be increasing for data that can show that a new technology improves patient outcomes. As Guyatt [19] has written:

> We must go beyond accuracy and try to determine if our patients are better off as a result of new technologies. Randomized trials focusing on patient outcomes are the only way to investigate these issues convincingly and definitively and should be conducted when the stakes are high enough.

Acknowledgments

I thank Peter Jucovy and Craig Blackmore for their insightful comments.

References

1. Fineberg H. Computerized cranial tomography: effect on diagnostic and therapeutic plans. *JAMA* **1977**;38:224–227
2. Fryback D, Thornbury J. The efficacy of diagnostic imaging. *Med Decis Making* **1991**;11:88–94
3. Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. (Eugene W. Caldwell lecture) *AJR* **1994**;162:1–8
4. Brook RH, Lohr KN. Efficacy, effectiveness, variations, and quality: boundary-crossing research. *Med Care* **1985**;23:710–722
5. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ* **1986**;134:587–594
6. Shannon CE, Weaver W. *The mathematical theory of communication*. Chicago: University of Illinois Press, **1949**
7. Cella DF, Tulsky DS. Measuring quality of life today: methodological aspects. *Oncology* **1990**;4:29–38
8. Dixon AK, Southern JP, Teale A, et al. Magnetic resonance imaging of the head and spine: effective for the clinician or the patient? *BMJ* **1991**;302:79–82
9. Drummond MF, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes*. Oxford, England: Oxford Medical, **1987**: 112–148
10. Gudex C, Kind P. *The QALY toolkit*. York, England: University of York, **1988**
11. Colice GL, Birkmeyer JD, Black WC, Littenberg B, Silvestri G. Cost-effectiveness of head CT in patients with lung cancer without clinical evidence of metastases. *Chest* **1995**;108:1264–1271
12. Jensen MC, Brant-Zawadzki MN, Obuchowski N, Modic MT, Malkasian D, Ross JS. Magnetic resonance imaging of the lumbar spine in people without back pain. *N Engl J Med* **1994**;331:69–73
13. Nevitt MC, Ettinger B, Black DM, et al. The association of radiographically detected vertebral fractures with back pain and function: a prospective study. *Ann Intern Med* **1998**;128:793–800
14. Roland M, Morris R. A study of the natural history of back pain. 1. Development of a reliable and sensitive measure of disability in low back pain. *Spine* **1983**;8:141–144
15. Deyo RA. Practice variations, treatment fads, rising disability: do we need a new clinical research paradigm? *Spine* **1993**;18:2153–2162
16. Black WC, Welch HG. Advances in diagnostic imaging and overestimations of disease prevalence and the benefits of therapy. *N Engl J Med* **1993**;328:1237–1243
17. Black WC, Welch HG. Screening for disease. *AJR* **1997**;168:3–11
18. Black WC, Ling A. Is earlier diagnosis really better? The misleading effects of lead time and

length biases. *AJR* **1990**;155:625–630

19. Guyatt GH. Critical evaluation of radiologic technologies. (editorial) *Can Assoc Radiol J* **1992**;43:6–7

Stephen J. Karlik[1]

How to Develop and Critique a Research Protocol

Imagine that I am working in the sonography suite. It is 2:35 A.M., and I have just spent a fruitless 45 min assessing the perfusion of a recently transplanted liver. If I do not detect any flow in the portal circulation, the patient must have an angiogram with the risk of hepatorenal toxicity or return to surgery. New contrast material is available, but expensive, and the hospital does not sanction its routine use. What are the criteria I would use to judge the effectiveness of this change in procedure to include contrast agents so that I can justify it to the hospital and for the examination of the patient? The manufacturer of the contrast agent has provided a variety of sales material that shows the apparent excellent ability of the contrast material to show perfusion at low flow rates. A recent refresher course about contrast media had no reference to portal venous assessment. However, at a specialty meeting, one of my residency classmates presented a case report in which she claimed to have had great success. I have heard about “evidence-based” medicine and realize a quick literature search may assist. Unfortunately, relevant citations in MEDLINE are virtually nonexistent.

A hypothetic example perhaps, but consider the outcome of this quandary. I could simply administer the contrast material, but do I know the limitations and actual measurable flow rates attainable with its use? What would be the outcome of negative findings? What patients would be the best subjects for this contrast material? Who would benefit the most from the injection? Is there sufficient scientific backup to identify this usage? Unfortunately, many choices in radiology rest on such slim justifications and unknowns. How many times have radiologists succumbed to a manufacturer’s glossy brochure or an impressive pilot study presented at a meeting by a colleague with the promise of the “holy grail” of imaging advances without solid statistically verified scientific support of the advantages of the latest and greatest? When faced with such a quandary, radiologists should consider all the options, evaluate the existing evidence, and possibly investigate the problem themselves. The purpose of this module is to introduce the concepts involved in turning an interesting and valuable question into a reasonable and effective research protocol. I will briefly introduce some essential concepts that will be expanded in detail in later modules in the series. At the end, I will use the preceding clinical scenario to focus my ideas and generate a summary of my research protocol.

Defining the Question

Research is a personal issue. A key feature in defining the question to be addressed is the value of the research to the discipline and practice of radiology. In some respects, the wider the applicability of the new technique, procedure, or algorithm, the greater the importance of the work to the discipline. However, there are certainly individual or location-specific problems that can only be settled by a rigorous scientific examination, no matter how limited the importance to others.

Motivation is a significant additional component of the personal nature of the research. The definition of a research question is based on knowledge, skills, and the perceived is-

Received October 20, 2000; accepted after revision December 4, 2000.

Series editors: Craig A. Beam, C. Craig Blackmore, Stephen J. Karlik, and Caroline Reinhold.

This is the third in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology.* The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1] Department of Diagnostic Radiology, London Health Sciences Center-University Campus, Rm. 2MR21, 339 Windermere Rd., London, Ontario, N6A 5A5 Canada. Address correspondence to S. J. Karlik.

AJR 2001;176:1375–1380

0361–803X/01/1766–1375

sue. An inquiring mind would probably see questions in the special areas of interest, asking "I wonder if there is a better way to do this?" Choosing the topic is a matter of interest, perceived need, and remembering the fact that research requires time, effort, and money (TEaM) to succeed.

Radiology research questions fall into four general categories: evaluation of equipment (e.g., technology assessment, as in the value of helical CT), discovery of and evaluation of techniques (e.g., platinum embolization coils or accuracy of an imaging sign), reevaluation of old techniques or procedures (e.g., the assessment of ionic and nonionic contrast agents or cost-effectiveness of an evaluative pathway), and application of radiologic techniques to investigate changes in treatment (e.g., the use of diffusion MR imaging in early stroke treatment). All topics can provide significant opportunities to contribute to the advancement of the discipline of radiology.

How can the question be evaluated and put in perspective? How is the "so what" challenge met? The questions can come from many places: an interesting patient, a new piece of equipment, a new contrast agent, or a clinical collaborator. Once the problem becomes interesting, radiologists must evaluate its value to the patient population and their discipline. The investment in TEaM places the decision directly on the potential investigators. A thorough review of all existing available literature is essential, and "Module 7" will address the issues related to an effective critical review of the literature. Obviously, existing studies should not be repeated if they are well done and give an adequate answer to the question. Unfortunately, the radiology research literature has often not met this criterion [1].

One of the good ways to approach a research inquiry is to think from the beginning about publication because peer review is a critical filter for research. Does the project warrant a paper to describe the results? Is the work trivial, predictable, or unoriginal? Sometimes the issue could be outdated or irrelevant. Does the study show true innovation? Similarly, a study with a narrow interest or directed at a highly specialized target population may be of less interest. All studies must have a clinical importance, whether directly or indirectly, with significant implications for patients. In the discipline of radiology, it is important to ask if a new technique or procedure carries additional risk factors that make a study of marginal importance a poor choice. A summary of the key considerations for assessing a research protocol includes the following: a strong personal interest and motivation; a determination of originality, relevance, and lack of triviality or predictability; wide potential interest; definite clinical importance; and risk factors addressed. In the selection of this list, other key factors beyond importance, novelty, and answerability have been emphasized [2]. A recent editorial in *Radiology*, written to offer a series of guidelines for manuscript review, addressed the elements of both substance and style [3]. It would be wise to consider the strengths and weaknesses of the protocol and advances in knowledge mentioned in this article when planning a project.

Sometimes the question just does not seem to warrant publication, yet is still important to the investigator. An example might be the usefulness of a new piece of equipment brought to the practice, such as an add-on stereotaxic unit for mammography. Does it improve diagnostic ability compared with the previous technique and equipment? This data could be valuable to practice management, perhaps without a wider range of applicability or publication. However, the same scientific skills required for publication-quality research should be used in this investigation.

Scientific Inquiry Loop

The formulation of a specific research topic involves scientific reasoning. The first goal is to express the question in a succinct way and to justify the query as a worthwhile expenditure of time, effort, and money. It is essential to evaluate thoroughly the existing evidence relevant to the question. Then the question must be formulated in a clear and succinct hypothesis. The study should then be designed with sufficient statistical power to be unequivocal. After evaluating the final results, choose the null (the statement that groups do not differ) or alternate hypothesis (the belief that the null hypothesis is unlikely). The selection of the correct one would lead directly to the formulation of a new testable hypothesis. Then the loop of science continues in these repeated small steps (Appendix).

This loop is the foundation for our research work. After a discussion of the individual background components below, I will return to the initial quandary about sonographic contrast material and use the information to structure an appropriate research protocol.

Making Generalizations

Generalizations are often used in science and in everyday life. Many day-to-day generalizations are based on statistical analyses that are performed casually and unconsciously on the basis of observations of the world. Generalizations are useful in daily life because they have predictive value. The highway home has been jammed at the end of nearly every day. If the highway is jammed at the end of every day, then it would be reasonable to predict that it will be jammed today and that avoiding it altogether would be faster. Predicting future events from past occurrences, is "statistical thinking," which can help make decisions about the future.

What is the best way to answer the question "What proportion of patients who receive contrast agents will have a serious reaction?" If "best" means most accurate, then logging every reaction for every procedure for every bottle of contrast agent manufactured would be the best way. Although this procedure would be ideal, it is obviously not practical. For most, "best" means as accurate as one can afford to be, and accuracy can be expensive in time and money (TEaM). Therefore, generalizations are usually made from incomplete information.

Why Statistics?

In a more formal sense, the primary objective of statistics is to infer the characteristics of a whole, on the basis of the characteristics observed in a part. Gaining a complete knowledge of the whole is usually impossible for practical, technical, or financial reasons. Although statistics may not reveal the absolute truth about the whole, they will allow the estimation of the truth. How close an estimate is to the truth is affected by many factors, and under certain conditions, the probability that the estimate is in error may be quantified. Statistics refers to methodologies used to interpret quantitative data with special calculated values that describe a collection of data and then to assess error in these values. Statistical methods are useful in scientific and clinical research because they include tools that can make accurate generalizations and meaningful comparisons between groups of observations [4]. Statistical methods enable the evaluation of treatment effectiveness and diagnostic test performance and assist in the development of new drugs or therapies. "Module 6" will examine these methods in detail.

The sensitivity (or more properly, the power) of statistical methods depends on the amount of data collected. Because statistical conclusions are based on incomplete information, studies with small samples can fail to determine that a large observed difference is statistically significant. Similarly, using a large sample size can also make a small difference statistically significant. After doing the statistical analysis, radiologists still must judge their results and those of others in terms of the clinical significance of the investigation. There might be highly important differences between our groups, but the sample size is too small to detect them. An example was the need to use large numbers of cases to compare the incidence of adverse effects in nonionic and ionic contrast agents because the actual incidences were small. In a paper that finds no significant difference, did the study have sufficient numbers to determine if a truly important difference existed? Conversely, studies with large samples can reveal significant results that have no substance. Thus, in a study reporting statistical significance, is the result statistical in origin and possibly not important [4, 5]? This latter scenario refers to the "so what" challenge on a completed protocol, but not on a new one.

Hypothesis Testing

Choosing the Right Hypothesis

A hypothesis is a fundamental basis for generating a successful research project. Generating a testable hypothesis from a question leads directly to a definition of the specific studies needed to prove or disprove the hypothesis [6]. The statistical tests to determine the potential differences between groups also directly follow. To formulate a test, usually some theory has been proposed as the truth, such as that MR imaging is better than CT for diagnosing spinal tumors or that an idea is proposed as true, but is unproven, such as claiming that a new barium contrast agent is superior to the old formulation.

Medical science has adopted the scientific method for determining differences between groups by testing statistical hypotheses. Usually, the question of interest is divided into two competing hypotheses, and a study must be designed to provide evidence for choosing between them. These are the null hypothesis (H0) and the alternative hypothesis (H1). Additionally, if the null hypothesis is to be disproved, studies must be designed so that it cannot be rejected unless the evidence is sufficiently strong. For example, the hypothesis that there is no difference in the adverse reactions between nonionic and ionic contrast agents (H0) is opposed to the hypothesis that there is a difference (H1).

Formulating Hypotheses for Testing

To simplify the interpretation of the results of any statistical test, what is being compared and the expected outcome, if possible, must be clearly defined. The rule to follow is to assume that no difference exists between treatments, groups, and procedures. Assume that any difference that does exist between the groups is entirely attributable to chance (sampling error, in particular) [7]. This assumption will be maintained until a statistical test can show that it is unlikely that chance alone can account for the difference. This rationality is analogous to a court of law in which someone is innocent until proven guilty. Because absolute proof is rare in the courts, guilt that is shown beyond a reasonable doubt is good enough. So it is in statistical analysis. Absolute proof that a difference between groups is not due to chance is rare, so thresholds are set beyond which one can no longer reasonably believe that the difference is due to chance alone. Conventionally, the scientific community has used a *p* value less than 0.05 as sufficiently small to call a result statistically significant.

The statement that the groups do not differ is called the null hypothesis (H0). If the null hypothesis is shown to be sufficiently unlikely, the belief to which one switches is called the alternate hypothesis (H1) [8]. The final outcome of a hypothesis test is to either reject or not reject H0. Statisticians give the null hypothesis priority over the alternative hypothesis as it relates to the statement being tested. Often the null hypothesis is set up as a straw man to be rejected in the study. However, if H0 is not rejected, the data from the experiment do not prove that the null hypothesis is true; the data only suggest that there might not be sufficient evidence against H0 in favor of H1.

A type I error occurs in a hypothesis test when a true null hypothesis is rejected (false-positive). An example would be if a study reported a difference between MR imaging and sonography for the evaluation of carotid stenosis when in fact, there was no difference. A type II error occurs when the null hypothesis is not rejected when it should be (false-negative). A type II error would occur if it were concluded that two MR imaging contrast agents produced the same enhancement when in fact, they produced different effects. A small sample size frequently leads to a type II error. Type I and type II errors are inversely related: that is, a smaller risk of one type is accompanied by a higher risk of the other. The objective is to obtain the lowest chance of a type I error, while minimizing the possibility of a type II error.

The type I error is more serious and, therefore, should be avoided. Thus, when an experiment is proposed, the hypothesis test procedure is adjusted to produce a low probability of incorrectly rejecting H0. The probability of a type I error is the "significance level" (commonly 0.05 or 5%). Therefore, a significance level of 0.05 defines the probability level that we accept to mistakenly reject the null hypothesis. The way statistical science limits a type I error to 5% is to reject the null hypothesis only if a statistic called the *p* value is less than 5%. The *p* value measures the likelihood of observing the data, or something further removed, and assuming that the null hypothesis is true. The null hypothesis is rejected when the data are a rare event (i.e., when *p* is small). The smaller the *p* value, the more it suggests that the null hypothesis is unlikely to be correct and should be rejected. How small is small? Because we consider the significance level as 5%, an event that occurs one in 20 times is rare enough to make us reject the null hypothesis. Examples of rare events are the following: being hit by lightening, one in 2,000,000; winning a state lottery, one in 14,000,000; or being killed in an automobile accident one in 5000. All these rare events are substantially less frequent than the one in 20 criteria for a rare event in scientific research.

Type II errors occur when the null hypothesis is accepted as true, although it is false. Suppose MR angiography was compared with angiography for detection of carotid stenosis. A type II error would occur if we concluded that the two imaging modalities were the same when in fact, the performance was different. A strategy to minimize the type II error is to have sufficient numbers of studies or patients. Obtaining larger study groups is a two-edged sword because the larger the numbers, the higher the risk of finding differences (or a type I error). The size of the risk of a type II error is β, and the power of the study (the probability of drawing a true-positive conclusion when the conclusion is true) is 1-β. Table 1 shows these concepts in a manner familiar to radiologists, the two-by-two diagram; the power of the study is analogous to the sensitivity of a

TABLE 1 **Labeling the Erroneous Conclusions from a Study**

Conclusion Drawn from Study	Reality	
	Test A Better Than Test B	Test A No Better Than Test B
Test A better than test B	True-positive Correct 1-β = power	False-positive Type I error Risk of error = α
Test A no better than test B	False-negtive Type II error Risk of error = β	True-negative Correct

Note.—Adapted from [7].

diagnostic test [7]. Because we have a convention that accepts an error of 5%, the standard acceptable β error is 20% (risk of finding no difference when one exists), and the power, 1-β, is an 80% chance of finding a statistically significant difference when one exists.

Primary and Secondary Hypotheses

The discussion so far has concentrated on the concept of testing one hypothesis. Scientific protocol is divided into primary and secondary hypotheses. A hypotheses can be expressed in terms of “guiding”:

CT is better than MR imaging for spinal disease.—or “testable”:

CT is superior to MR imaging for lumbar spinal stenosis in asymptomatic individuals.

A study can be designed to investigate more than one hypothesis. For example, a study comparing the effectiveness of sonography versus MR angiography for carotid stenosis could have a primary null hypothesis that

MR imaging and sonography are equivalent for the diagnosis of carotid stenosis.

Perhaps secondary hypotheses could include a comparison of enhanced sonography and enhanced MR imaging on the evaluation:

Enhanced sonography is equivalent to enhanced MR angiography for the evaluation of carotid stenosis.—or that there is equivalence only for certain degrees of stenosis:

Enhanced sonography is equivalent to enhanced MR angiography for the evaluation of carotid disease in the range of 50–80% stenosis.

Perhaps the patient’s medical condition or symptoms could also be the focus of a secondary hypothesis:

Enhanced sonography is equivalent to enhanced MR angiography for the evaluation of carotid stenosis in patients with bruits.

Each one of these new ideas potentially adds to the TEaM. Sometimes, simpler is better. Answer one hypothesis, go entirely through our scientific loop as shown in the Appendix, propose a second hypothesis on the basis of the results, and continue the scientific progression [6]. A statistician collaborator should assist in making that determination on the basis of the study in question.

Similarly, specific aims should be identifiable for each of the protocol hypotheses. For example, if we hypothesize that enhanced sonography is equivalent to enhanced MR angiography for the evaluation of carotid stenosis, then we need to understand that a specific aim also should be considered, perhaps something like the following: to perform contrast-enhanced MR angiography and sonography on 100 consecutive patients with suspected carotid stenosis using carotid angiography as a standard of reference (previously called the gold standard). Subsequent other secondary hypotheses should also have identifiable associated aims.

Defining a Protocol

Remember the opening scenario, assessing the perfusion of a recently transplanted liver. The steps to produce a summary of the research protocol are the following: identify the problem, answer the question of whether it is generalized or specific, evaluate the existing evidence, construct an appropriate hypothesis, establish one or more aims to test the hypothesis, and define a research plan that provides sufficient statistical power to answer the hypothesis.

The researcher has a valid clinical question and a specific and relevant issue in the practice of sonography. Evaluation of portal perfusion posttransplantation is a reasonable and valuable clinical diagnostic test for an important patient population.

Our Basic Query

Can enhanced sonography help detect low flow rates in vessels that are apparently below the detection threshold for conventional Doppler sonography? Why else would the manufacturers invest so much time and money in their development? However, is the clinical use scientifically proven?

Some Other Relevant Questions

What is the minimal flow level at which the contrast agent will work? Is the effectiveness of the contrast machine dependent? What are the best techniques for visualization of low flow? Does the contrast agent work for all vessels, or are there anatomic limitations? Are there specific patients who should not have this contrast material?

Assessing the Existing Evidence

A contrast agent that permits visualization and quantification of low flow velocities could potentially improve examination on sonography of the patient with a transplanted liver. Unfortunately, portal venous thrombosis is a common complication of liver transplantation, leading to high mortality rates, difficult surgeries, and more postoperative complications [9]. In the diagnostic armamentarium, contrast-enhanced studies have proved effective in the assessment of hepatic allografts with MR imaging and angiography [10]. Although MR angiography has been compared with unenhanced sonography in the examination of liver transplants [10], only preliminary studies have been performed with sonographic contrast agents to determine the blood flow in the portal circulation [11, 12]. Contrast-enhanced MR imaging has already been used with Doppler sonography in the preoperative assessment of the portal venous system [13]. Clearly, potential exists for the use of sonographic agents for the examination of the portal venous system after transplantation in the patient. Therefore, this new technique should be applied to the assessment of the transplanted hepatic allograft, especially in the patients in whom a conventional unenhanced sonogram detects low flow or fails to detect perfusion at all. Such an added discrimination could prevent the unneeded surgical procedures, such as mesoportal jump graft or splanchnic tributary, in lieu of thrombectomy [9].

Honing the Hypothesis

Hypothesis 1: Enhanced sonography is better than unenhanced sonography for the detection of low flow rates.

This statement seems reasonable; however, this hypothesis can be tested only with great difficulty because the statement is too generic. Some defining questions are the following: in what patients, tissue, or structures? What does low flow mean? These issues are addressed in hypothesis 2.

Hypothesis 2: Enhanced sonography is better than unenhanced sonography for the detection of greater than 50% thrombosis in the portal venous system.

This hypothesis is better, but questions remain. For example, what does "better" mean? Does it mean less expensive, faster, more specific, more sensitive, easier, or less risky to the patient? In the discipline of radiology, the value of a diagnostic test must rest solidly on the concepts of sensitivity and specificity (to be discussed in "Module 11") [14, 15]. A procedure is valueless if it does not show significant sensitivity and specificity. In this instance, the technique must be sensitive to flow rates currently undetected by conventional means—a valuable extension of the existing technology. This consideration leads us to hypothesis 3.

Hypothesis 3: Enhanced sonography is more sensitive than unenhanced sonography for the detection of stenotic vessels (greater than 50% stenosis) in portal venous vessels.

If the determination of sensitivity and specificity is added to the protocol, it is essential to propose some type of a standard of reference. This can be a difficult issue in radiology; a discussion of this topic will be found in "Module 9." The determination of a standard of reference for a diagnostic procedure usually involves postsurgical examination of the relevant tissues. However, other diagnostic tests with established sensitivity and specificity have also been used. In appropriate conditions, follow-up clinical diagnosis may also be appropriate. These considerations speak directly to the relevant knowledge of the investigators and their ability to choose an appropriate standard of reference and leads to hypothesis 4:

Hypothesis 4: Enhanced sonography is more sensitive than unenhanced sonography for the detection of greater than 50% stenosis in portal venous vessels, in which angiography is used as the standard of reference.

Do normal livers have stenoses? The original inquiry and postulate was concerning a transplanted liver. This problem is relevant and gives the opportunity to generate a final testable hypothesis.

Hypothesis 5: Enhanced sonography is more sensitive than unenhanced sonography for the detection of greater than 50% stenosis in liver allograft portal vessels, in which conventional angiography is used as the standard of reference.

With this hypothesis, the specific aim can be defined, incorporating a patient population with a transplanted liver, sonographic investigation with and without contrast agents, quantification of stenosis with sonography and angiography, and determination of sensitivity and specificity. As experts in the field, radiologists know the patients and appropriate radiologic measures. However, the statistical methods and sample size that will achieve the desired power must be determined. This stage is absolutely critical in the design of the study. If an investigator does not have the competence in statistical design, a statistician should be consulted to determine how the observations will be compared and how many subjects will be needed.

Aim 1.—to determine and compare the sensitivity and specificity of enhanced and unenhanced sonography for the detection of portal venous stenosis in patients with transplanted livers with angiography as the standard of reference.

Additional aims from the same study could be the following:

Aim 2.—to determine the highest degree of stenosis on sonography and contrast-enhanced sonography when flow is still visible.

Aim 3.—to evaluate the incremental cost and benefit of the addition of contrast material to the routine examination of newly transplanted livers.

Aim 4.—to determine the predictive value of the detection of stenoses below the threshold of conventional Doppler sonography to the failure of hepatic allografts.

The final step in the definition of our research protocol is to generate a research plan that incorporates the relevant experiments needed to fulfill the aims of the study. The following is an example of a research plan to fulfill the primary aims our project.

Research Plan

Consecutive patients referred to the sonographic service for routine examination of a liver allograft will have a conventional Doppler sonogram, a contrast-enhanced sonogram (with 10 mg/kg Dopplerview), and a conventional angiogram with the administration of 30 mL radiographic contrast agent. The percentage of stenosis will be determined on all three modalities, and the sensitivities and specificities for enhanced and unenhanced sonography will be determined and compared. Scatterplots of detection and degree of stenosis will be used to establish lower cutoff levels for stenosis detection with and without contrast medium administration. A cutoff level based on angiography will be established to perform a receiver operator characteristic curve analysis of the ability of contrast-enhanced sonography to reveal pathologically important lower levels of liver flow. A significance level of p less than 0.05 will be used to evaluate the differences with receiver operator characteristic curve, chi-square, regression, and t tests, if appropriate.

Protocol Summary

Background

A contrast agent that permits visualization and quantification of low flow velocities could potentially improve examination on sonography of the patient with a transplanted liver. Although contrast-enhanced MR imaging has been compared with unenhanced sonography in hepatic allografts, only preliminary studies have been performed with contrast agents to determine the blood flow in the transplanted liver. This added discrimination could significantly improve the care of the patient with a liver transplant by preventing unneeded surgical intervention.

Hypothesis

Enhanced sonography is more sensitive than unenhanced sonography for the detection of greater than 50% stenosis in liver allograft portal vessels, whereas conventional angiography is used as the standard of reference.

Specific Aims

Aim 1.—to determine and compare the sensitivity and specificity of sonography with and without contrast agents for the detection of portal venous stenosis in patients with transplanted livers with angiography as the standard of reference.

Aim 2.—to determine the largest stenosis on sonography and enhanced sonography when flow is still visible.

Aim 3.—to evaluate the incremental cost and benefit of the addition of contrast medium administration to the routine examination of newly transplanted livers.

Aim 4.—to determine the predictive value of stenoses below the threshold of conventional Doppler sonogram for the failure of hepatic allographs.

Research Plan

In consecutive patients referred to the sonography service for routine examination of a liver allograft, a conventional Doppler sonogram, a contrast-enhanced sonogram (with 10 mg/kg Dopplerview), and a conventional angiogram with 30 mL radiographic contrast agent will be obtained. Inclusion and exclusion criteria for patient participation will be defined. The degree of stenosis will be determined for all three modalities, and the sensitivities and specificities for enhanced and unenhanced sonography will be determined and compared with an angiogram as the standard of reference. Scatterplots of detection and stenosis will be used to establish lower cutoff levels for stenosis detection with and without contrast administration. A percentage stenosis cutoff level will be established to perform a receiver operator characteristic curve analysis of the ability of contrast-enhanced sonography to reveal pathologically important portal flow. The costs of the procedures will be established and compared. Patients will be followed up clinically for 6 months to determine the relationship between allograft survival and stenosis detected. A significance level of p less than 0.05 will be used to evaluate the differences with receiver operator characteristic curve, chi-square, regression, and t tests, if appropriate.

Conclusion

The generation of this protocol has addressed a number of the key issues that define the scientific approach to radiologic investigation. In this module, an important question has been raised, the relevant background information has been examined, a testable hypothesis has been honed, a series of aims have been generated, and a possible set of experimental studies to test the hypothesis has been produced. This exercise has illustrated the thinking behind the scientific method, the basis of which is statistical hypothesis testing. The purpose of this module is, therefore, to give the structural basis to take a question, evaluate its importance, and structure it in a manner suitable for testing. The other modules in this series will address various aspects of defining and understanding the ideas behind specific techniques for specific research protocols.

Acknowledgments

We thank Donal Downey and Craig Beam for their helpful comments.

References

1. Kent DL, Haynor DR, Longstreth WT Jr, Larson EB. The clinical efficacy of magnetic resonance imaging in neuroimaging. *Ann Intern Med* **1994**; 120:856–871
2. Eng J, Siegelman SS. Improving radiology research methods: what is being asked and who is being studied? *Radiology* **1997**;205:651–655
3. Proto AV. Radiology 2000: reviewing for radiology. *Radiology* **2000**;215:619–621
4. Giere RA. Justifying statistical hypothesis. In: *Understanding scientific reasoning*. Fort Worth, TX: Holt, Rinehart & Winston, **1984**: 230–272
5. Gilbert N. Scientific tests. In: *Biometrical interpretation*. Oxford, England: Oxford University Press, **1989**:69–79
6. Medina LS. Study design and analysis in neuroradiology: a practical approach. *AJNR* **1999**;20: 1584–1596
7. Sackett DL, Haynes RB, Tugwell P. Deciding on the best therapy. In: *Clinical epidemiology*. Boston: Little Brown, **1985**:162–165
8. Clarke GM. *Statistics and experimental design*. London: Edward Arnold, **1994**:171–197
9. Yerdel MA, Gunson B, Mirza D, et al. Portal vein thrombosis in adults undergoing liver transplantation: risk factors, screening, management, and outcome. *Transplantation* **2000**;69:1873–1881
10. Glockner JF, Forauer AR, Solomon H, Varma CR, Perman WH. Three-dimensional gadolinium-enhanced MR angiography of vascular complications after liver transplantation. *AJR* **2000**;174:1447–1453
11. Venz S, Gutberlet M, Eisele RM, et al. The diagnosis and imaging of the a. hepatica after orthoptic liver transplantation: a comparison of frequency-modulated and amplitude-modulated color Doppler sonography [in German]. *Rofo Fortschr Geb Rontgenstr Neuen Bildgeb Verfahr* **1998**;169:284–289
12. Leutloff UC, Scharf J, Richter GM, et al. Use of the ultrasound contrast medium levovist in after-care of liver transplant patients: improved vascular imaging in color Doppler sonography [in German]. *Radiologe* **1998**;38:399–404
13. Naik KS, Ward J, Irving HC, Robinson PJ. Comparison of dynamic contrast enhanced MRI and Doppler sonography in the pre-operative assessment of the portal venous system. *Br J Radiol* **1997**;70:43–49
14. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Is this evidence about a diagnostic test important? In: *Evidence-based medicine*. New York: Churchill Livingston, **1997**:118–128
15. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* **1994**;271:703–707

APPENDIX: Loop of Science Algorithm

1. Ask a question
2. Assess the importance: motivation, originality, innovation, significance
3. Evaluate the existing evidence
4. Generate a specific testable hypothesis
5. State specific aims
6. Design the study
7. Evaluate data with appropriate statistical methods
8. Choose null or alternative hypothesis
9. Return to 3

Fundamentals of Clinical Research for Radiologists

Philip E. Crewson[1]
Kimberly E. Applegate[2,3]

Data Collection in Radiology Research

This paper introduces the basic principles essential for a successful data collection effort. Data collection must begin with a clear research question. The researcher should then carefully identify data needs, anticipate problems with data measurement and missing data, design and pilot test a data collection system, establish quality control, and plan both data entry and statistical analyses. To be successful, all aspects of data collection must focus on the goal of obtaining substantively important data that are consistent, accurate, and unbiased.

"On being asked to talk on the principles of research, my first thought was to arise after the chairman's introduction, to say, 'Be careful', and to sit down…" by J. Cornfield [1].

Universally lamented by experienced clinical researchers as an important but often ignored aspect of medical research, good study design and data collection are critical to the success of any clinical study [2, 3]. Although most researchers are, by their very nature, excited by experimentation and analysis, few find enjoyment in the design and implementation of data collection, although these factors are critical to successful research. Too often, researchers pay little attention to how data will be collected, if the data are available or can be measured, or how much data will be incorrect or missing. Even fewer researchers carefully train the data collectors and periodically check their work.

This paper outlines seven basic elements of data collection. We discuss defining the research question, deciding on what data to collect, obtaining institutional review board (IRB) approval, planning statistical analyses, designing the data collection system establishing quality control, and organizing data entry. This article is by no means comprehensive but provides guidelines that we believe will improve clinical research in radiology.

Three general rules of data collection underlie this discussion. First, researchers should assume they will underestimate the amount of time and effort involved in data collection. Second, the more complex the data collection process is, the longer it will take to acquire and enter the data. Finally, systematic and individual data collection errors must be addressed early in the process, because it is unwise to trust human memory or a statistician's creativity to resolve errors in the data.

Define the Primary Research Question

The first step in designing data collection is formulating the research question or questions [4]. The research question should identify the study's end points, also known as the response or outcome variables (see Appendix 1 for a glossary of terms). Examples of common end points in diagnostic imaging studies are diagnostic accuracy, patient quality of life, patient satisfaction, patient comfort, safety, morbidity, impact on patient care, and costs.

The end point is the dependent variable, the variable you wish to better understand. Identifying other variables becomes an exercise in determining what factors might explain variations in the study's end point [4, 5, 6]. These factors, known as independent variables, usually include basic demographics such as age, sex, and race. Other independent variables could include comorbidity, stage of disease, signs or symptoms, laboratory test results, imaging test results, clinician experience and training, type of imaging equipment, and patient movement, to name only a few. To be worthy of inclusion in the study, independent

Received April 3, 2001; accepted after revision April 24, 2001.

Series editors: Craig A. Beam, C. Craig Blackmore, Stephen J. Karlik, and Caroline Reinhold.

This is the fourth in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1]American College of Radiology, 1891 Preston White Dr., Reston, VA 20191. Address correspondence to P. E. Crewson.

[2]Department of Radiology, Rainbow Babies & Children's Hospital, 11100 Euclid Ave., Cleveland, OH 44106-5056.

[3]Present address: Department of Radiology, Riley Hospital for Children, 702 Barnhill Dr., Indianapolis, IN 46202-5200.

AJR 2001;177:755–761

0361–803X/01/1774–755

variables should either relate directly to the research question or provide useful controls for defining the study population and sample.

Identify Data Requirements

Deciding what, how, and when to collect data may be the most difficult part of the data collection design [2]. Every principal investigator will be faced with the dilemma of either collecting too little data, thereby weakening the study's results, or attempting to collect too much data, and becoming so overwhelmed that the study is never completed or participation of institutions and personnel wanes from exhaustion. Collecting insufficient data may also significantly impact the statistical analyses. Few studies have the luxury of retrospectively obtaining data after the study has been closed. In turn, endeavoring to collect too much data can result in lack of participation by patients and institutions, excessive amounts of missing data, fatigue of support personnel, and cascading delays in patient accrual, data cleaning, and analysis [2]. This trade-off is particularly important to address in multidisciplinary research where there will be greater demands to collect extraneous data. In characterizing these trade-offs, one author suggests that the right amount of data are "as many as necessary and as few as possible" [4].

Three additional elements must also be considered when determining what data will be collected. Essential for designing data collection forms and creating data files, these elements are the unit of analysis, data precision, and the collection sequence. The involvement of a statistician at this stage of data collection design cannot be overemphasized. They can provide guidance on determining the unit of analysis, data precision, and many other research design issues essential to a defensible statistical analysis.

Unit of Analysis

Determining the unit of analysis is a basic task in designing a study, not only for methodologic reasons, but also because it affects the design of data collection forms, the storage and linking of documentation, and the design of electronic data files. The most common unit of analysis is the individual patient, but there are many other possibilities, such as the institution, the type of procedure, the images, or in the case of reader studies, even individual radiologists.

Data Precision

The degree of accuracy needed in the collected data also deserves early attention [2, 4]. There are likely to be several different ways to measure the data you collect. For example, when recording carotid stenosis is it sufficient to record stenosis to one decimal place (0.4), two decimal places (0.44), or three (0.435)? Obviously, the more precise the measure the better, but the goal of precision may need to be tempered by consideration of the cost in both time and money and the substantive importance of the measure.

Whenever possible, use well-established measurements and common terminology to reduce design time and improve comparability with other studies [6]. In addition, good research design must address reliability (consistency and reproducibility, such as the extent to which a measure obtains similar results on identical patients) and validity (how often the positive test result is correct) [2, 4, 5, 6]. Both are important for establishing the accuracy of the study outcome.

Collection Sequence

Finally, study collaborators should consider the sequence of data collection early in the study design. This will allow for thoughtful preparation of data forms, the design of an adequate data file format, and development of a suitable analysis plan. Many studies incorporate patient follow-up, often at multiple intervals. Follow-up measurements should be recorded on well-focused forms that are coded with a common linking identifier (generally the case identification number) to ensure they can be aggregated with previously collected data.

IRB Approval

Local institutional review board approval is required in most, if not all, clinical studies and deserves to be a major consideration when designing data collection (Appendix 2). Regardless of whether it is a retrospective analysis of collected data or a prospective clinical trial, no data collection should be initiated until all ethical, procedural, and legal requirements are satisfied [7, 8]. IRB requirements will vary, but you should be prepared to address patient confidentiality, potential risks to the patient, and procedures for obtaining informed consent and monitoring for adverse events. Some IRBs will give quick administrative approval for a retrospective study of medical records, whereas others require both full IRB committee review and patient informed consent for this type of study. There is increasing public concern over confidentiality of medical records and increased scrutiny of medical research by the federal government. This concern cannot be overemphasized and is exemplified by a recent New York Times report that a computer hacker accessed thousands of medical records in a cardiology research database at the University of Washington [9].

The Statistical Analysis Plan

The precision and scale of the data (i.e., nominal, ordinal, and interval) will determine or limit the statistical techniques used in the data analysis portion of the study. Collecting patient age in years is fairly precise, but collecting date of birth allows computation of age to years, weeks, and days. Similarly, collecting age by grouping (<35, 36–55, ≥56) converts the scale of data from interval (computing mean age in years) to ordinal (not useful for computing a mean). As a result, developing a clear statistical analysis plan at this early stage can be very useful [10] not only in providing focus for the data collection effort (such as specifying sample size estimation) but also pointing out weaknesses in the scale and precision of the data before data collection begins.

The statistical analysis plan is a detailed outline of what data will be analyzed and how. This plan should include clear definitions of variables and statistical end points (descriptive or inferential), a description of the required subgroup analyses, and identification of the most appropriate statistical techniques and their relationship to the research hypotheses. Although a poor research methodology, data are frequently collected without a clear understanding of how it will be analyzed or the scale of data necessary for a particular statistical technique. A useful but time-consuming tool in designing a statistical plan is to draft the tables you will use to present the results of your analysis [5]. This approach is helpful in identifying important comparisons while clarifying statistical method requirements and data needs.

Designing the Data Collection System

Most data are either collected from secondary data sources such as patient records and other administrative databases or from primary sources such as patient interviews, patient surveys, and interpretation of imaging studies by clinical personnel. Collection instruments can be as simple as handwriting data on a paper ledger or as complex as creating a complete computerized internet-accessible direct entry system. Do not assume, however, that sophisticated electronic collection systems will produce error-free data [11]. They will be susceptible to the same problems as paper forms, such as entry errors and missing information. In addition, errant programming could lead to a multitude of problems such as improperly formatted data fields, unsaved entries, and a confusing data file design that requires extensive and time-consuming manipulation.

Regardless of the complexity, factors to consider in designing the data collection system include creating data forms, avoiding systematic bias, and preparing a plan for data administration.

Data Collection Forms

The case report form is a common tool used to collect multiple sources of data (patients, physicians, records) into one document. In developing and designing this type of data form, it is wise to allow for detailed notes, regardless of the number of investigators involved in the study [12]. These notes may or may not be entered into an electronic data file, but they can become invaluable in explaining otherwise unexplainable variations in the data later in the study. Examples could be patient movement or other uncooperative activity, equipment malfunctions, previously undisclosed comorbidities, or exceptions to protocol guidelines that can occur for many reasons including human error and clinical necessity.

Form development is both an art and a science, but there are a few basic rules to follow. First, forms should be self-explanatory to the person entering the data. Second, data should not require extensive interpretation before recording. Third, the unit of measurement should be defined. Using time as an example, specify which unit of measurement is required (hours, days, weeks, months, or years). Level of precision should be evident (fractions of hours, round to nearest full day).

Also, consistent and complete responses should be required for each section of the form. Never leave a section blank. Leaving a section blank may mean the issues are not applicable (which is important to code) or the originator of the data forgot to respond. In the case of missing data, an assumption of irrelevance may be entirely wrong.

Finally, the form should be visually appealing, easy to navigate, and conducive to data entry [2, 4]. It is often helpful to have the coding conventions for data entry included directly on the form (female [0], male [1]).

Pilot Testing

Pretest the forms on individuals who are characteristically the same as those who will fill out the forms in the study; have physicians fill out physician forms, technologists fill out technologist forms, and someone who is not a physician or healthcare worker fill out the patient questionnaires. Involve everyone who will be handling the data collection, data entry, or analysis in the form design and testing process. The initial data collection form may be piloted on a small sample of potential patients to determine whether the desired data are available and whether the data form is easy to complete and enter into the database. The data form can then be revised before the full study has begun. Finally, the principal investigator should not rely on memory to recall study design issues such as units of analyses, data measurement techniques, definitions of each measurement and variable, and time sequence. All members of the research team need a copy of the methodology and a “code book” to serve as a reminder of the methods established for data collection.

Avoid Systemic Bias

Although some biases can be corrected in the analysis, some are fatal and may render the study invalid. Therefore, it is always best to design the study to avoid or minimize these biases. There are many sources of bias to consider, however; some are more closely related to the data collection effort than others. In particular, steps should be taken to maintain objectivity in the data collection system while avoiding bias in patient recruitment and minimizing the effects of “interpretation bias” and “response bias.”

Objective measurement of data reduces the likelihood of collection bias, but the degree of objectivity can vary, depending on the means of measurement. As an example, measures of body weight on a calibrated digital scale are unlikely to vary depeding on who weighs the patient. In contrast, if surveyors are asked to indicate whether a patient is underweight, normal in weight, or overweight, much will depend on individual perceptions. It is a subjective measure.

Subjective measures are particularly susceptible to prior knowledge of the treatment arm of a clinical trial [10]. In blinded studies neither the patient nor the data collector know who is in the control group or in the treatment group, therefore minimizing bias. In open studies both the patient and the data collector know who is given treatment. In the event complete blinding is not possible, a blinded clinician could be used to review the data from both groups for consistency. Although a complete discussion is beyond the scope of this paper, be aware that a clinical study may require procedures that fail to completely mirror clinical practice, such as having all available patient information before making an assessment.

If your study calls for patient randomization into multiple study arms, it is essential that the randomization is, if at all possible, either done by a third party or automation. Before obtaining patient consent, the study monitor should have no knowledge of that patient’s study arm placement.

In diagnostic imaging, comparisons often involve the same patient receiving two diagnostic tests [13]. Ensuring that the technologists and radiologists are unaware of the competing test results is essential to prevent interpretation bias. This sort of blinding will require two separate, and probably different, data report forms.

Response bias can occur during the follow-up stage of a study because of incomplete responses or patients lost to follow-up [14]. Ill patients may be more likely to complete a follow-up quality of life survey than patients who are not ill. As a result, aggregate quality of life estimates may be lower than what they would have been if all participants responded.

Plan for Data Management

Investigators must develop a plan for ensuring the confidentiality and storage of paper forms and documentation. The preparation of a coding system that provides individual identifiers (case IDs) for each patient is one way to keep data confidential and to support blinding efforts. The same case ID should be used consistently for all data forms related to a particular patient. Maintaining consistent case IDs is especially important for follow-up efforts, because these data may have to be entered separately and then merged with the primary data file at a later date.

In general, all personal identifiable information should be collected and stored separately from the case report form. Even though patient identifying information should be separate from the case report form, each set of data should be stored in a secure location with limited access. Assign responsibilities for data storage and maintenance of the master list that contains both patient information (names and addresses) and assigned case IDs. Similarly, a computer specialist must secure the confidentiality of the electronic files.

Quality Control

Data integrity is the bedrock of any clinical study [2, 10, 14]. Early and ongoing review and cleaning of data during the collection process, while being alert for systematic biases in data collection and processing, is a critical element of ensuring quality control [15]. Preventable errors should be identified and, ideally, corrected early in the study, instead of consuming expensive resources and time cleaning the database after the study has been closed.

Data Cleaning

Cleaning data requires developing a scheme for ensuring that the data are consistent and accurate. Much depends on the study design, but consider monitoring for the following: out-of-range data values; missing data; lack of variability (survey questionnaires can include reversed questions to see whether the respondent is using the scale appropriately); logic traps (check combinations of responses for inconsistency, such as a female record that lists chronic prostatitis as a comorbidity); and date checking (verify forms are completed in sequential order) [12]. An entry error in the year field is much easier to catch early on than it is after the data are entered and combined with other participants.

All members of the research team should understand the goals and design of a study so that they may flag questionable data [2]. The study design should clearly identify the target and study populations and patient selection criteria so that variability among centers and the individual investigators who enroll patients is minimized. Develop and enforce consistent rules for data review and cleaning, including specification of how to handle missing dating. These rules should be delineated before data aggregation in which the temptation to justify certain decisions in favor of a particular outcome is strongest [10]. The principal investigators should also determine whether “interim analysis” is necessary and determine prospectively when and what is analyzed and identify the decision rules for discontinuing the study [10, 16]. An interim analysis is generally done when it is important to monitor the efficacy or safety of two treatments.

Once the data are clean, the “database lock” occurs, the point at which no additional cases or data will be added to the data file. Always assume, however, that there will be data errors even after a complete quality control plan is used [4, 10]. During the statistical analysis, do not be surprised if it becomes necessary to pull original case report forms to answer questions from the analysis. Outliers can be very revealing in a statistical analysis and it is not unusual to want to verify data integrity when the results run counter to theory or prior experience.

Amoral Consequence of Dishonesty

Our discussion of bias thus far has assumed that errors in data collection are the result of unintentional practices, such as misunderstanding instructions or rationalizing postprotocol changes in study design that result in collecting and reporting inaccurate information. In contrast, dishonesty biases data collection through deliberate falsification of either the raw data or the conditions essential for maintaining the integrity of the clinical study (such as proper patient recruitment). Regardless of whether data collection errors are accidental, well-intentioned, or the result of a deliberate fabrication, the amoral consequence is bias [3]. Some will conclude that quality control is a necessary evil to prevent the errors caused by others involved in the study. For most studies, however, the danger of instituting error in data collection rests less with the dishonest than with those well-intentioned researchers who fail to recognize and take steps to mitigate their own potential for bias.

Electronic Data File

It would be hard to conceive of a clinical study that does not require statistical analysis. Regardless of whether the statistical needs are modest (counts, percents, means) or more demanding (multivariate techniques, survival analysis, complicated error estimation), most forms of statistical analysis require the creation of an electronic data file. Therefore, planning the format of the data file at the beginning of the study is essential [1]. There can be a disconnect, however, between what the clinician visualizes as data and what the statistician needs. Although most clinicians are likely to view data in its raw form as patient records, lab results, responses on case report forms, and interview sheets, statisticians view data as numbers in an array of rows and columns.

Although there are many available data file formats and complex organizational structures, such as relational databases, most statisticians prefer the traditional rectangular data file (Table 1). Analogous to the common spreadsheet, each row typically represents one case (a patient) and each column represents a variable (a data element). Ideally, most data entries are numerical codes [1, 4]. As an example, although it is possible to enter “male” or “female” for the sex variable, data entry and subsequent statistical programming are much simpler if numbers (numerical fields) are used in place of words (string fields). In Table 1, female patients are coded “0” and male patients are coded “1.” If possible, avoid open-ended entries (e.g., free text comment or description fields) because they will inevitably lead to interpretation error. Popular electronic data files include delimited text files, Excel (Microsoft, Redmond, WA), SAS (SAS Institute, Cary, NC), SPSS (SPSS, Chicago, IL), Access (Microsoft), and Epi Info (Centers for Disease Control and Prevention, Atlanta, GA).

Conclusion

Our discussion has been based primarily on experiences related to collecting data from traditional sources, such as paper case report forms and questionnaires. Much of what has been presented, however, will also be useful as technol-

TABLE 1 Data File Design and Formatting

Case ID	Treatment Group	Recruiting Site	Start Date	Sex	Age	Imaging Result	Pathology Report
1001	1	23	10212000	1	45	3	0
1002	0	15	10252000	0	32	1	1
1003	0	7	12032000	1	56	5	0
1004	1	12	01052001	1	28	3	1

Note. —The data are coded numerically so that statistical analyses can be easily performed. Female patients are coded “0” and male patients are coded “1.” A code explanation book allows all members of the research team, including the statistician, to understand each numeric code in the database. Confidentiality is maintained by removing the patient names and assigning a case identification number.

ogy shifts to more internet-based collection efforts [17] and alternative software collection systems [18]. We are confident that, regardless of advances in technology and the potential for increased automation in data collection, the basics will remain important.

Data collection requires thoughtful preparation and consistent implementation. To be successful, all aspects of data collection must be focused on the goal of obtaining substantively important data that are consistent, accurate, and unbiased. Data collection begins with a clear research question and is followed by careful attention to identifying data needs, anticipating missing or incorrect data, planning statistical analyses, designing a data collection system, establishing quality control, and planning for data entry. Considerable misspent effort can be avoided if the principal investigators, data managers, and statisticians work together early in the design of a data collection effort.

We have presented elements of a data collection checklist that should be addressed in most, if not all, clinical research. This list is not comprehensive; much will depend on the specifics of a particular study, but recognition of the seven primary issues can dramatically improve the quality of research in radiology.

References

1. Feigal D, Black D, Grady D, et al. Planning for data management and analysis. In: SB Hulley, SR Cummings, eds. *Designing clinical research: an epidemiologic approach*. Baltimore: Williams & Wilkins **1988**:159–171
2. Friedman LM, Furberg CD, DeMets DL. Data collection and quality control. In: *Fundamentals of clinical trials.* New York: Springer **1998**:156–169
3. Altman DG. Statistics and ethics in medical research: collecting and screening data. *BMJ* **1980**; 281:1399–1401
4. Crombie IK, Davies HTO. Issues in data collection. In: *Research in health care: design, conduct and interpretation of health services research.* West Sussex, England: Wiley, **1996**:199–222
5. Goldin J, Sayre JW. A guide to clinical epidemiology for radiologists. I. Study design and research methods. *Clini Radiol* **1996**;51:313–316
6. Grady KE, Wallston BS. Research in health care settings. *Sage* **1998**;14:84–100
7. Office for Human Research Protections. *Regulations.* Department of Health and Human Services Web site. Available at: http://ohrp.osophs.dhhs.gov. Accessed April 18, **2001**
8. Department of Health and Human Services Commission on Research Integrity. *Integrity and misconduct in research.* Washington: United States Printing Office, **1995**. Publication no. 1996-746–425
9. Hackers from abroad obtain data on Washington patients. *New York Times*, Dec 8, 2000. Avaiable at: www.nytimes.com. Accessed on December 8, **2000**.
10. Meyerson LJ, Wiens BL, LaVange LM, et al. Quality control of oncology clinical trials. *Hematol Oncol Clin North Am* **2000**;4:953–971
11. McManus B. A move to electronic patient records in the community: a qualitative case study of a clinical data collection system problems caused by inattention to users and human error. *Top Health Inf Manage* **2000**;20:23–37
12. Fisher LD, Van Belle G. Data collection: design of forms. In: *Biostatistics: a methodology for the health sciences.* New York: Wiley, **1993**:24–34
13. Valk PE. Clinical trials of cost effectiveness in technology evaluation. *Q J Nucl Med* **2000**;44: 197–203
14. Kane RL. Miscellaneous observations about outcomes research: practical advice. In: *Understanding health care outcomes research.* Gaithersburg, MD: Aspen, **1997**:243–255
15. Begg CB. Biases in the assessment of diagnostic tests. *Stat in Med* **1987**;6:411–423
16. Knatterud GL. Comment. *Control Clin Trials* **1996**;17:285–293
17. Wright S, Neill K. Using the World Wide Web for research data collection. *Clin Excell Nurse Pract* **1999**;3:362–365
18. eDict Systems, Inc. eDict Systems, Inc. Web site. Available at: http://www.edictation.com. Accessed April 18, **2001**

The reader's attention is directed to earlier articles in the Clinical Research series: Introduction, which appeared in the February 2001 issue; Framework, April 2001; and Protocol, June 2001.

APPENDIX 1: Glossary of Terms

Case report form A standardized form used to collect and organize data for analysis.

Dependent variable A measure not under the control of the researcher that reflects responses caused by variations in another measure (the independent variable).

Descriptive statistic A statistic that classifies and summarizes sample data.

Independent variable A measure that can take on different values that are subject to manipulation by the researcher.

Inferential statistic A statistic that uses characteristics of a random sample along with measures of sampling error to predict the true values in a larger population.

Institutional review board An independent group of reviewers responsible for determining if the appropriate clinical, legal, and ethical safeguards have been incorporated into a study.

Interpretation bias An error in data collection that occurs when knowledge of the results of one test affects the interpretation of a second test.

Interval data Objects classified by type or characteristic, with logical order and equal differences between levels of data.

Measurement scale A reflection of how well a variable or concept can be measured. Generally categorized in order of precision as nominal, ordinal, interval, and ratio data.

Nominal data Objects classified by type or characteristic.

Ordinal data Objects classified by type or characteristic with some logical order.

Patient randomization Assignment to a treatment group that is independent of the person recruiting the patient and the patient's characteristics.

Precision The degree of accuracy used in measuring a variable.

Reliability The extent to which a measure obtains similar results over repeated trials.

Research question A question that defines the purpose of the study by clearly identifying the relationship(s) the researcher intends to investigate.

Response bias Errors in data collection caused by differing patterns and completeness of data collection that are dominated by a specific subgroup within the sample.

Response variable The measure not controlled in an experiment. Commonly known as the dependent variable.

Unit of analysis The object under study, which could be patients, radiologists, images, institutions, etc.

Validity The extent to which a measure accurately represents an abstract concept such as the presence of disease.

Variable A characteristic that can form different values from one observation to another.

APPENDIX 2: Data Collection Checklist

Determine primary research question and key end point (the dependent variable)
- Determine primary and secondary research questions and key end points
- Specify target population and sample selection criteria

Identify data needed to measure end points and provide statistical controls
- Identify the unit of analysis (patients, procedures, images, etc.)
- Determine scale and precision needed for each data element
- Identify collection sequence (pre- and postintervention, follow-up, etc.)

Obtain institutional review board approval
- Informed consent form
- Patient confidentiality
- Identify potential risks
- Adverse event monitoring

Create statistical analysis plan
- Establish statistical methods used for each research hypothesis
- Create tables used for reporting results

Design data collection system
- Specify data sources (patients, physicians, records)
- Design case report forms to collect data
- Pilot test case report forms
- Review for systematic bias
- Develop a case numbering system for data entry and record management
- Establish system for securing data forms and maintaining confidentiality

Establish quality control
- Establish a data cleaning procedure and assign responsibilities
- Establish acceptable data ranges
- Create a timeline for quality control
- Require complete entries (removes doubt about reason for missing data)

Organize data entry
- Determine data format and design electronic data file
- Develop coding and data entry guidelines
- Set data checking procedures

Fundamentals of Clinical Research for Radiologists

Ella A. Kazerooni[1]

Population and Sample

The design of clinical research begins with the formulation of a research question. As radiologists, we ask many questions about the diagnostic imaging tests we perform and interpret, particularly as new tests are introduced. Can we see a disease on an imaging test at all (technical efficacy)? What are the imaging findings of that disease (description)? Can these findings be used to distinguish between the disease in question and the condition of no disease (accuracy) or distinguish between different diseases (discrimination)? Is a newly introduced imaging test as good as or better than existing tests (comparison)? Can the test be performed in a technically adequate manner in most clinical circumstances (technical reproducibility)? Will the same radiologist interpreting an imaging study today and the same study again next month come to the same conclusion (intraobserver agreement), and will a group of radiologists of varying expertise interpret the same study the same way (interobserver agreement)? What is patient preference when given the option of two or more competing tests? How cost-effective is the test? How does the test affect treatment outcome?

Substantial research questions deal with matters of vital relevance to important groups, or populations of individuals. However, important populations are generally large and, because of numerous practicalities (economy, time, and ethics), researchers often find they cannot afford to study all members of interesting populations. The time-honored scientific solution to this problem is to draw a representative subset, or sample, from the population and to base conclusions about the population on conclusions drawn from the sample. Statistical science is then used to assess and manage the uncertainties inherent in this process of scientific inference.

The goal of this article is to review the distinction made by modern scientific thought between population and sample, and to review considerations applicable to the identification and selection of population and sample in clinical radiology research.

Conventional science distinguishes three groups of individuals (Fig. 1). The goal of the series that includes this article is to bring clinical research in radiology more in line with mainstream medical research. Researchers in radiology should therefore adhere to the modern concepts of target population, study population, and sample when designing and writing about their research. Introductory statistical texts serve to codify current concepts in mainstream scientific thinking. The following excerpt, representative of many, is taken from one such widely used text [1].

> We must also carefully distinguish between the TARGET POPULATION and the STUDY POPULATION. The target population is the whole group of [individuals] to which we are interested in applying our conclusions. The study population, on the other hand, is the group of [individuals] to which we can legitimately apply our conclusions. Unfortunately the target population is not always readily accessible and we can only study that part of it that is available. If, for example, we are conducting a telephone interview…we do not have access to those individuals without a telephone.

Received April 6, 2001; accepted after revision April 24, 2001.

Series editors: Craig A. Beam, C. Craig Blackmore, Stephen J. Karlik, and Caroline Reinhold.

This is the fifth in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1]Department of Radiology, 2910 Taubman Center, University of Michigan Medical Center, 1500 E. Medical Center Dr., Ann Arbor, MI 48109-0326. Address correspondence to E. A. Kazerooni.

AJR 2001;177:993–999

0361–803X/01/1775–993

Further on in the same text, the authors identify "sample"[1]:

> There are many ways to collect information about the study population. One way is to conduct a complete CENSUS, by collecting data for every [individual] in it.… A more practical approach is to study some fraction, or SAMPLE, of the population.

Before selecting a sample, the investigator first must determine whether a need really exists for the information that will come from the investigation. The question being asked is intimately related to the selection of a sample that can provide the answer, and to the size of the sample needed to answer the question. The sample composition impacts the generalizability of the results to the study population; the composition of the study population impacts further generalization to the target population. The biases that might be introduced in the selection of the sample impact the confidence in the conclusions that can be drawn from a research study. In discussing the sample necessary to answer different questions, examples have been taken from this author's subspecialty of thoracic radiology, particularly the use of CT pulmonary angiography for the diagnosis of acute pulmonary embolism and lung cancer.

Definition of Sample

The sample is described thoroughly in terms of clinical and demographic characteristics in the methods section of a research article so that others can draw conclusions, apply the results, and compare one investigation with another. It is not the target population, but rather a group of patients or individuals who are actually studied. The target population consists of all the individuals in the world, or in the United States, with the same characteristics as the sample to which we would like to apply the conclusions of a study. Because it is unrealistic to perform research on all individuals on earth or in the United States or in one state, we settle on a subset, or a sample, with defined inclusion and exclusion criteria. However, the results drawn from the investigation of the sample are interpreted and applied directly only to the study population. For example, to evaluate the accuracy of CT and MR imaging for lung cancer staging, it is not possible to perform CT and MR imaging on all patients diagnosed with lung cancer in the United States. The Radiologic Diagnostic Oncology Group [2] reported the accuracy of CT and MR imaging in 170 patients with "known or suspected" non–small cell lung cancer who were "considered to be surgical candidates on the basis of general health and pulmonary function." The sample was the 170 patients, and the target population was all patients with known or suspected lung cancer who were surgical candidates in the United States. A third group must be defined, however: the study population. This population includes the sample and all other patients with the same characteristics as the sample who did not participate in the study, but are in the same geographic location during the same time period of the study. For example, in the Radiologic Diagnostic Oncology Group study of 170 patients, 250 patients in total met eligibility criteria. The study population includes those 80 patients who were excluded for various reasons. Some patients might have declined to be studied, others might have dropped out after enrollment. How they differ from those who agreed to participate might introduce bias, which is discussed later.

If a group of patients in clinical practice meets the same inclusion and exclusion criteria as the sample, then we apply the conclusions drawn from the sample to these patients from the study population with confidence. The more a patient differs from the sample, the more likely it is that the results from the sample do not apply to this patient.

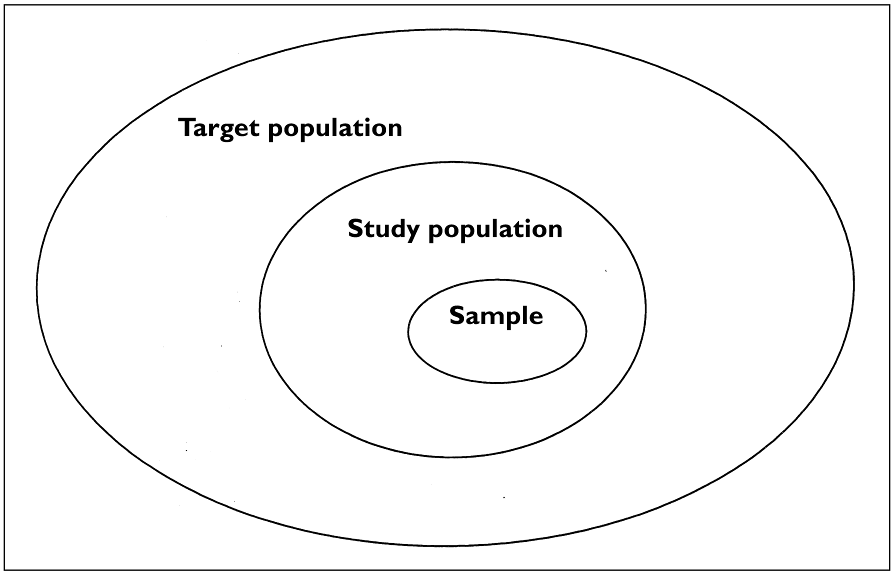


Fig. 1.—Graphic shows relationships among target population, study population, and sample. Conventional science distinguishes three groups of individuals. Target population is population of ultimate clinical interest. But, because of practicalities, entire target population often cannot be studied. Study population is subset of target population that can be studied. Samples are subsets of study populations used in clinical research because often not every member of study population can be measured.

Can a Disease Be Detected on an Imaging Test, and What Does It Look Like?

If the intended purpose of proposed research is to introduce a new concept to the literature, then a sample of one or a few might be sufficient. This approach might be useful when a new technology is applied to a disease or clinical circumstance, or when the imaging findings of a specific disease are being described. This type of research is called descriptive research, and it is used in most of the published radiology articles [3–6]. Descriptive research is the lowest on the hierarchy of studies at providing information that can be used to evaluate the efficacy of a diagnostic test in actual clinical practice [7], but for rarely occurring diseases it might be difficult to do anything more. However, these studies are a necessary first step along the way to evaluating efficacy. They are the easiest to perform, use the least amount of resources, and in the circumstance of a single case report, are usually the hardest to publish. Without knowing what a disease looks like, the next step—determining whether a test can distinguish between disease or no disease, can discriminate between diseases, and, if so, how accurately and reproducibly—cannot be done.

For example, in the early to mid 1980s, several groups of researchers reported on CT and pulmonary embolism [8–13]. Those articles were case reports and small case series that for the first time documented that pulmonary embolism could be seen on IV contrast-enhanced CT. Although this simple concept might appear obvious to someone looking at the CT technology of today, it was not apparent before that time. The purpose of these reports by several investigators was to confirm the observation and to generate a database of knowledge that could lead to the generation of more complex scientific hypotheses. The early observations did not show the technical limitations of the technique or reveal the parameters necessary to optimize the technique. They did not show the accuracy of CT compared with a known reference standard such as conventional pulmonary angiography, and they did not show the accuracy of CT compared with other diagnostic tests, such as ventilation–perfusion scintigraphy alone or in combination with lower extremity sonography. They did not show whether observers of varying expertise could agree on the diagnosis reproducibly or evaluate patient preference for one diagnostic test or another. These observations were simply the first step in a series of steps that need to occur before it can be determined if and what the role of a new technology is in medical practice.

Selection Bias and How to Select an Unbiased Population

When looking for a population of patients with a specific disease for which the findings of that disease are to be described, or to compare the accuracy of one test against another, it might seem straightforward to generate a list of all patients with the disease who have undergone the test or tests of interest over a specified period of time. However, who is chosen impacts to whom the results can be generalized. Many times in descriptive series a statement is made in the methods section that all patients with a specific disease imaged with a specific test formed the sample. Or, when comparing one test against another, such as CT versus MR imaging, all patients who underwent CT were compared against all patients who underwent MR imaging. What does this really mean? It is important that the population studied is thoroughly described, so that readers can compare the results of one study against another, particularly when results appear to be in conflict. Several biases can be introduced; the major issues of concern are sampling bias, the exclusion of patients, the use of a retrospective sample versus a prospectively collected sample, consecutive versus nonconsecutive patient enrollment, and selection based on the availability of imaging rather than the clinical presentation or clinical question.

Sampling Bias

The best sample is one that has the same characteristics as the study population to which the investigator wishes the results to be applied. The choice of a control group might introduce bias. A control group made up of normal volunteers recruited from a newspaper advertisement or a notice on a bulletin board is likely to be healthier than disease-free patients being seen in a medical clinic, which will make a diagnostic test appear more specific [14]. For example, if the intent is to investigate the diagnostic accuracy of a test, such as positron emission tomography, to distinguish between lung cancer and no lung cancer, the appropriate group to study is all patients with suspected lung cancer, not patients with lung cancer and healthy volunteers. In actual clinical practice, the diagnostic test would not be applied to normal healthy volunteers but instead to patients with, for example, a solitary nodule detected on a chest radiograph, some of whom will have lung cancer and some of whom will not.

No matter what population is studied, it is important to thoroughly describe them. It is equally important to describe the sample. Although age and sex are usually specified, other factors, such as racial mix, inner city versus rural setting, or type of medical center in which the investigation was performed, often are not. Diseases might look different in populations of different ethnic backgrounds, and therefore diagnostic tests might perform differently. Patients referred to a tertiary academic medical center might have more severe disease than patients treated for the same disease in a community hospital. This factor might make a diagnostic test appear to be more sensitive than it is in actual community practice, because more severe disease is generally easier to detect [14]. It is also important to report comorbidities. For example, the accuracy of CT pulmonary angiography for pulmonary embolism might be different in outpatients, who in general are less sick and more likely to be able to hold their breath for a CT examination, than in hospitalized patients, particularly intensive care unit patients, who are more likely to have lung disease. In this example, reporting the frequency of pleural effusions, lung abnormalities, pulmonary function test results, and the percentage of patients who are ventilator-dependent might be crucial to understanding the population studied and how the results could be applied in clinical practice.

Exclusions and Omission of Uninterpretable Results

As important as it is to describe who was studied, it is also important to describe patients who were excluded from the study or who declined to participate, because they might be different from the patients actually studied [15]. Some exclusions are random: for example, an optical disk on which a CT scan of a patient was stored is corrupted and the hard- copy images for that case are lost, or a patient died an unrelated death as a result of an airplane crash. Other exclusions are not random, and might introduce bias. For example, if patients with early stage lung cancer manifesting predominantly as a solitary pulmonary nodule declined to participate in a CT study designed to evaluate lung cancer staging, the sensitivity of CT staging might be artificially high and the population studied might be biased to patients with relatively obvious metastatic disease. On the other hand, if patients with advanced metastatic lung cancer declined to participate in the study because they felt too sick, then the sensitivity of CT staging might be artificially low because the patients with the most obvious disease were not included. For these reasons, it is important to describe the patients studied as well as the patients who were not studied, and to compare them to determine whether inherent differences exist.

Consider the Radiologic Diagnostic Oncology Group lung cancer staging study [2] in which 80 of the 250 eligible patients were excluded from the analysis. The report states that 43 of these patients did not undergo a surgical staging procedure, and “20 of these were considered to have extensive disease on the basis of imaging studies (six of these had T3 or T4 lesions).” Therefore, six (7.5%) of 80 patients excluded had T3 or T4 lesions, compared with 48 of the 170 studied, or 28% [2]. In general, the higher the T level, the more likely that metastatic lymph nodes are present and that these lymph nodes are larger in size and greater in number than for lower level T lesions, and therefore easier to identify. If the sample is skewed toward patients with more severe disease, then the sensitivity might be overestimated. On the other hand, for the other 14 of 20 excluded for extensive disease, it is not stated in the published report what the extensive disease was. It is logical to think it might have been metastatic disease or M1 disease because patients with all levels of nodal or N disease were reported. If this is correct, then 14 (17.5%) of 80 excluded patients had metastatic disease. Because it is more likely that patients with metastatic disease have larger lymph nodes of greater size than patients without metastatic disease, selecting out more obvious cases of lymph node metastases might artificially reduce the reported sensitivity for lymph node staging compared with a group of all patients with known or suspected lung cancer selected to undergo imaging. So within the same study there are reasons to think that the reported sensitivity of CT and MR imaging for staging the lymph nodes is exaggerated and underestimated. The more thoroughly the sample and the excluded patients are defined, the easier it is to know whether they are similar or dissimilar and how that might impact these reported measures of test performance.

Omitting the results of studies that are technically inadequate and therefore uninterpretable, or including in a study only patients who can cooperate sufficiently to produce a technically optimal diagnostic test can lead to an overestimate of the test's sensitivity. For example, one cause of suboptimal-quality CT pulmonary angiography for acute pulmonary embolism is respiratory motion, because many patients with suspected pulmonary embolism are short of breath. If the sample is selected using clinical and demographic characteristics, and then the examinations of suboptimal quality are excluded from the final analysis, the reported sensitivity will be higher than if these patients were included in the analysis as cases in which no pulmonary embolism was detected on these studies (i.e., as negatives).

Using another CT pulmonary angiography example, Remy-Jardin et al. [16] compared the findings in 20 patients who underwent pulmonary angiography studies using 3-mm collimation, pitch of 1.7, and 1.0 sec per rotation with findings in 20 patients who underwent CT pulmonary angiography studies using 2-mm collimation, pitch of 2, and 0.75 sec per rotation. Remy-Jardin et al. stated the purpose of their study was to "analyze the influence of collimation on identification of segmental and subsegmental pulmonary arteries." The frequency of arteries that were sufficiently well seen to be analyzable for emboli was reported for both groups, with statistically significantly more segmental and subsegmental arteries seen with the thinner collimation protocol. When the sample is scrutinized, the scans included in the study had to be "technically acceptable," with strict inspiratory apnea and good or excellent arterial contrast opacification. Patients with prior lung surgery, lung distortion, or parenchymal infiltration on CT were excluded. Thirty-five patients were evaluated for suspected pulmonary embolism, all of whom had negative findings for pulmonary embolism on CT pulmonary angiography; the other five patients (12.5%) were not scanned because of suspected pulmonary embolism. In other words, the CT scans were much more ideal than they would be in a consecutive group of patients being scanned for pulmonary embolism, who are commonly short of breath and might have lung parenchymal or pleural abnormalities, or alterations in cardiac function that might reduce the technical adequacy of the study. Although this study of collimation showed that with thinner collimation more small vessels were well seen, it is unclear whether this finding would translate to a more realistic clinical population.

Retrospective Versus Prospective Selection

When patients are selected retrospectively, it is important to know why they were selected for imaging. Rather than representing all patients with a suspected disease or all patients in a specific clinical circumstance who presented for evaluation, it is more likely that patients might have been sent for imaging for clinical reasons that make them different than if the diagnostic test had been applied to all patients with the same disease or symptoms. Biases will be introduced by such patient selection that might overestimate the value of the diagnostic test being studied or the frequency with which specific abnormal findings are reported.

When looking at pulmonary embolism, the sensitivity of CT pulmonary angiography for small emboli has been questioned, leading investigators to look at the frequency with which isolated subsegmental or smaller pulmonary embolisms occur. Reported percentages have ranged broadly from 4% to 36% [17–20]. In one study, consecutive patients undergoing conventional angiography were studied, and 30% were found to have emboli in only subsegmental or smaller pulmonary arteries [20]. As the methods stated, these were consecutive patients undergoing pulmonary angiography, not consecutive patients with suspected pulmonary embolism. In fact, Oser et al. [20] stated in the discussion of their publication that

> … the vast majority of our patients had intermediate-probability lung scans; thus, the patients with a larger embolic burden, namely, those with high-probability scans, were potentially excluded. This selection bias is difficult to avoid in a retrospective series, as it reflects the hospital referral pattern.

With regard to CT and pulmonary embolism, in order to know the sensitivity of CT pulmonary angiography for pulmonary embolism in the general population of patients presenting with suspected pulmonary embolism, a prospective investigation of all patients with suspected pulmonary embolism is necessary, using a reference standard such as conventional pulmonary angiography. The goal should be to prospectively recruit all patients with suspected pulmonary embolism and have all patients undergo the test under evaluation— CT pulmonary angiography, and the reference test—conventional angiography. Consider the impact of retrospective selection of the sample on diagnostic accuracy in the following scenarios. If all patients undergoing both CT pulmonary angiography and conventional pulmonary angiography over the previous 2-year period formed the sample, the reasons that patients underwent both tests, and not just CT pulmonary angiography, impact sensitivity. If a large proportion of the conventional angiograms were obtained because of inconclusive findings or a technically poor CT pulmonary angiogram, then the sensitivity of CT pulmonary angiography will appear artificially low compared with sensitivity in the general population. If a normal CT pulmonary angiography is the predominant reason for obtaining conventional angiograms, the sensitivity of CT pulmonary angiography will again be low. In this case, the frequency of subsegmental emboli found at angiography will also be higher than would be found in the general population of patients with pulmonary embolism because patients with larger and more obvious emboli will not have undergone conventional angiography.

Which physicians accept and begin to use a new imaging test might also bias the results. For example, if physicians in the emergency department began using CT pulmonary angiography before most of the physicians taking care of inpatients, then the sensitivity of CT pulmonary angiography might be high, but would be biased by the type of patients that are seen in the emergency department, who in general might be healthier, younger, able to hold their breath better, or have less lung disease than hospitalized patients. On the other hand, if critical care medicine physicians accept CT pulmonary angiography earlier for intensive care unit patients, the sensitivity of CT pulmonary angiography might appear low because of the extensive parenchymal consolidation and pleural effusions that are often present in this population of patients who are often ventilator-dependent. In this way, the spectrum of disease or the case mix in the sample impacts the measured accuracy of the diagnostic test in question. This point reinforces the need to thoroughly describe the patient population studied.

Retrospective studies also suffer from recall bias. Suppose an investigator wants to determine the severity of dyspnea in patients with suspected pulmonary embolism, hypothesizing that patients with more severe dyspnea have a higher frequency of pulmonary embolism than patients with lesser degrees of dyspnea or no dyspnea at all. The investigator might be approaching this as a way to evaluate the likelihood of a patient's having pulmonary embolism and thus to triage patients to a diagnostic test within 1 hr versus within 4–6 hr, given the available imaging facilities. If an investigator questions all patients evaluated over the past year for suspected pulmonary embolism about their dyspnea, it is likely that the patients who were diagnosed with pulmonary embolism and hospitalized for treatment will remember their dyspnea more vividly and rate it as more severe than patients not diagnosed with pulmonary embolism who were sent home. This would exaggerate the difference in reported dyspnea in the two groups, compared with what would be seen if all of the patients were asked about dyspnea before undergoing any diagnostic test for pulmonary embolism and would thereby increase the likelihood that the investigator's hypothesis would be proven correct on analysis.

Consecutive Versus Nonconsecutive Selection

If patients are selected in a nonconsecutive manner, they might be inherently different from a population of all patients who meet inclusion criteria for a study. Suppose that the strategy were to recruit only the first patient seen each day who met the inclusion criteria for the study. It is possible that patients who are able to come for a 7:00 A.M. clinic appointment are different from patients who come later in the day. Perhaps they are less sick, resulting in a bias toward milder disease. Suppose that the strategy were to recruit only those patients meeting the inclusion criteria who are seen Monday to Friday between 8:00 A.M. and 5:00 P.M. If the study were looking at lung cancer staging accuracy, there might be little, if any, bias. However, in other circumstances, the patients might be inherently different from patients presenting to the emergency department in the evening with the same symptom complex. For example, if the study involved suspected myocardial infarction, the patients coming to the emergency department in the evening after a day of work might have had chest pain all day long and sought medical attention hours after the onset of the acute event, whereas patients coming during the day might have had symptoms of shorter duration. Because the time from onset of symptoms is critical to outcome after a myocardial infarction, patients presenting during the day might have a better outcome than patients presenting at night, independent of any therapeutic intervention.

Reference Standard

The choice of a reference standard impacts measurements of test accuracy. In contrast to the ideal scenario for evaluating the accuracy of CT pulmonary angiography described in the previous section, a methods section might read: "All patients with pulmonary embolism confirmed at autopsy who had undergone CT pulmonary angiography formed the sample." In this case, the sensitivity of CT pulmonary angiography might be higher than in the general population because patients dying from pulmonary embolism might have larger emboli than patients not dying from pulmonary embolism.

Another problem is commonly referred to as "workup bias" [21]. Whenever the reference test is selectively applied only to patients with a positive result on the test in question—for example, only patients with a positive CT pulmonary angiography—the reported sensitivity of CT pulmonary angiography will be artificially high at 100%, whereas the specificity will be artificially low.

When a new technology is compared with accepted reference tests or gold standards, the accuracy of the reference test is often called into question [22–27]. In the example of CT pulmonary angiography, the validity of conventional pulmonary angiography has been questioned. Several studies have reported poor interobserver agreement as to the presence or absence of emboli in subsegmental pulmonary arteries on conventional angiography. The Prospective Investigation of Pulmonary Embolism Diagnosis investigators (PIOPED) [28] found only 66% agreement among observers for isolated subsegmental emboli, compared with 98% at the lobar level and 90% at the segmental artery level. Similarly, Diffin et al. [17] reported interobserver agreement of only 45% for isolated subsegmental emboli at conventional angiography. If observers cannot agree on the gold standard, how can the new test, CT pulmonary angiography, be compared with it? This problem might lead investigators to look for a new sample population and apply a new gold standard. To do so might require an animal study with autopsy confirmation as the reference standard. For CT pulmonary angiography, Baile et al. [27] did just that. To compare the accuracy of CT pulmonary angiography and conventional angiography, these investigators instilled colored methacrylate beads into the pulmonary artery circulation of pigs, with a methacrylate cast of the pulmonary arteries used as the reference standard. These researchers found no statistically significant difference in CT pulmonary angiography and conventional angiography for the detection of emboli. However, if conventional angiography were used as the reference standard to which 1-mm CT pulmonary angiography was compared, conventional angiography would, by definition as the reference test, be 100% sensitive with a 100% positive predictive value, whereas CT pulmonary angiography would be considered only 76% sensitive with a positive predictive value of only 86%. If the sensitivity of a test is in question, surrogate measurements might be used to support the value of a negative test, such as patient outcome. For CT pulmonary angiography, most investigators have looked at series of patients gathered retrospectively with negative findings for pulmonary embolism on CT pulmonary angiography, and looked at the incidence of pulmonary embolism over the next 3–12 months. These studies have shown that pulmonary embolism occurs with the same frequency after negative findings on CT pulmonary angiography as after negative findings on conventional angiography [29, 30].

Imaging-Based Selection

It is often convenient to select patients who have undergone an imaging test, or patients who are going to be sent for imaging, to form a sample. This is referred to as imaging-based selection. However, patients who undergo imaging might not be representative of all patients with a specific diagnosis or symptom. Consider describing the appearance of lung cancer on MR imaging. Investigators could generate a list of all patients at their facility who underwent thoracic MR imaging in the past or will be undergoing MR imaging over the next year, who have a diagnosis of lung cancer. A fairly high proportion of these patients will likely have masses that abut or invade the mediastinum. This does not mean that this proportion of all patients presenting with lung cancer have mediastinal invasion, because the patients undergoing MR imaging for lung cancer are usually preselected because of a suspicion of mediastinal invasion on CT, and therefore the high incidence should not be surprising. To know what the appearance of lung cancer is on MR imaging or to determine the accuracy with which MR imaging can detect lung cancer requires that all consecutive patients with a diagnosis of lung cancer over a specified period of time undergo MR imaging. Although this example might seen fairly obvious, the literature is full of examples in which this type of selection bias impacts study results, although the impact on the results might be less obvious than in the example and not initially apparent.

Generalizability

Who was studied impacts to whom the results can be applied. If all patients presenting with suspected pulmonary embolism undergo a diagnostic test, the results will be different than if only patients with acute right heart failure and suspected massive pulmonary embolism are studied, or if patients who have an inconclusive result from another diagnostic test, such as a ventilation–perfusion scan, are studied. Similarly, how the test performs on inpatients or intensive care unit patients might be different from how it performs in outpatients or patients presenting to an emergency department, who are less likely to have coexisting lung disease or abnormal chest radiographic findings. In selecting a population to study for an investigation, it is important to consider to whom the information derived from that investigation can be applied.

For example, recently the prevalence of isolated subsegmental pulmonary embolism has been debated as part of the question of how accurate CT pulmonary angiography needs to be for the detection of subsegmental pulmonary embolism. If isolated subsegmental pulmonary embolism rarely occurs, then the technology might not need to be accurate for vessels of this size. However, if isolated subsegmental emboli are commonly seen, then the technology might need to be accurate. In one study, isolated subsegmental pulmonary embolism was reported to occur in 36% of patients diagnosed with pulmonary embolism [19]. In another study, isolated subsegmental pulmonary embolism was reported to occur in only 6% of patients diagnosed with pulmonary embolism [18, 31]. Which more realistically represents a population of all patients with suspected pulmonary embolism? The former study was performed to prospectively compare helical CT with pulmonary angiography for the detection of pulmonary embolism in patients with an unresolved clinical and ventilation–perfusion scan diagnosis of pulmonary embolism. Patients with either a normal perfusion scan or a high-probability scan, the two groups for whom no pulmonary embolism and definite pulmonary embolism were diagnosed, and perhaps the easiest patients for CT to evaluate, were not studied with CT. Therefore, it is likely that 36% is an overestimate of the frequency with which isolated subsegmental pulmonary embolism occurs. The latter study was the PIOPED study [18, 28], in which patients with suspected pulmonary embolism were prospectively enrolled at multiple medical centers, and all patients underwent ventilation–perfusion scannning and conventional pulmonary angiography.

The results described by Goodman et al. [19] can be generalized only to patients with an unresolved clinical suspicion for pulmonary embolism after ventilation–perfusion scanning who underwent CT, as the title of that investigation states clearly. The results can also be generalized only to patients undergoing CT with the technique that was reported (5-mm collimation, pitch of 1:1, covering 12 cm of the thorax, and viewed on hard-copy film). Imaging technology rapidly evolves. Several researchers after Goodman et al. have reported on CT pulmonary angiography at 3-mm collimation [32–34]. The ability to perform multidetector CT pulmonary angiography using 1.25-mm collimation of the entire thorax is now possible, and interpretation on workstations has been shown to improve detection of pulmonary embolism compared with film-based interpretation [35]. However, the published literature lags behind what the technology of today is capable of. As investigators plan to study a new technology, they should consider ways to recruit a larger number of patients more quickly to answer the question they propose before the technology is outdated [36].

Several studies have reported the findings of pulmonary embolism detected incidentally on CT scans obtained for other reasons [13, 35, 37–39]. It would be incorrect to draw a conclusion that the anatomic distribution of pulmonary emboli in these patients is the same as in a population of patients presenting with clinical signs or symptoms of pulmonary embolism. In one series of nine patients, no incidentally detected emboli were seen beyond the segmental arteries [39]. This result does not mean that subsegmental pulmonary embolism does not occur as an incidental finding. The CT scans in this study might have been done with protocols used for general thoracic CT, rather than using a thin-section, rapid IV–contrast injection protocol CT, or the researchers may not have used a workstation for interpretation—both factors that improve the accuracy of CT pulmonary angiography for pulmonary embolism, particularly for small arteries.

Conclusion

This article has reviewed the current concepts of target population, study population, and sample. These terms need to be used appropriately in the design, execution, and reporting of clinical research in radiology. The article also has discussed considerations for the definition and selection of these entities. Other considerations, such as randomization, statistical power, and sample size, that are relevant specifically to the selection of sample, will be the subject of future articles in this series.

References

1. Elston R, Johnson W. Populations, samples and study design. In: *Essentials of biostatistics.* 2nd ed. Philadelphia: Davis, **1994**:15–16
2. Webb WR, Gatsonis C, Zerhouni EA, et al. CT and MR imaging in staging non-small cell bronchogenic carcinoma: report of the Radiologic Diagnostic Oncology Group. *Radiology* **1991**;178:705–713
3. Blackmore CC, Black WC, Jarvik JG, Langlotz CP. A critical synopsis of the diagnostic and screening radiology outcomes literature. *Acad Radiol* **1999**;6[suppl 1]:S8–S18
4. Hillman BJ. Research in radiology departments. *Invest Radiol* **1993**;28[suppl 2]:S44–S48
5. Applegate KE. Study design: pros and cons. In: *2000 annual meeting scientific session.* Oak Brook, IL: Society of Health Services Research in Radiology, **2000**
6. Holman BL. The research that radiologists do: perspective based on a survey of the literature. *Radiology* **1990**;176:329–332
7. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Stat Med* **1984**;3:361–373
8. Godwin JD, Webb WR, Gamsu G, Ovenfors CO. Computed tomography of pulmonary embolism. *AJR* **1980**;135:691–695
9. Sinner WN. Computed tomography of pulmonary thromboembolism. *Eur J Radiol* **1982**;2:8–13
10. Ovenfors CO, Godwin JD, Brito AC. Diagnosis of peripheral pulmonary emboli by computed tomography in the living dog. *Radiology* **1981**;141:519–523
11. Cholankeril JV, Ketyer S, Ramamurti S, Millman AE. Pulmonary embolism demonstrated by computerized tomography. *J Comput Assist Tomogr* **1982**;6:135–139
12. Breatnach E, Stanley RJ. CT diagnosis of segmental pulmonary artery embolus. *J Comput Assist Tomogr* **1984**;8:762–764
13. Allen BT, Day DL, Dehner LP. CT demonstration of asymptomatic pulmonary emboli after bone marrow transplantation: case report. *Pediatr Radiol* **1987**;17:65–67
14. Browner WS, Newman TB, Cummings SR. Designing a new study. III. Diagnostic tests. In: Hulley SB, Cummings SR, eds. *Designing clinical research.* Baltimore: Williams & Wilkins, **1988**:87–97
15. Hulley SB, Gove S, Browner WS, Cummings SR. Choosing the study subjects: specification and

sampling. In: Hulley SB, Cummings SR, eds. *Designing clinical research.* Baltimore: Williams & Wilkins, **1988**:10–30

16. Remy-Jardin M, Remy J, Artaud D, Deschildre F, Duhamel A. Peripheral pulmonary arteries: optimization of the spiral CT acquisition protocol. *Radiology* **1997**;204:157–163
17. Diffin DC, Leyendecker JR, Johnson SP, Zucker RJ, Grebe PJ. Effect of anatomic distribution of pulmonary emboli on interobserver agreement in the interpretation of pulmonary angiography. *AJR* **1998**;171:1085–1089
18. Stein PD, Henry JW. Prevalence of acute pulmonary embolism in central and subsegmental pulmonary arteries and relation to probability interpretation of ventilation/perfusion lung scans. *Chest* **1997**;111:1246–1248
19. Goodman LR, Curtin JJ, Mewissen MW, et al. Detection of pulmonary embolism in patients with unresolved clinical and scintigraphic diagnosis: helical CT versus angiography. *AJR* **1995**;164:1369–1374
20. Oser RF, Zuckerman DA, Gutierrez FR, Brink JA. Anatomic distribution of pulmonary emboli at pulmonary angiography: implications for cross-sectional imaging. *Radiology* **1996**;199:31–35
21. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology* **1988**;167:565–569
22. Chugh SK. Stress echo training: need for a better gold standard—the invasive viewpoint. *Eur Heart J* **2000**;21:859–860
23. Shah A, Wagner GS, Granger CB, et al. Prognostic implications of TIMI flow grade in the infarct related artery compared with continuous 12-lead ST-segment resolution analysis: reexamining the “gold standard” for myocardial reperfusion assessment. *J Am Coll Cardiol* **2000**;35:666–672
24. Koretz RL. Prospective randomized controlled trials: when the gold in the gold standard isn’t pure. (commentary) *J Parenter Enteral Nutr* **2000**;24:5–6
25. Rolfe MW, Solomon DA. Lower extremity venography: still the gold standard. (editorial) *Chest* **1999**;116:853–854
26. Kalodiki E, Nicolaides AN, Al-Kutoubi A, Cunningham DA, Mandalia S. How “gold” is the standard? interobservers’ variation on venograms. *Int Angiol* **1998**;17:83–88
27. Baile EM, King GG, Muller NL, et al. Spiral computed tomography is comparable to angiography for the diagnosis of pulmonary embolism. *Am J Respir Crit Care Med* **2000**;161:1010–1015
28. The PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA* **1990**;263:2753–2759
29. Goodman LR, Lipchik RJ, Kuzo RS, Liu Y, McAuliffe TL, O’Brien DJ. Subsequent pulmonary embolism: risk after a negative helical CT pulmonary angiogram—prospective comparison with scintigraphy. *Radiology* **2000**;215:535–542
30. Garg K, Sieler H, Welsh CH, Johnston RJ, Russ PD. Clinical validity of helical CT being interpreted as negative for pulmonary embolism: implications for patient treatment. *AJR* **1999**;172:1627–1631
31. Worsley DF, Alavi A. Comprehensive analysis of the results of the PIOPED study: prospective investigation of pulmonary embolism diagnosis study. *J Nucl Med* **1995**;36:2380–2387
32. Garg K, Welsh CH, Feyerabend AJ, et al. Pulmonary embolism: diagnosis with spiral CT and ventilation-perfusion scanning—correlation with pulmonary angiographic results or clinical outcome. *Radiology* **1998**;208:201–208
33. Mayo JR, Remy-Jardin M, Muller NL, et al. Pulmonary embolism: prospective comparison of spiral CT with ventilation-perfusion scintigraphy. *Radiology* **1997**;205:447–452
34. Remy-Jardin M, Remy J, Deschildre F, et al. Diagnosis of pulmonary embolism with spiral CT: comparison with pulmonary angiography and scintigraphy. *Radiology* **1996**;200:699–706
35. Gosselin MV, Rubin GD, Leung AN, Huang J, Rizk NW. Unsuspected pulmonary embolism: prospective detection on routine helical CT scans. *Radiology* **1998**;208:209–215
36. Baum RA, Rutter CM, Sunshine JH, et al. Multicenter trial to evaluate vascular magnetic resonance angiography of the lower extremity: American College of Radiology Rapid Technology Assessment Group. *JAMA* **1995**;274:875–880
37. Verschakelen JA, Vanwijck E, Bogaert J, Baert AL. Detection of unsuspected central pulmonary embolism with conventional contrast-enhanced CT. *Radiology* **1993**;188:847–850
38. Winston CB, Wechsler RJ, Salazar AM, Kurtz AB, Spirn PW. Incidental pulmonary emboli detected at helical CT: effect on patient care. *Radiology* **1996**;201:23–27
39. Romano WM, Cascade PN, Korobkin MT, Quint LE, Francis IR. Implications of unsuspected pulmonary embolism detected by computed tomography. *Can Assoc Radiol J* **1995**;46:363–367

The reader’s attention is directed to the earlier articles in the Clinical Research Series: Introduction, which appeared in the February 2001 issue; Framework, April 2001; Protocol, June 2001; and Data Collection, October 2001.

Fundamentals of Clinical Research for Radiologists

Craig A. Beam[1]

Statistically Engineering the Study for Success

A scientific study is a dynamic endeavor the outcome of which can never be wholly determined in advance. However, over years of experience, the art and science of engineering a scientific study have evolved so that the savvy investigator can dictate the limits of risk and the likelihood of outcomes from this dynamic process of discovery. This particular form of art and science is commonly referred to as "experimental design."

When reading the scientific literature or designing studies, every clinical radiologist should be aware of and concerned about three main considerations of modern experimental design that apply to research in clinical radiology (Fig. 1). The first consideration is the extent to which the findings of the study might mislead ("bias"). Another consideration is the ability of the study to reveal something important ("power"). The final consideration is the desire to create useful information ("precision") from the research. The deceptively simple statistical concept of the "average" will be shown to be central to many of these considerations.

In this article, I will review these three key considerations, each of which needs to be adequately appreciated and addressed by investigators seeking to design a successful diagnostic radiology study. Because successfully engineering the scientific study requires drawing heavily on both clinical and statistical sciences, interdisciplinary collaboration should be encouraged and nurtured. In this way, research in clinical radiology will mature into a modern scientific discipline. Motivating such collaborations is a goal of this series of articles.

Minimizing Bias

Statistical Meaning of the Word "Bias"

As with many other words, the word "bias" is interpreted differently by different individuals. However, statistical science has a definite and precise meaning for this word, and because statistical science provides the foundation of modern experimental design, it is this interpretation that must be addressed by successful scientific studies in clinical radiology.

Statistically, bias is a property of averages. A statistical measure is said to be biased if, on average, it does not equal what it is intended to estimate. To say that a study is biased is to say that it was conducted in such a fashion that, on average, the measurements from the study are biased.

What Is the Weight of a 1-Oz Marble?

Suppose that a group of researchers had a reliable spring scale with which to measure the weight of marbles. Reliable means that the researchers generally get the same value each time they weigh the same marble. Now suppose that the researchers have a marble that they know weighs exactly 1 oz and thus that marble becomes the gold standard. They weigh this marble five times and get the following values: 1.1, 1.2, 1.2, 1.1, and 1.1 oz. The values are always slightly more than the marble's true weight of 1.0 oz. Sometimes the "error" is 0.1 oz, and other times it is 0.2 oz. The average of these errors is 0.14, and so, on average, the scale errs by 0.14 oz.

Statistically, this measurement would be described as biased: it tends to overestimate true weight by 0.14 oz. Knowing this bias, the researchers could correct the scale by ad-

Received November 15, 2001; accepted after revision January 22, 2002.

Series editors: Craig A. Beam, C. Craig Blackmore, Stephen J. Karlik, and Caroline Reinhold.

This is the sixth in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1] Department of Radiology, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226. Address correspondence to C. A. Beam.

AJR 2002;179:47–52

0361–803X/02/1791–47

vising users to always subtract 0.14 from the reading. Then, although individual measurements may be a little off, on average, the users will get the correct value. Thus, the corrected measuring device would be said to be "unbiased" for the weight of marbles.

The previous case is an example of measurement bias. Studies can be affected by other biases as well. Studies of diagnostic technologies have their own special biases [1] with which the reader of the literature of diagnostic radiology should be familiar. These specific biases will be the subject of a subsequent article on the clinical assessment of diagnostic technologies in this series. For the present discussion, however, I will focus on two biases that affect every type of clinical study. These biases come about by the way subjects are selected for, and participate in, a study.

Selection Bias

The article in this series by Ella Kazerooni [2], "Population and Sample," makes it very clear that subjects selected for a study must be representative of some clinically relevant population. One of several statistical motivations for this notion has to do with bias: We want the measures from our study to reflect the value of the measures in the general population. We do not want to be off the mark, so to speak. To accomplish this objective, we must have a sample that in some way reflects the population being studied.

Recalling that the statistical meaning of bias involves averages, we can restate our consideration as seeking to sample from the study population in such a way that our measurements on average equal the value in the population. Luckily, this goal can be accomplished by the well-known mechanism of random sampling.

By randomly sampling, we follow a procedure that guarantees that every sample has the same chance of being selected for our study. If we decided to do a study with a sample of 100 randomly selected subjects from our study population, we would have to follow a method of sampling so that every possible sample of 100 subjects would be equally likely to be selected. How does random sampling ensure that our results will not be biased? The answer to this question requires the logic of statistical science. However, an intuitive answer is that measures that are simple averages will be unbiased for the population average when the measures are based on random samples.

Are simple averages relevant to clinical radiology research? Thankfully, the answer in many cases is yes. Many published clinical studies report means (which, of course, are averages) of measurements. Measures of diagnostic accuracy such as sensitivity and specificity, which are frequently reported, are averages as well. Other commonly used measures in clinical radiology are not simple averages but do enjoy the property of being unbiased when based on random samples. Examples of these are the slope in linear regression and the nonparametric receiver operating characteristic curve area.

Participation Bias

When conducting research that compares groups of subjects, care should be taken to ensure that the group assignments are free of bias. In other words, the way in which subjects participate should not bias the findings of the study. The mechanism by which this bias is typically eliminated is randomization. In the valuable reference book *Statistics in Medicine*, Theodore Colton [3] writes, "Randomization ensures that the personal judgment and prejudices of the investigator and of the patient do not influence treatment allocation."

Randomization, in fact, has become the gold standard for the clinical trial. For example, popular guidelines for evaluating the quality of research are based on the assumption that the controlled randomized trial is the epitome of study design. Some scientists advise using randomization simply because it is a good strategy for success in publication: "Without proper randomization, the investigator is immediately on the defensive and increases his vulnerability to the critical onslaught of his peers." [3].

What actually is randomization? First, let us specify what it is not. Colton [3] admonishes:

> It is worthwhile to point out that one should not confuse randomization… with haphazard assignment…. The pattern of assignment to treatment may appear to be haphazard, but this arises from the haphazard nature with which digits appear in a table of random numbers, and not the haphazard whim of the investigator in allocating patients.

Randomization is an objective process that takes group assignment out of the hands of humans and gives the responsibility to the random number generator. Once the human factor in group assignment is eliminated, we can make the important assertion that the process of allocation was unbiased. The statistical significance of this step is that each possible allocation had an equal chance of occurring so that, on average, the findings from the study are not affected by the way the subjects participated in the study.

It is widely held that randomization "averages out" the effect of influencing factors

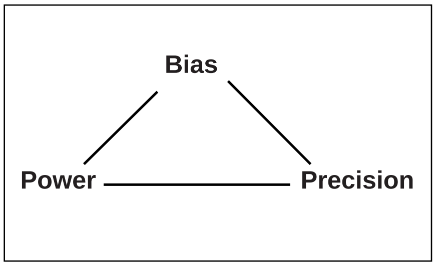


Fig. 1.—Diagram illustrates the three elements of study design.

TABLE 1 **Contrast-to-Noise Ratios (CNRs) for Six Subjects Assigned to Unenhanced or Enhanced MR Imaging Groups Using Randomization**

Subjects	CNR in Unenhanced Group	CNR in Enhanced Group
1 and 2	8	9
3 and 4	13	7
5 and 6	15	20
Means	12	12

Note.—CNRs given in second and third columns apply to subjects retrospectively.

TABLE 2 **Contrast-to-Noise Ratio (CNR) Data Grouped by Presence of Cirrhosis in Enhanced and Unenhanced Imaging Groups**

Presence of Cirrhosis in Subjects ($n = 6$)	CNR in Unenhanced Group	CNR in Enhanced Group
Yes ($n = 3$)	8	9 and 7
No ($n = 3$)	13 and 15	20

that are unknown to the investigator. This tenet is true and provides another example of how the concept of the average is fundamental to our modern understanding of experimental design. However, the benefit of randomization is realized only if the averaging is performed across the many different ways of allocating subjects to treatment. In any one study, which can have only one such allocation, an imbalance of factors could influence the findings. Randomization does not guarantee an equitable allocation in any particular study; its benefits accrue as we consider the process of averaging across studies.

Consider the following study: Six subjects are selected for a clinical study of gadolinium enhancement of breath-hold T2-weighted MR imaging of hepatic lesions. Suppose that enhancement will be measured with a contrast-to-noise ratio (CNR) determined by dividing the difference between the lesion and liver signal intensities by the standard deviation of the background noise. Now suppose that the researchers wish to compare the CNR in the unenhanced section of the liver with the CNR in the enhanced section of the liver. However, the institutional review board requires the use of separate groups of subjects. Therefore, the subjects must be assigned to one of two "treatment" groups. How should the assignments be made?

If the investigators were to use randomization in this study, they would have to apply a mechanism that would give each possible allocation of three subjects the same chance to be in the enhanced MR imaging group. Note that randomization does not mean assigning individuals to treatments according to no discernible plan or pattern. For example, randomization would not occur if the first three patients who showed up at clinic were assigned to the gadolinium-enhanced imaging group and the next three to the unenhanced imaging group. That is not randomization because the researchers have not ensured that every allocation of three individuals to the enhanced imaging group was equally likely. The researches cannot feign ignorance either. Perhaps those three individuals who were assigned to the enhanced imaging group always show up early in the morning, and so the others would never have a chance to be in the group that undergoes enhanced MR imaging. In sum, to say that subjects were randomly assigned to treatments is to say that complete control had been exercised over the allocation mechanism in a quite definite way.

Randomization controls the bias of allocating individuals to treatments by the same averaging seen with random sampling. To say that randomization averages out the influence of unknown effects is to say that, on average, the values resulting from a study will equal the average of the values resulting from every possible experimental allocation of subjects to treatments.

Suppose that randomization was followed, and the data in Table 1 were observed. One would probably conclude from this study that the use of gadolinium does not improve the CNR because the mean CNR of the two treatment groups are equal. Because randomization was used, researchers would trust that any effects that might have biased the findings have been averaged out. "Trust" is the operative word: Randomization does not guarantee that the group allocation actually realized in this particular instance was equal with respect to characteristics that might be important. Randomization is only a property of averages. Any one particular randomization can, by chance, lead to severe disparities between the two groups in some characteristic.

Actually, the principal investigator of this supposed study was wise enough to design into it the collection of extra information about the subjects. One extra (or concomitant) variable measured was whether the subject had cirrhosis of the liver (determined independently of the measurement of the CNR). Table 2 presents the raw data from this study categorized by the treatment received (i.e., enhanced or unenhanced MR imaging) and by the presence of cirrhosis in the six subjects.

Examination of this table shows that three of the subjects selected for the study had cirrhosis and that two of these subjects were assigned by the process of randomization to the treatment (gadolinium-enhanced MR imaging) group. Conversely, two of the subjects without cirrhosis were assigned to the "control" (unenhanced MR imaging) group. Obviously, the occurrence of cirrhosis was not equally represented in the two groups. Did randomization fail? No. The allocation used in this study was just one possible allocation of the six subjects to the two treatment groups. There are, in fact, 20 different ways to assign these six subjects to the two groups. The investigators used a method that picked one of these assignments at random—that is, in a way that each assignment was equally likely (one in 20) to be picked. Thus, they randomly assigned subjects to the groups. This time, randomization just happened by chance to come up with the assignment of two subjects with cirrhosis to the treatment group and two subjects without cirrhosis to the control group.

The investigators are concerned because they believe that the presence of cirrhosis is likely to have dampened the enhancement of the gadolinium. What can they do? They consult their statistician who then generates Table 3. From this analysis, it becomes obvious that there is no benefit for subjects with cirrhosis but a big benefit for other subjects.

The need to be cautious with the results from even the most carefully planned randomized trial is appreciated by experienced researchers. Colton [3], for example, observes:

> Randomization achieves a balance in the long run. However, with a small series of patients, randomization may not always produce groups that are alike in every respect…. [A]s a general rule, a report of a clinical trial should include among its first tables one in which the treatment and control groups are compared on the several important characteristics relating to the disease under study.

In sum, the gadolinium-enhanced imaging example shows that successful study design requires collection of data that could plausibly influence the outcome of the study, good statistical methods by which to adjust the outcomes for these concomitant variables, and randomization of subjects to average out the possible influence of unrecognized factors.

Power in Comparisons

A successful study finds something. If a study does not find something, then the re-

TABLE 3 Mean Contrast-to-Noise Ratios (CNRs) of the Unenhanced and Enhanced Imaging Groups Controlling for Cirrhosis

Cirrhosis Present	Mean CNR in Unenhanced Group	Mean CNR in Enhanced Group	Difference in Means
Yes	8	8	8 – 8 = 0
No	14	20	20 – 14 = 6

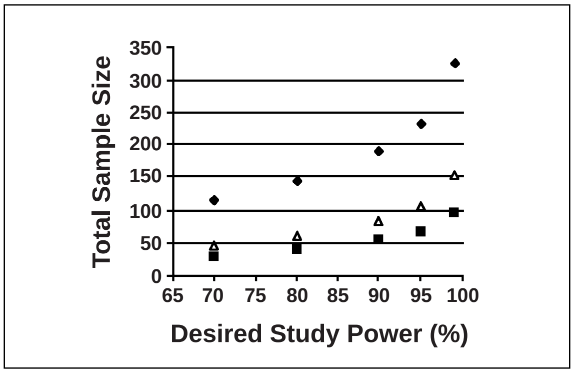


Fig. 2.—Graph plots relationship between study design and power comparing two designs commonly used in diagnostic test evaluation: independent groups and paired groups. Paired groups study design is shown as requiring fewer subjects than independent groups design for any desired power in study. ◆ = independent groups study, ■ = minimal disagreement in paired groups study, ▲ = maximal disagreement in paired groups study.

searchers in a successful study have the confidence to say that if there had been something to find, they probably would have found it. The ability of a study to detect a specific difference among study groups is its power. The logical expectation is that the power of any study is greater when measuring greater differences. For example, collecting data to show that two imaging modalities differ by 50% in their sensitivities should be easier than collecting data to show that they only differ by 1%.

To be clinically useful, a successful study must have the power to detect the smallest difference that is deemed clinically important. If a difference in sensitivity as small as 1% leads to clinically important differences in patient outcome, we then are required to design a study that has adequate power to detect a difference as small as 1% in the sensitivities of the two modalities. If, however, our study was able to detect only a larger difference—for example, 20%—and gave negative results, we could not say with confidence that no clinically significant difference exists between the modalities. The difference might, indeed, lie between 10% and 20%, a range we consider clinically important. We would have to regard our study as unsuccessful.

Statistically, power is expressed as the probability of rejecting the hypothesis of no difference (the “null” hypothesis) when, in fact, a specific, clinically important difference does exist. The concept of power depends on specification of hypotheses and definition of a specific, clinically important difference. To assess the power of a study, it is not enough to say that the sensitivity of the new test is greater than that of the standard. A definite value for this difference must be specified.

Two important aspects of study design determine the power of a study. One is sample size, and the other is the design itself. A successful study is one that has sufficient power to detect the smallest clinically significant difference. The sample size that ensures this power is thus a requirement for the successful study. Determination of the sample size is the purview of statistical science, and so the required sample size for a study is often the contribution of the collaborating statistician. However, determination of sample size and power also requires specification of the smallest clinically important difference for the problem at hand. This determination is the purview of clinical medicine. Thus, statistically engineering the study for power should be a collaborative undertaking between clinical and statistical scientists.

Although the role of sample size and power is well known in medical circles, I do not think the role of experimental design and power is as well appreciated. The graph in Figure 2 illustrates the importance of the relationship. This graph depicts sample size requirements for two basic study design types that one might consider when comparing the diagnostic accuracies of two modalities.

Our scenario is that a clinical radiologist seeks to compare the sensitivity of a new diagnostic modality against that of an established modality. Based on her understanding of the medical literature, and of the costs and benefits to her patients in testing for this particular condition, the clinical researcher has determined that the smallest clinically relevant difference in sensitivities for this diagnostic problem is 5%.

The two basic study designs for this sort of clinical trial are the “independent groups” design and the “paired groups” design. The independent groups design specifies that the assignment of each of the study’s subjects to

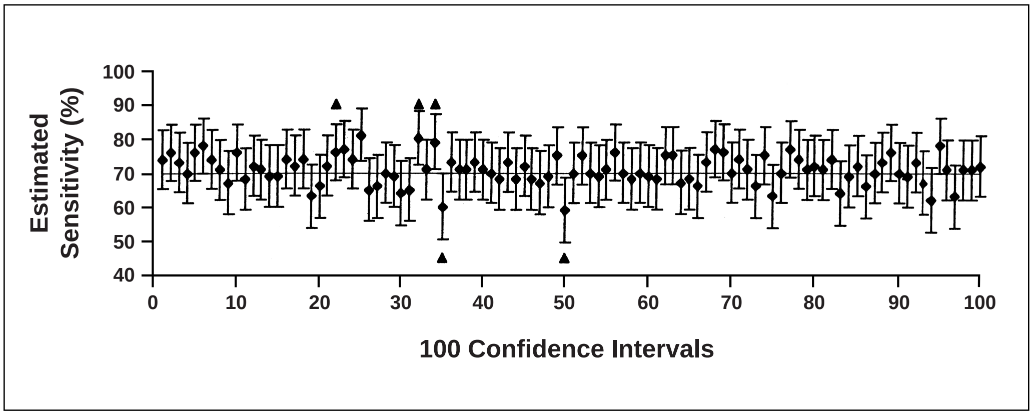


Fig. 3.—Graph shows computer-simulated sampling of 100 confidence interval (CI) point estimates of test sensitivity to illustrate term “95% CI.” For 100 samples of subjects from large population, the sensitivity and 95% confidence interval are plotted in order. Horizontal line at 70% represents true sensitivity. Point estimates (◆) fall around true value. Results from some samples overestimate and some underestimate. Approximately 11 of 100 simulated point estimates appear to be exactly correct. Bars around each point represent associated 95% CI. In estimates in which bars overlap horizontal line (true sensitivity), CI contains true value of quantity being estimated. In estimates in which bars do not overlap line, CI failed to capture true value. (Intervals that failed to capture true value are represented by ▲.) Of 100 CIs randomly generated, five failed to capture true value and 95 did capture it. In large series of such intervals, CIs will give range that captures true value in 95% of cases.

one of two groups should be randomized. One group will be imaged using the reference modality, and the other group will be imaged using the new modality. In the paired group design, each of the subjects is imaged using both of the modalities being studied. Preferably, the interpretation of each modality is done independently of the result of the other modality, and the order in which the subjects are imaged with each modality is also randomized.

Figure 2 shows the total sample size required to achieve various levels of statistical power for the two designs. In fact, there are two sets of points for the paired groups design because the power of this design also depends on the extent to which the two modalities disagree (i.e., the proportion of patients for whom one modality is positive and the proportion for whom one is negative and vice versa). One set of points shows sample size required when the disagreement between the modalities is minimal, and the other set shows the power of the study when the modalities disagree as much as possible. (More details about these considerations and computations can be found in an earlier article that I wrote for *AJR* [4].)

Figure 2 provides confirmation of the intuitive realization that greater power in a study requires a larger total sample size, or, conversely, the intuitive realization that a larger sample size means greater power. This relationship between power and sample size is true regardless of which study design is chosen.

However, note that the paired groups design requires a smaller total sample size for any power we may wish to achieve. For example, to achieve 90% power requires approximately 200 subjects with the independent groups design but only approximately 50 subjects when the paired groups design is used and the measures of the two modalities under study have the lowest level of disagreement possible. Even in the worst-case scenario, in which there is maximal disagreement between the modalities, the paired groups design requires only approximately 80 subjects.

The previous example illustrates that study design can greatly increase power for a given sample size or, conversely, that the study design can reduce the sample size needed to obtain a specific power. Successful studies are ones that achieve the desired power economically. Knowledge of study design is, therefore, essential to the engineering of powerful and economical scientific studies.

Precision in Estimation

In recent years, the trend in the medical literature has been to pay less attention to tests of hypotheses and give more attention to estimations. In fact, several authors have debated the issue. In the context of diagnostic radiology, the seminal article by James Hanley [5] "The Place of Statistical Methods in Radiology (and in the Bigger Picture)" is worthy of special attention. In that article, Hanley notes:

> The biggest objection to a statistical test is that it answers with a "yes" or a "no" an overly simplistic question: Is there some difference? The emphasis on significant differences…distracts from the real (issue), which is how big is the difference….

Given this recent trend, the design of the modern study in clinical radiology research must ensure success in estimation. The phrase "success in estimation" means that, statistically, the study has been designed to achieve sufficient precision in estimation with a desired level of confidence. Understanding how to design a study to be successful in estimation requires, then, an understanding of the statistical concepts of precision and confidence.

To estimate the sensitivity of a new diagnostic technology, we would do well to follow the direction given by Kazerooni in "Population and Sample" [2] and perform the test on a random sample from the study population. Because our sample is, of necessity, not the complete population of interest, we would expect imprecision in our estimate of sensitivity from this one sample. Being scientifically sophisticated, we are not satisfied in reporting only the estimated sensitivity but also want to assess the probable error in our estimate. The standard way to both report an estimate and provide an assessment of probable error is through the use of statistical confidence intervals (CIs).

A statistical CI of a quantity is a range of values along with a statement of the level of confidence. Usually, the CI accompanies a single value (or point) estimate of the quantity. For example, if the sample previously discussed yielded an estimated sensitivity of 75% with an accompanying 95% CI for values ranging from 67% to 83%, how should we interpret these values?

The observed sensitivity is 75%, so that is our point estimate. However, we estimate the value might be within the range of values from 67% to 83% with 95% confidence. The adjective "confidence" in the phrase "confidence interval" is not an assertion of personal belief. The term has an explicit statistical meaning that, not surprisingly, is related to the long-term process of sampling. To say that the interval is a 95% CI means that the interval was formed by a statistical method in such a way that if a large number of random samples were taken from the study population and an interval were computed for each sample, 95% of these intervals would contain the true value of the sensitivity of the test. Figure 3 is a graph depicting this concept using a computer-simulated experiment.

Another important feature of a confidence interval is its width. Wide confidence intervals are less informative than narrow ones. For example, to say that the sensitivity of a test falls between 68% and 72% is much more informative than saying the sensitivity falls somewhere between 0% and 100%.

The width of a confidence interval is its precision. Successful studies provide precise estimates. Therefore, engineering the successful study requires first specifying the precision the investigators wish to obtain. As

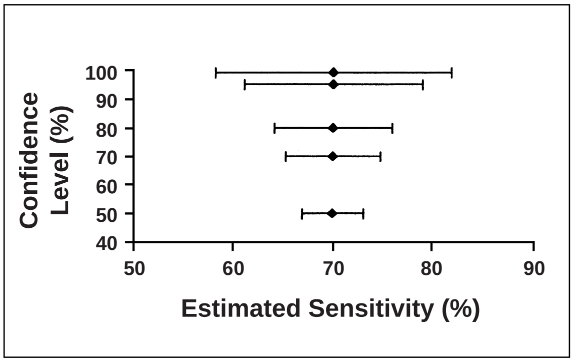


Fig. 4.—Graph depicts relationship between confidence interval (CI) precision, or width, and confidence level. Bars represent CIs (estimated sensitivity) and ◆ represents true sensitivity. As confidence level decreases, precision increases.

in considering power, specification of precision should in some way reflect a clinically relevant definition of precision. For example, if the researchers want to estimate sensitivity, it might be relevant clinically to require the precision of estimation to be within 5% of the true value if the researchers conclude that sensitivities this similar are virtually equal for clinical purposes.

Having specified the precision to be achieved by the study, the researchers have basically two design considerations by which to achieve this goal. One consideration is sample size. As expected, the precision of a confidence interval increases (i.e., its width decreases) with a larger sample size. Therefore, when designing a study for estimation, one must select a sample size large enough to achieve a desired precision in the confidence intervals.

Another way by which to achieve greater precision is by manipulating the level of confidence. Although the standard, by and large, for CIs is the 95% level, there is nothing sacred about this number. Other levels of confidence could be considered. The problem is, however, justifying this break from tradition.

Figure 4 shows the impact of changing the level of confidence on the precision of CIs based on the same sample size. Precision (interval width) is greatest for smallest confidence. In other words, precision and level of confidence exist in a trade-off relationship. Precision can be increased by decreasing confidence. In most cases, choosing precision at the expense of confidence will probably not be an acceptable trade-off. To alter the confidence level, one has to argue effectively that not following the status quo 95% level was appropriate. Generally, however, people set the level at 95% and find the sample size required to obtain adequate precision in estimation.

In this article, I have reviewed some of the key considerations in modern experimental design as they apply to diagnostic radiology. Each of these considerations—bias, power, and precision—should be addressed by investigators who want to design a successful study in diagnostic radiology. Because engineering a successful scientific study requires the expertise of both the clinical and statistical sciences, collaboration between these disciplines should be nurtured. In this way, research in clinical radiology will mature into a modern scientific discipline.

References

1. Begg CB. Assessment of radiologic tests: control of bias and other design considerations. *Radiology* **1988**;167:565–569
2. Kazerooni E. Population and sample. *AJR* **2001**;177:995–999
3. Colton T. *Statistics in medicine*. Boston: Little, Brown, **1974**
4. Beam CA. Strategies for improving power in diagnostic radiology research. *AJR* **1992**;159:631–638
5. Hanley JA. The place of statistical methods in radiology (and in the bigger picture). *Invest Radiol* **1989**;24:10–16

The reader's attention is directed to the earlier articles in the Clinical Research Series: Introduction, which appeared in the February 2001 issue; Framework, April 2001; Protocol, June 2001; Data Collection, October 2001; and Population and Sample, November 2001.

Cheryl R. Herman[1]
Harmindar K. Gill
John Eng
Laurie L. Fajardo

Screening for Preclinical Disease: Test and Disease Characteristics

Screening is the application of a test to detect a potential disease or condition in an individual who has no known signs or symptoms of that disease or condition [1]. In general, screening has two major objectives. One is the early detection of disease at a point when treatment is more effective, less expensive, or both. Here, the implicit assumption underlying the concept of screening is that early detection—before the development of symptoms—will lead to a more favorable prognosis because intervention initiated before the disease is clinically manifested will be more effective than treatment provided at a later stage of the disease [2, 3]. The second objective in screening is to identify risk factors that render an individual at a higher than average risk for developing a disease, with the goal of modifying the risk factors to prevent or minimize the disease [4–6]. The application of imaging examinations for disease screening is most often based on the first objective.

Although medical imaging is used in the diagnosis of most human ailments, mammography is the only diagnostic imaging examination currently in widespread use as a screening tool [7]. Multidetector CT is being evaluated as a means of detecting early-stage lung carcinoma [8, 9] and colorectal adenomatous polyps [10, 11], but it is not yet an accepted routine screening examination. Indeed, the concept of disease screening, including its appropriateness and evaluation, is not as straightforward as it may first appear. Even the basic assumption that early treatment will improve prognosis may not be true in all circumstances. Moreover, even if this assumption is justifiable for a particular condition, the risks or costs that are associated with any screening test (and any consequent "induced" procedures) must be weighed against the benefits. Thus, any new application of an imaging procedure to screen for disease should be considered an unproven method of disease control until its risks, benefits, and costs have been rigorously evaluated. Ideally, such evaluations should be completed before widespread use of the procedure for disease screening is undertaken or recommended [12].

Making and evaluating recommendations on the use of imaging studies for disease screening is one of the more difficult problems in medical imaging and clinical medicine. This article will discuss the use of screening tests for detecting early disease or for detecting risk factors for developing disease. Consideration will be given to the appropriateness criteria for two major elements of health screening programs: the condition or disease for which screening is being performed and the screening test itself. Within the context of these two elements, potential biases in the evaluation of screening programs and other critical issues in the evaluation of screening programs will be presented.

Appropriateness Criteria: The Disease or Condition Being Screened

To be appropriate for screening, a disease should be serious, and the preclinical phase of the disease (Appendix 1) should have a high prevalence among the population targeted for screening. Furthermore, screening initiated before a critical point in the natural history of the disease should result in treatment being initiated before the onset of symptoms (Fig. 1). This treatment should be more beneficial in reducing morbidity or mortality than treatment given after symptoms develop. Finally, the

Received March 20, 2002; accepted after revision April 22, 2002.

This is the seventh in a series designed by the American College or Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

[1] All authors: The Russell Morgan Department of Radiology and Radiological Sciences, JHOC Rm. 4155, P. O. Box 0814, Johns Hopkins Medical Institutions, 601 N. Caroline St., Baltimore, MD 21287. Address correspondence to L. L. Fajardo.

AJR 2002;179:825–831

0361–803X/02/1794–825

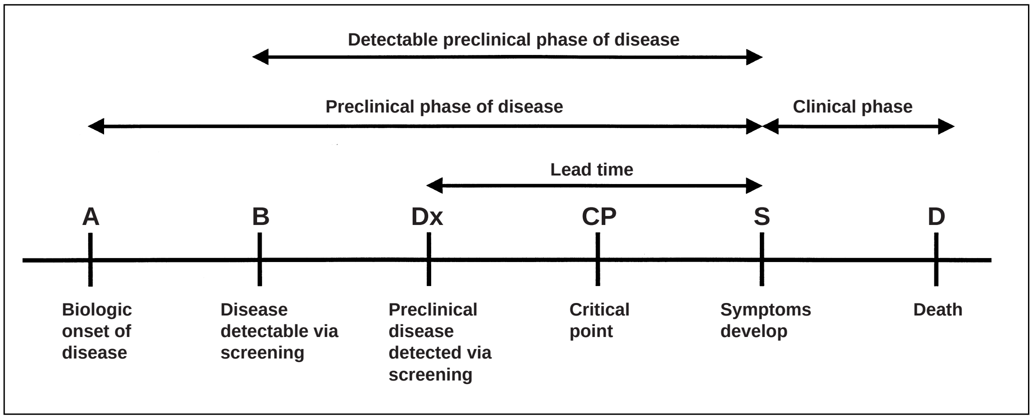

Detectable preclinical phase of disease
Preclinical phase of disease
Clinical phase
Lead time
A
B
Dx
CP
S
D
Biologic onset of disease
Disease detectable via screening
Preclinical disease detected via screening
Critical point
Symptoms develop
Death

Fig. 1.—Diagram shows natural history of disease. Progression from biologic onset of disease to death is divided into preclinical and clinical phases. Detectable preclinical phase of disease is period during which screening tests are applied to detect a condition early in its natural history, before onset of symptoms.

screening for the disease should not result in a significant incidence of pseudodisease.

Substantial Morbidity or Mortality If Untreated

The criterion of seriousness relates primarily to issues of both cost-effectiveness and ethics. The elimination or amelioration of adverse health consequences must justify resource expenditures on radiologic imaging for disease screening. Likewise, the consequences of failing to detect and treat the disease early must be sufficiently grave to ethically warrant exposing individuals to the risks (e.g., radiation exposure or false-positive diagnosis) and discomforts of the screening procedure itself. Life-threatening conditions, such as heart disease and cancer, and those known to have serious and irreversible consequences, such as congenital hypothyroidism and phenylketonuria, clearly meet the criterion of seriousness. On the other hand, medical imaging tests should be thoroughly evaluated for risks and benefits before being used to screen for certain asymptomatic conditions, such as gallstones. Although asymptomatic gallstones are fairly prevalent, rarely are they life-threatening and, in fact, the condition may never become symptomatic.

High Preclinical Prevalence

For a screening test to be effective, it must reveal a sufficient number of preclinical disease cases to justify the testing costs. Thus, the prevalence of preclinical disease must be high in the population for which screening is recommended. Targeting high-risk populations can increase the prevalence of the detectable preclinical phase of the disease and thus the number of cases detected on screening. This strategy will likely be applied to the emerging approaches to lung cancer screening using multidetector CT. Exceptions to the criterion concerning high prevalence of the detectable preclinical disease should be made if screening for rare conditions can be accomplished using tests that are accurate, inexpensive, and noninvasive. Although phenylketonuria occurs in only one of 15,000 neonates, widespread screening is justified by the effectiveness and low cost of the test and by the serious public health consequences of not detecting the disease in its preclinical phase.

Existence of a Critical Point and Appropriate Therapy

Screening tests are only effective if the condition or disease has a critical point (point CP in Fig. 1) so that treatment instituted before the critical point is more efficacious than treatment provided later. In the case of screening for preclinical neoplastic conditions, the critical point coincides with the onset of metastasis [12]. Thus, the critical point must occur during the detectable preclinical phase of the disease because screening is ineffective (and, indeed, unnecessary) after the onset of symptoms (i.e., during the clinical phase of the disease). If the critical point occurs soon after the onset of the detectable preclinical phase, screening may be too late to be useful. Conversely, screening may also be less effective early in the onset of the detectable preclinical phase if lesions are extremely small and are just at the threshold of detectability.

For screening to improve patient outcomes, an effective treatment for the disease must be available. A critical question in evaluating the importance of screening for a condition is whether treatment of the preclinical disease detected on screening is more effective than intervention initiated after the disease becomes symptomatic. Here, the natural history of the disease should be carefully considered. Figure 1 illustrates that the natural history of disease can be divided into preclinical and clinical phases. The preclinical phase is the period from the biologic onset of disease to the onset of clinical manifestations of the disease. During this phase, the condition is asymptomatic but detectable on a screening test. The detectable preclinical phase of disease is defined as the interval between the point at which the disease can be detected on screening (point B in Fig. 1) and the point at which symptoms develop [13] (point S in Fig. 1).

For screening to be beneficial, treatment initiated during the detectable preclinical phase must result in a better prognosis than therapy given after symptoms develop. For example, some subtypes of breast cancer develop for 3–8 years before becoming palpable at routine clinical breast examinations. During this stage, nonpalpable breast carcinomas may be detected on mammography. Many of these carcinomas are confined to the breast and are not associated with lymph node metastasis. Diagnosing and treating breast cancer during the preclinical phase result in a higher percentage of the cases remaining noninvasive (i.e., ductal carcinoma in situ), a lower percentage of cases of axillary lymph node metastasis, and a better 5-year patient survival rate than when breast cancer is diagnosed during the clinical phase [14].

Conversely, if early treatment engenders no difference in the patient's prognosis or health outcome, then the application of a screening test is neither necessary nor effective. For example, screening for lung carcinoma with chest radiography has historically been discouraged because the disease has a poor prognosis regardless of the phase during which treatment is initiated. Similarly, little justification exists in screening for conditions that are completely curable during the clinical phase of their natural history.

Low Incidence of Pseudodisease

A pseudodisease is a disease that does not require treatment because it does not affect patients' length or quality of life in a significant way. Screening for a disease will be ineffective if the screening test reveals substantial pseudodisease. Two sources of pseudodisease have been described [12, 15]. A type I pseudodisease is a condition that is diagnosed via a screening test and does not progress to symptomatic disease; it may even regress over time. This is a recognized phenomenon in screening for breast carcinoma; not all cases of ductal carcinoma in situ progress to invasive or metastatic disease [16, 17]. A type II pseudodisease is an indolent, slowly progressive disease found in conditions with long detectable preclinical phases or among patients with short life expectancies who may die from other causes [12]. This latter type of pseudodisease has been described in prostate carcinoma. Although the prevalence of clinically apparent prostate carcinoma in men aged 60–70 years is only about 1% [18], more than 40% of men in their 60s who have normal findings at rectal examinations have histologic evidence of disease [19] when prostate tissue is removed during cystectomy performed for bladder cancer. Because patients with pseudodisease do not die from the disease for which screening is performed, the survival of these patients is erroneously attributed to early treatment. If adjustments are not made for the detection of pseudodisease in a screening program, an overdiagnosis bias occurs [12]. For both types of pseudodisease, a screening test with positive results may cause the patients to undergo unnecessary tests and therapy. For these reasons, screening for conditions with a high frequency of pseudodisease is not cost-effective.

Appropriateness Criteria: The Screening Test

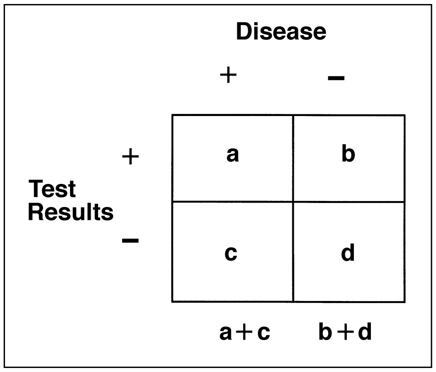
A successful disease-screening program requires not only that the disease have characteristics appropriate for screening but also that a valid screening test be available. Ideally, the test should be widely accessible, simple to administer, inexpensive, and associated with minimal discomfort and morbidity to the population screened. Moreover, the screening test results must be valid and reproducible. Finally, as discussed earlier, the test should be able to reveal the detectable preclinical phase of the disease accurately before the critical point of the disease.

Test Accuracy

A screening test is 100% accurate if it can be used to correctly classify individuals having preclinical disease as test-positive and those without preclinical disease as test-negative. In its simplest form, the assessment of the accuracy of a diagnostic technology involves two dichotomies: disease that is present (+) or absent (–) and test results that are positive (+) or negative (–). A 2 × 2 matrix (Fig. 2) is frequently used to illustrate the four outcome combinations in which n, the total number of test results examined, is expressed by the equation $n = a + b + c + d$. Two of the counts, a and d, correspond to correct test results (true-positive and true-negative, respectively), whereas b is the number of false-positive results and c is the number of false-negative results.

Because the counts for the four outcomes are highly dependent on the sample size, it is customary to express them as rates. For example, a / (a + c) is equal to the proportion of individuals who have the disease and who have positive test results, or the rate of true-positives, also known as the sensitivity of the test; d / (b +d) is equal to the proportion of individuals who do not have the disease and who have negative test results, or the rate of true-negatives, also known as the specificity of the test; c / (a + c) is equal to the proportion of individuals who have the disease but have falsely negative test results, or the rate of false-negatives; and b / (b + d) is equal to the proportion of individuals who do not have the disease but who have falsely positive test results, or the rate of false-positives. Thus, sensitivity is the probability of an individual having positive test results when the disease is truly present, and specificity is the probability of an individual having negative test results when the disease is truly absent.

The usefulness of a screening test is evaluated by its positive and negative predictive values. The predictive value of a negative test (d / [c + d]) is the probability that a patient with a negative result on the diagnostic test truly does not have the disease for which the screening was conducted. Conversely, the predictive value of a positive test (a / [a + b]) is the probability that a patient with a positive result on the screening test truly has the disease for which the screening was conducted. The positive and negative predictive values of a test are dependent on the prevalence of the disease.



		Disease	
		+	−
Test Results	+	a	b
	−	c	d
		a+c	b+d

Fig. 2.—Diagram of 2 × 2 matrix illustrates test outcomes and test accuracy for individuals with and without disease. Disease + = disease present, disease – = disease absent, a = number of true-positive results, b = number of false-positive results, c = number of false-negative results, and d = number of true-negative results.

As the sensitivity of a screening test increases, the number of individuals with preclinical disease not diagnosed by the test decreases. A highly specific test has a low percentage of healthy individuals who are misclassified as having positive test results. Decisions regarding specific criteria for acceptable levels of sensitivity and specificity for a given preclinical disease involve weighing the consequences of leaving cases undetected (false-negatives) against erroneously classifying healthy persons as having the disease (false-positives). In general, sensitivity should be increased at the expense of specificity if the consequences of missing preclinical disease are great, such as when the disease is serious, detectable during its preclinical phase, and curable. Conversely, high specificity is desirable when the costs or risks associated with further diagnostic tests (i.e., surgical biopsy) are substantial. In this circumstance, ethics require that the screened population be informed that a negative result on the screening test does not absolutely guarantee that the disease is not present, only that the likelihood of having the disease is low.

One way to address the problem of the trade-off between the sensitivity and specificity is by administering several screening tests in parallel or sequentially. The former involves performing all the screening tests at the same time and considering individuals with positive results on any of the tests to be true-positive cases. This approach gives greater sensitivity than that achievable by performing each test alone because the condition is less likely to be missed; however, the approach lowers specificity because false-positive diagnoses are also more likely. When screening tests are administered sequentially, an initial screening test is performed, and only those individuals with positive test results undergo an additional screening procedure. Generally, sequential testing results in higher specificity than that achievable with a single test because positive results on a series of tests are more likely to represent a true-positive finding. This method, however, also lowers sensitivity.

Test Reproducibility

Any test being considered for use in a screening program must have reproducible results. For imaging tests, four important sources of variability can affect the reproducibility of results. The first relates to a biologic variation that might affect the performance of the test (i.e., patient size or cardiac motion). The second relates to the reproducibility of the test itself (i.e., patient positioning or film processing in the acquisition and production of mammograms). Third, intraobserver variability refers to differences in the way the same radiologist interprets a specific screening test at different times. Finally, interobserver variability refers to inconsistencies attributable to differences in the way different radiologists interpret the same screening examination. Interobserver variability is minimized if the interpretation criteria and end points are defined and quantifiable and is greater if the criteria are vague and subjective. Both intra- and interobserver variabilities have been reported [20–22] in the interpretation of screening mammograms, description of specific lesions, and recommendations for follow-up examinations, using the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) [23].

A common but flawed approach to measuring the accuracy of a potential screening test is to extrapolate data on tests performed in populations with symptomatic disease to screening populations [13]. However, using an asymptomatic population involves testing many subjects to identify a group with disease and following up those subjects to ascertain the true disease status. Both positive and negative test results in the subjects should be verified by acceptable methods such as histopathology and clinical or imaging follow-up. With respect to the latter, a follow-up period of sufficient length is critical. If the follow-up period is too short, false-negative cases may be missed; if it is too long, new cases of disease (e.g., "interval cancer") may be inaccurately classified as false-negatives.

Test Safety, Availability, and Cost-Effectiveness

Because screening tests are performed on asymptomatic individuals—most of whom are healthy and do not have preclinical disease—the tests must not be associated with significant morbidity or mortality. Even a minor side effect or adverse consequence to the screened population will likely offset the benefits of screening [12]. Radiation dose and the likelihood that the screening test itself may induce malignancies are frequently considered adverse consequences of screening tests involving imaging [24–26]. Other sources of morbidity that affect an individual's decision to undergo or forego screening include the discomfort associated with the test (e.g., compression with screening mammography or bowel preparation for screening barium enema examinations).

The screening test should be accessible to the population for whom it is indicated. Screening cannot be effective if the screening test is available only at large medical centers. Likewise, if the examination is costly, insurers may choose not to provide screening coverage, and patients may be unwilling or unable to pay for the test out of pocket.

Evaluating the Effectiveness of Screening

Evaluations of effectiveness of a screening program should be based on outcomes and measures reflecting the impact of the program on the course of the disease. Here, the critical outcomes of interest are the assurance that the screened and unscreened populations are comparable, the estimates of lead-time and length-time biases, a comparison of cause-specific mortality rates between the screened and unscreened groups, and the measurement of relative and absolute risks.

Comparability of the Screened and Unscreened Groups

In determining the efficacy of a screening test, the screened and unscreened groups must be comparable with regard to all factors affecting the end point under evaluation, with the exception of the screening experience. In this regard, patient recruitment and self-selection bias (volunteer bias) should be taken into account. People who choose to participate in a screening program are likely to differ from those who do not volunteer in several ways that may affect survival [27, 28]. Volunteers tend to have better health and lower mortality rates than the general population and are more likely to adhere to prescribed medical regimes. Consequently, an observational study design comparing mortality rates of screened and unscreened groups is likely to show that those who volunteer to undergo screening have lower mortality rates, regardless of any effect of screening. On the other hand, those who volunteer for screening programs may represent the "worried well," or asymptomatic individuals who are at higher risk of developing disease because of medical or family history or lifestyle factors. Such individuals might have an increased risk of mortality regardless of the efficacy of the screening program. Thus, the direction of potential patient selection bias may be difficult to predict and the magnitude of such events even more difficult to quantify. Randomization schemes are used to overcome self-selection bias in studies evaluating potential screening tests by assigning individuals to screened and unscreened study groups after they agree to participate in the study.

Lead-Time and Length-Time Biases

Showing the benefit of treatment initiated during the preclinical phase of a disease is surprisingly difficult. Two widely recognized problems that arise when the benefits of screening are evaluated by comparing screened to unscreened populations are lead-time bias and length-time bias.

Lead time is the interval between the diagnosis of a disease at screening and the time at which it would have been detected via the onset of clinical symptoms [29]. Lead time, therefore, is the amount of time that the diagnosis was advanced as a result of screening (Figs. 1 and 3). Because screening is applied to asymptomatic individuals, every case of disease detected at screening has had its time of diagnosis advanced. Whether that lead time is a matter of days, months, or years varies by disease, individual, and screening procedure. For a disease that progresses rapidly from the preclinical to the clinical phase, less lead time will be gained from screening than for a disease that develops slowly and has a longer preclinical phase.

Lead time also varies with how soon the screening test is performed after the preclinical disease becomes detectable. For screened patients, cause-specific survival is measured as the length of time from disease detection on the screening test to death from the disease. For patients not screened, cause-specific survival is measured as the length of time from clinical diagnosis to death from the disease. For example, Figure 3 illustrates the hypothetical histories of two women with breast cancer. We assumed that the age of both women at the biologic onset of disease was 35 years and that the disease was detectable on screening when the women were 44 years old. One women (A) was screened at age 47, and her breast cancer was detected at that time. The other woman (B) did not undergo screening mammography; her breast cancer was diagnosed when she was 50 after she discovered a lump in her breast. Both women died at the age of 53. Because woman A survived 3 years longer after detection of breast cancer than woman B, screening appears to

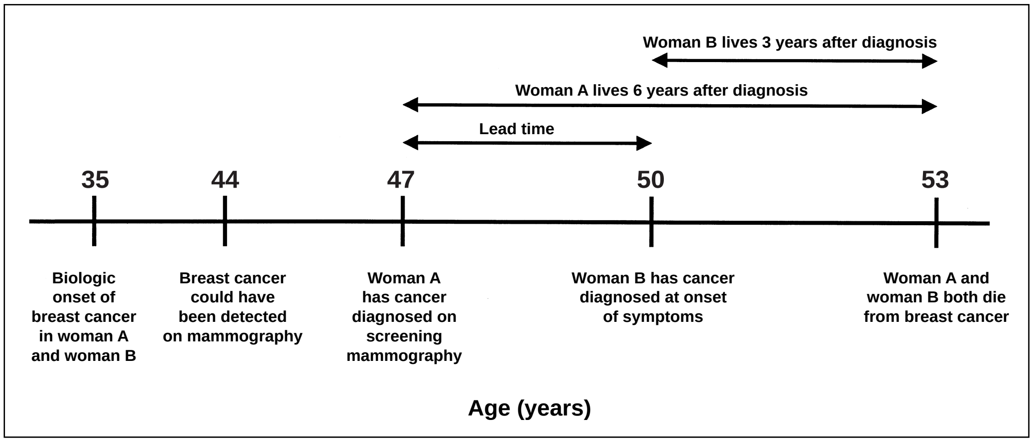

Woman B lives 3 years after diagnosis
Woman A lives 6 years after diagnosis
Lead time
35
44
47
50
53
Biologic onset of breast cancer in woman A and woman B
Breast cancer could have been detected on mammography
Woman A has cancer diagnosed on screening mammography
Woman B has cancer diagnosed at onset of symptoms
Woman A and woman B both die from breast cancer
Age (years)

Fig. 3.—Diagram depicts how lead-time bias can result in apparent increase in survival attributable to screening. Shown are hypothetical case histories of two women with breast cancer. Screening appears to be beneficial when, in fact, it only pushed time of diagnosis forward.

be beneficial when in fact it only pushed the time of diagnosis forward. This phenomenon is commonly referred to as lead-time bias [30–36]. If an estimate of lead time is not taken into account when comparing mortality between screened versus unscreened populations, survival will be erroneously overestimated for the screening-detected cases simply because the diagnosis was made earlier in the natural history of the disease. A second way to account for the effect of lead time on the efficacy of a screening program is to compare the age-specific death rates in the screened and unscreened groups rather than the length of survival from diagnosis to death.

Length-time bias refers to the overrepresentation among screening-detected cases of those diseases with long preclinical phases and thus more favorable prognoses. Diseases with a long preclinical phase are more readily detected on screening tests than are the more rapidly progressing diseases with shorter preclinical phases. An assumption underlying the concept of length-time bias is that diseases with long preclinical phases are more indolent and would have more favorable prognoses, regardless of any effect of the screening program itself. Thus, length-time bias could lead to an erroneous conclusion that screening is beneficial when, in fact, observed differences in mortality rates resulted merely from detection of cases of less rapidly fatal diseases, whereas cases of diseases that are more rapidly fatal were diagnosed after symptoms developed. Length-time bias is difficult to quantify. Its effect is greatest for cases detected at the initial screening; thus, one method of controlling for length-time bias is to compare cases detected at a subsequent screening (i.e., after the initial screening) to those detected clinically (when the patient develops symptoms).

Comparison of Cause-Specific and All-Cause Mortality Rates

The most definitive measure of the efficacy of the screening program is a comparison of the cause-specific mortality rates of those whose disease was diagnosed on screening and those whose diagnosis was made after the development of symptoms. Because the target disease causes only a small proportion of deaths in a screening-eligible population, a statistically precise estimate of differences in mortality rates or a statistically significant effect of screening on all-cause mortality rates can rarely be shown. However, evaluating the all-cause mortality rates may help to ensure that a major harm or benefit is not being missed. An all-cause mortality rate is all-inclusive and provides data relevant to the question of whether other risks are somehow changed along the continuum of the application of the screening test, the diagnosis of a disease, and the treatment. Second, an all-cause mortality rate provides an important perspective on the magnitude of benefit from screening. It puts cause-specific mortality reduction in the context of other competing risks and thus permits an estimate of the overall benefit to be reasonably expected by a particular individual who undergoes a screening evaluation [35]

Absolute Risk Versus Relative Risk

The effectiveness of screening can be expressed in terms of the relative risk, which is the ratio of the cause-specific mortality rate in the study group to that in the control group, or to the relative risk reduction, which is 1 minus this ratio. Although calculations of relative risk are valid, they can be misleading because they convey no information about an individual's baseline risk. The absolute risk reduction is increasingly recognized as a more appropriate measure of effectiveness of screening interventions [37]. Absolute risk reduction is expressed as the product of risk and relative risk reduction. For example, suppose a screening-eligible individual has a 2% probability of dying of a particular disease over the next 20 years. If the relative risk reduction from screening is 50%, the absolute risk reduction is 1%. Reporting absolute risk reduction is especially appropriate for screening because the overall risk to be averted is usually small. The absolute risk reduction puts the potential benefit in proper perspective so that an individual or his or her health care provider can weigh it against the potential side effects and costs. The reciprocal of the absolute risk reduction is the number of individuals who must be screened to prevent one death or adverse event. In our example, this number is 100 or 1/0.01. The perception of the absolute risk reduction from screening may be significantly affected by the detection of a pseudodisease that, as discussed previously, falsely increases the perceived risk of developing the disease and the perceived effectiveness of earlier treatment.

Study Designs for Evaluation of Screening Tests

Many epidemiologic design strategies are used to evaluate the efficacy of screening tests, including correlational studies, observational studies, and randomized trials. Correlational studies are used to examine trends in disease rates relative to screening frequencies in a population or to compare the relationship between frequencies of screening and disease rates for different populations. Such descriptive studies are useful in suggesting a relationship between screening and a decline in the morbidity or mortality rate. However, correlational studies have inherent limitations. First, because information from such studies concerns populations rather than individuals, it is not possible to establish that those experiencing the decreased mortality rate are in fact the same persons who received the screening tests. Moreover, such studies do not allow control of potential confounding factors, such as socioeconomic status. Finally, the measure of screening frequency used is usually an average value for the population, so identifying the optimal screening strategy for an individual is impossible. Thus, correlational studies can suggest the possibility of a benefit from a screening test, but they cannot test that hypothesis.

Observational analytic studies, both case-control and cohort, are also used to evaluate the efficacy of screening programs. In the case-control design, individuals with and without the disease are compared with respect to their prior exposure to the screening test. As with any case-control study, the definition and selection of the cases and controls are of critical importance to the validity of the findings [38, 39]. In a cohort study, the case-fatality rate of those who chose to be screened is compared with the case-fatality rate among those whose diagnoses were made due to the onset of symptoms. Interpretion of the results of cohort studies requires consideration of the potential effects of the self-selection of participants as well as lead-time and length-time biases [40].

Because the chief threat to validity is that screened and unscreened cases cannot be compared, the optimal assessment of the efficacy of a screening program derives from randomized trials. If the sample size is sufficiently large, the process of randomization controls any potential confounding variables. Patient self-selection or volunteer bias, a problem when comparing screened and unscreened groups in observational studies, does not influence the validity of randomized trials: after a group of volunteers agrees to participate in the study, individuals who are to undergo screening are chosen at random from the group by the investigators. Adjusting for the lead-time average can eliminate lead-time bias in comparisons of survival rates of patients whose disease was detected via screening versus those whose disease was detected clinically or, preferably, in comparisons of the age-specific mortality rates for the screened and the unscreened groups. Trials can also address the potential for length-time bias by comparing the mortality experience of the groups after repeated screenings.

In the United States, few randomized trials have evaluated programs that use imaging to screen for preclinical disease. The Health Insurance Plan Breast Cancer Screening Project [41] was a randomized trial conducted to evaluate whether periodic breast cancer screening with mammography and physical examination would result in reduced breast cancer mortality rates among women whose ages ranged from 40 to 64 years old. After 9 years of follow-up, an overall statistically significant reduction in breast cancer mortality was found among women who were offered screening compared with women who were assigned to usual medical care.

Although randomized trials provide the best and most valid data on the efficacy of screening programs, a fair amount of evidence on screening programs has come from nonexperimental study designs. Cost, feasibility, and ethical concerns can make randomized trials controversial. As radiologic screening for disease becomes more common, considerations of new evaluation methodologies to determine costs and benefits may be needed. The challenge for the future is to better identify which screening tests are appropriate for which populations. Emerging quantitative techniques of eliciting patient preferences [42] and of analyzing benefits, harms, and costs over time [43, 44] may help radiology meet this challenge.

References

1. Eddy DM. How to think about screening. In: Eddy DM, ed. *Common screening tests*, 1st ed. Philadelphia: American College of Physicians, **1991**:1–21
2. Eddy DM. Screening for cervical cancer. In: Eddy DM, ed. *Common screening tests*, 1st ed. Philadelphia: American College of Physicians, **1991**:255–285
3. Eddy DM. Screening for breast cancer. In: Eddy DM, ed. *Common screening tests*, 1st ed. Philadelphia: American College of Physicians, **1991**:229–254
4. Garber AM, Sox HC Jr, Littenberg B. Screening asymptomatic adults for cardiac risk factors: the serum cholesterol level. *Ann Intern Med* **1989**; 110:622–639
5. Sox HC Jr, Garber AM, Littenberg B. The resting electrocardiogram as a screening test: a clinical analysis. *Ann Intern Med* **1989**;111:489–502
6. Sox HC Jr, Littenberg B, Garber AM. The role of exercise testing in screening for coronary artery disease. *Ann Intern Med* **1989**;110:456–469
7. Smith RA, Mettlin CJ, Johnston DK, Eyre H. American Cancer Society guidelines for the early detection of cancer. *CA Cancer J Clin* **2000**;50:34–49
8. Henschke CI, McCauley DI, Yankelevitz DF, et al. Early lung cancer action project: overall design and findings from baseline screening. *Lancet* **1999**;354:99–105
9. Kaneko M, Eguchi K, Ohmatsu H, et al. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. *Radiology* **1996**;201:798–802
10. Winawer SJ, Fletcher RH, Miller L, et al. Colorectal cancer screening: clinical guidelines and rationale. *Gastroenterology* **1997**;1123:594–642
11. Frazier AL, Colditz GA, Fuchs CS, Kuntz KM. Cost-effectiveness of screening for colorectal cancer in the general population. *JAMA* **2000**; 284:1954–1961
12. Black WC, Welch HG. Screening for disease. *AJR* **1997**;168:3–11
13. Cole P, Morrison AS. Basic issues in population screening for cancer. *J Natl Cancer Inst* **1980**; 64:1263–1272
14. Bassett LW, Lui TH, Giuliano AE, Gold RH. Prevalence of carcinoma in palpable vs. impalpable, mammographically detected lesions. *AJR* **1991**; 151:21–24
15. Morrison AS. *Screening in chronic disease*. New York: Oxford Univ. Press, **1992**:125–127
16. Page DL, Dupont WD, Rogers LW, Landenberger M. Intraductal carcinoma of the breast: follow-up after biopsy only. *Cancer* **1982**;49:751–758
17. Rosen PP, Braun DW Jr, Kinne DE. The clinical significance of pre-invasive breast carcinoma. *Cancer* **1980**;46:919–925
18. Feldman AR, Kessler L, Myers MH, Naughton MD. The prevalence of cancer: estimates based on the Connecticut Tumor Registry. *N Engl J Med* **1986**;315:1394–1397
19. Montie JE, Wood DP Jr, Pontes E, Boyett JM, Levin HS. Adenocarcinoma of the prostate in cytoprostatectomy specimens removed for bladder cancer. *Cancer* **1989**;63:381–385
20. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. *J Natl Cancer Inst* **1998**;90: 1801–1809
21. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med* **1994**; 331:1493–1499
22. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by U.S. radiologists: findings from a national sample. *Arch Intern Med* **1996**;156:209–213
23. American College of Radiology. *Breast imaging reporting and data system (BI-RADS)*, 3rd ed. Reston, VA: American College of Radiology, **1998**
24. Dixon AK, Dendy P. Spiral CT: how much does radiation dose matter? *Lancet* **1998**;352:1082–1083
25. Faulkner K, Moores BM. Radiation dose and somatic risk from computed tomography. *Acta Radiol* **1987**;28:483–488
26. Mossman KL. Analysis of risk in computerized tomography and other diagnostic radiology procedures. *Comput Radiol* **1982**;6:251–256
27. Greenlick MR, Bailey JW, Wild J, Grover J. Characteristics of men most likely to respond to an invitation to be screened. *Am J Public Health* **1979**;69:1011–1016
28. Wilhelmsen L, Ljungberg S, Wedel H, Werko L. A comparison between participants and non-participants in a primary preventive trial. *J Chronic Dis* **1976**;29:331–339
29. Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little Brown, **1985**:172–176
30. Hutchison GB, Shapiro S. Lead time gained by

diagnostic screening for breast cancer. *J Natl Cancer Inst* **1968**;41:665–681

31. 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**;111:261–267
32. Shapiro S, Goldberg JD, Hutchison GB. Lead time in breast cancer detection and implications for periodicity of screening. *Am J Epidemiol* **1974**;100:357–366
33. Prorok PC. The theory of periodic screening. I. Lead time and proportion detected. *Adv Appl Prob* **1976**;8:127–143
34. Prorok PC. The theory of periodic screening. II. Doubly bounded recurrence times and mean lead time and detection probability estimation. *Adv Appl Prob* **1976**;8:460–476
35. 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
36. Shwartz M. Estimates of lead time and length time bias in a breast cancer screening program. *Cancer* **1980**;46:844–851
37. 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
38. Morrison AS. Case definition in case-control studies of the efficacy of screening. *Am J Epidemiol* **1982**;115:6–8
39. Weiss NS. Control definition in case-control studies of the efficacy of screening and diagnostic testing. *Am J Epidemiol* **1983**;116:457–460
40. 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**; 111:261–267
41. Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer* **1977**;39[suppl 6]:2772–2782
42. Nease RF, Tsai R, Hynes LH, Littenberg B. Automated utility assessment of global health. *Qual Life Res* **1996**;5:175–182
43. De Koning HJ, Ineveld BM, van Oortmarssen GJ, et al. Breast cancer screening and cost effectiveness: policy alternatives, quality of life considerations, and the possible impact of uncertain factors. *Int J Cancer* **1991**;49:531–537
44. Black WC, Welch HG. A Markov model of early diagnosis. *Acad Radiol* **1996**;3[suppl 1]:S10–S12

APPENDIX 1. Screening for Preclinical Disease: Glossary of Terms

Screening—The application of a test to detect a potential disease or condition in an individual who has no known signs or symptoms of that disease or condition.

Preclinical phase of disease—The period of time from the biologic onset of disease to the onset of clinical manifestations of the disease.

Sensitivity—The probability of having a positive test result when the disease is truly present.

Specificity—The probability of having a negative test result when the disease is truly absent.

Lead time—The interval between the diagnosis of a disease at screening and the time it would have been detected via the onset of clinical symptoms.

Length-time bias—The overrepresentation among screening-detected cases of diseases with long preclinical phases and thus more favorable prognoses.

Relative risk—The ratio of the incidence rate of a disease among individuals exposed to a particular risk factor to the incidence rate among unexposed individuals.

Correlational study—A study conducted to examine trends in disease rates relative to screening frequencies in a population or to compare the relationship between frequencies of screening and disease rates for different populations.

Cohort study—A comparative study of two or more groups that differ according to their exposure to a risk factor or other characteristic (such as whether or not they have undergone screening). The groups are then followed up prospectively to assess the incidence of a disease or other outcome hypothesized to be associated with the risk factor or characteristic.

Efficacy—The magnitude of the beneficial effect produced by a specific intervention or procedure under ideal conditions. It is ideally determined by a randomized controlled trial.

Recommended reference for standard epidemiologic terms: Last JM, ed. *Dictionary of epidemiology,* 4th ed. New York: Oxford Univ. Press, **2001**.

Stephen J. Karlik[1]

Exploring and Summarizing Radiologic Data

In this series, we have been learning about the use of statistics to plan, execute, and analyze our research. This module is designed to help define and categorize data into conventional measures for display and analysis. Display, or visualization, of the data is an important concept and one that is at the root of our understanding of various types of data. Before addressing which types of graphs, presentations, or analyses are useful and appropriate, we need to define exactly what type of data to analyze. In our studies, we choose different variables with which to collect the data that can be divided into two primary types by quantity or category. The quantity types are continuous (measuring) and discrete (counting), and the category types are nominal (named) and ordinal (ordered). The following section defines and gives examples of each.

Quantitative Variables

Continuous Data

Continuous data are probably the least frequently reported in the radiology literature because our work has been traditionally one of dichotomous interpretation: either an imaging study successfully reveals an abnormal from a normal finding or it does not. Continuous data are found in which the data of interest exist in a quantifiable range of values that can take any conceivable value in that range. The degree of precision is based on the technology used for its measurement. Some examples are blood pressure (mm Hg), size of a tumor (cm), serum cholesterol (µg/mL), length of an MR imaging sequence (sec), and amount of contrast material (mL). Each of these variables can have a wide range of values whose precision of measurement can vary significantly. Another way to think about continuous data is that a possible value between two other values always exists. An example would be a patient with a systolic blood pressure of 111.5 mm Hg that lies between two other patients with pressures of 111.2 and 111.9 mm Hg.

Discrete Data

A discrete variable is characterized by having only certain values (usually integers). For example, a patient can have only a whole integer representing the number of breast tumors. There are never cases of “2.7 tumors detected on a mammogram” (although a group of patients might have a mean of 2.7 tumors). Another example might be, “The study used eight radiographs for archiving the images for a study.” In the previous example, it seems obvious that we use only a whole radiograph, not 7.5 or 8.5. The distinction may not be so apparent: consider WBC. Because one counts the number of cells per millimeter cubed, the data appear (e.g., 33 cells/mm^3) like a ratio scale (which is discussed in the next section of this article). Because there are never partial cells, the data are defined as discrete.

Comparing Ratio and Interval Scales

Ratio scales of measurement have a constant interval size and a true zero point. If one patient has a 6-cm kidney tumor and a second has a 3-cm tumor, then we can state that the second tumor is half as large as the first. Ratio scales also include capacities (mL), volumes (cm^3), rates (mL/min), weights (kg), and lengths of time (min).

Received June 13, 2002; accepted after revision July 9, 2002.

Series editors: Craig A. Beam, C. Craig Blackmore, Stephen J. Karlik, and Caroline Reinhold.

This is the eighth in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project Coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1] Diagnostic Radiology and Nuclear Medicine, Rm. 2MR21, University of Western Ontario, London Health Sciences Center-University Campus, 339 Windermere Rd., London, Ontario, Canada N6A 5A5. Address correspondence to S. J. Karlik.

AJR 2003;180:47–54

0361–803X/03/1801–47

TABLE 1 Contrast Agent Transit Time for Maximal Renal Enhancement

Patient No.	Transit Time (sec)
1	16
2	27
3	19
4	24
5	28
6	17
7	13
8	8
9	15
10	23
11	16
12	13
13	18
14	13
15	21
16	15
17	15
18	21
19	14
20	21
21	9
22	14
23	18
24	17
25	17
26	16
27	18
28	12
29	13
30	16

TABLE 2 "Stem-and-Leaf" Plot of Contrast Agent Times

Stem	Leaf
0	8, 9
1	2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9
2	1, 1, 1, 3, 3, 4, 7, 8

Interval scale data are those derived from a measurement scale that possesses a uniform interval, but interval scale data have no true zero, as, for example, the centigrade temperature scale (degrees Celsius). Although the difference between 20°C and 25°C is the same as between 5°C and 10°C, 50°C cannot be considered twice as hot as 25°C because the zero point is arbitrary. Actually, the temperature scale of kelvin is a ratio scale, because the zero value is real at absolute zero.

Categoric Variables

Nominal Data

Nominal variables often describe characteristics, such as male and female, and are commonly used in radiologic studies. Nominal scales name the values of the nominal variable. For example, a breast tumor type could be classified as benign, malignant, or containing calcifications.

Ordinal Data

This type of data deals with comparisons that are relative, rather than quantitative. Thus, the data consist of an ordering or a ranking of measurements. When one orders the finding, then the scale becomes ordinal, even if the steps in the order are different. An example is the Kurtzke [1] expanded disability scale (0–10) for the neurologic assessment of patients with multiple sclerosis. In this widely used scale, a worsening in the patient status of one unit from 1 (minor signs) to 2 (elevated thresholds) is dramatically different from 6 (walks with assistance) to 7 (wheelchair bound). A common form used in radiology is to classify image interpretability as poor, moderate, or excellent and perhaps grade as 1, 2, and 3.

It is also possible to have exactly the same original data portrayed in several different data types. Using an example of examination marks, we can have raw marks of 97, 75, 68, and 51 (discrete data) that can be expressed as the grades A, B, C, and D (ordinal data) or pass, pass, pass, and fail (nominal data). Although this latter example appears trivial, this exact type of data reduction is common in radiology, in which a complex data set is reduced to presence or absence to facilitate the common 2 × 2 chi-square analysis of diagnostic accuracy. The problem with data reduction is that it can result in a loss of information.

Plotting Methods

Let us take the different types of measurement in turn and examine exploring, summarizing, and presenting each type.

Continuous data have no discrete divisions between elements apart from those imposed by our measuring technique. Some examples are time, the size of a tumor, and blood pressure. Table 1 lists the time taken for a bolus injection of radiographic contrast material to reach a maximum in the kidney with a range of 8–28 sec for 30 patients. These data are raw in the sense that they are unadulterated, unmodified, and untransformed. Time is a continuous measurement, which can take any value whatsoever, but the precision of its measurement is dependent on our measurement tool (wall clock vs stop watch, accurate to a millisecond). By the established rules of science, a reported time of 21 sec is actually all times from 20.5 sec up to and including 21.4 sec. The next section illustrates a variety of ways of exploring these data.

A preliminary and easy way to look at continuous (raw) data is to use the "stem-and-leaf" plot. Although likely unfamiliar to the radiologist, it is easy to construct without computerized graphing packages and shows the distribution of the data in a rudimentary way. The common "stem" is along the left for each decade (0 for units = 0–9, 1 for teens = 10–19, and 2 for twenties = 20–29), and the different values are sorted by increasing values in the second column (Table 2). Most values are in the decade from 10–19, and there are no values exceeding 28 sec. This plot style would clearly identify a highly unusual value (84 sec) from a large number of points—for example, if a value of 8 in the stem and 4 in the leaf were seen. Although a stem-and-leaf plot can allow an easy appreciation of a data set, the details of the distribution are missing.

To obtain a more detailed examination of our example of enhancement time, we created a dot plot that shows the frequency of occurrence of any individual data values (Fig. 1). In a way analogous to the stem-and-leaf plot, the dot plot uses a stem value for each unique time value in the whole set. Then, a single dot is plotted for each occurrence of that value in the data set—for our example, one dot for 8 or 9 sec and 4 dots for 16 sec. Although possibly also unfamiliar, this method is another way to picture the raw data and is analogous to a histogram for each time point. In this type of plot, each data point simultaneously shows the actual value, occupies space, and represents one counting unit. Compared with the stem-and-leaf visual, the dot plot permits a more detailed appreciation of the variability in the data and is close to a histogram (albeit one that has been stood on its end).

Time (sec)	No. of Occurrences
8	*
9	*
12	*
13	***
14	**
15	***
16	****
17	***
18	***
19	*
21	***
23	**
24	*
27	*
28	*

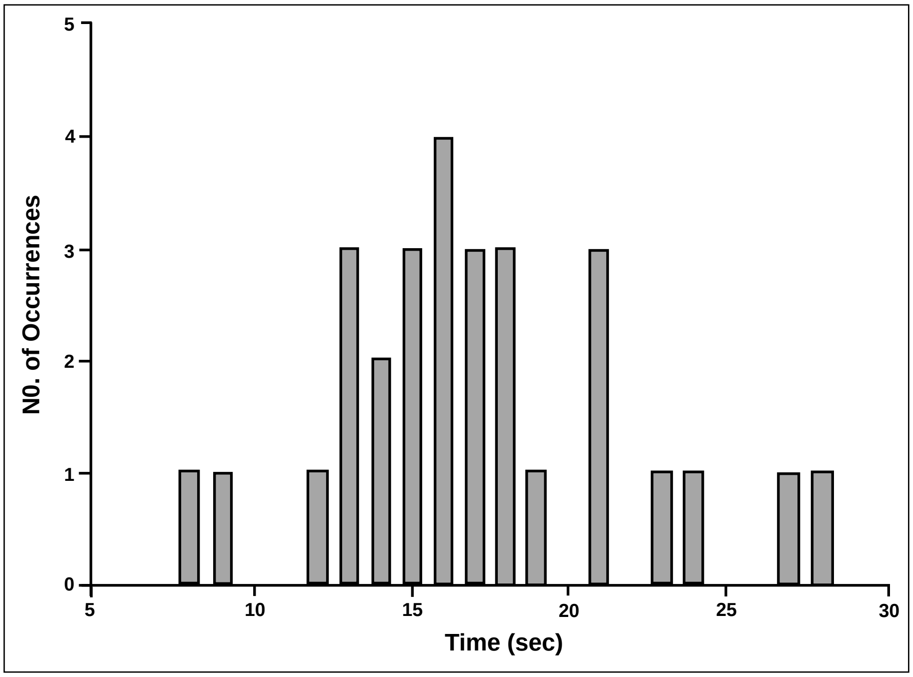


Fig. 1.—Dot plot of transit time data (found in Table 1) shows each asterisk as representing actual occurrence of specific time (sec) beside it.

Fig. 2.—Conventional frequency histogram shows all raw data for transit time (found in Table 1).

The data set that is organized as a conventional histogram shows the frequency of each data value as a bar (Fig. 2). When the data are scattered or the data intervals are too numerous, it is customary to reduce the number of intervals, remembering that there should be enough intervals or bins to show any relevant pattern. Because the data in Table 1 consist of 30 values, one published rule is to use approximately $\sqrt{n}$ (square root) intervals, where n is the total number of values [2]. With an n value of 30 and a $\sqrt{n}$ value of 5.5, five or six intervals are appropriate. If we choose six intervals, then the resulting histogram shows a maximum in the 15- to 17-sec interval (Fig. 3). Although this reduction of the data by decreasing the number of intervals loses some of the details of the exact measurements seen in Figure 2, the essential character of the data is illustrated, in that the maximal enhancement time values are identified in the 12- to 21-sec area in intervals containing 12–14, 15–17, and 18–21 sec. If the transit time variability is important, then you might prefer to

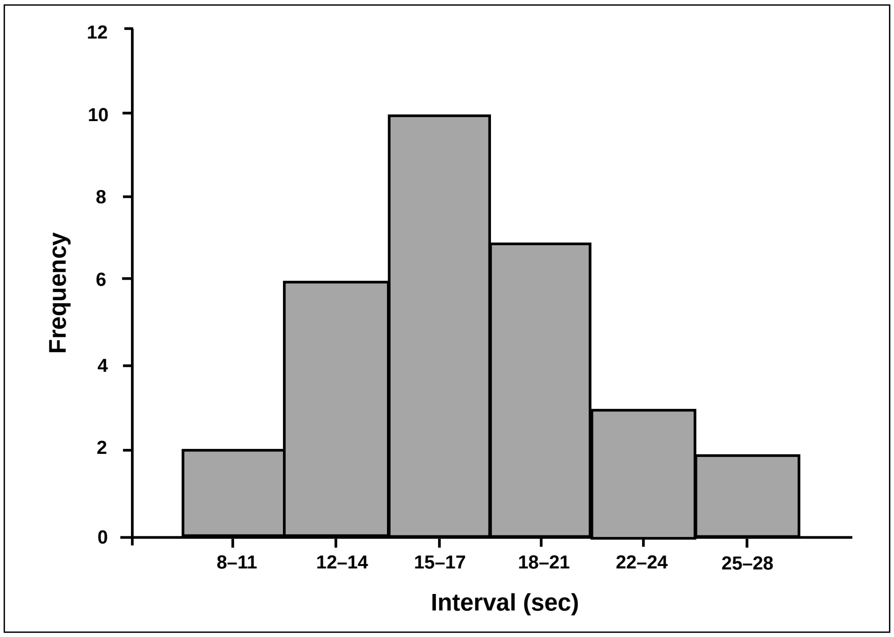


Fig. 3.—Frequency histogram (for data found in Table 1) shows that number of bins has been decreased to $\sqrt{n}$ (square root).

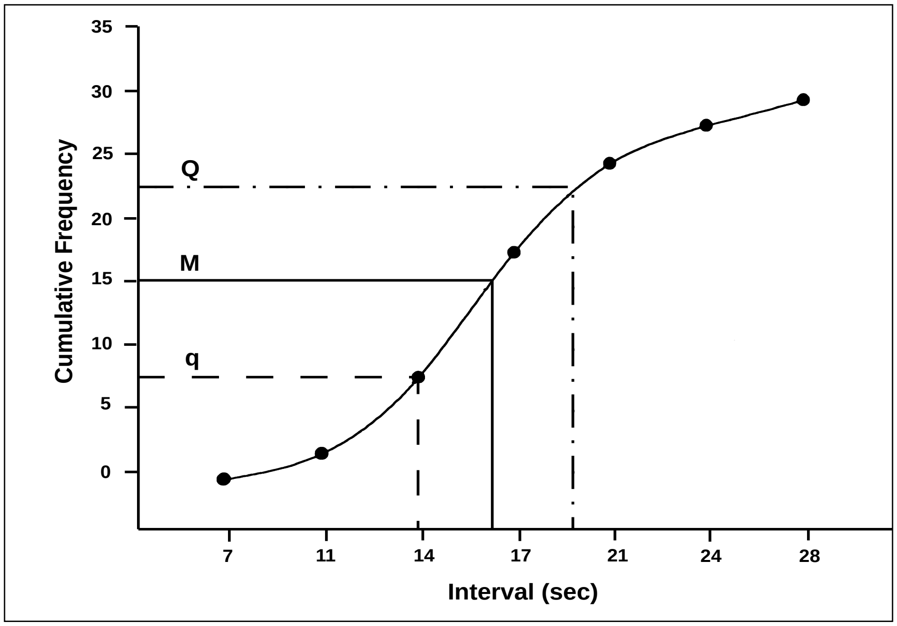


Fig. 4.—Graph shows distribution of enhancement data converted to cumulative data (see Table 2). Conversion from histogram format permits easy visualization of quartiles; Q = third quartile, M = median, q = first quartile.

TABLE 3 Cumulative Frequency of Contrast Agent Transit Times

Frequencies	Interval (sec)					
	8–11	12–14	15–17	18–21	22–24	25–28
Frequency	2	6	10	7	3	2
Cumulative frequency	2	8	18	25	28	30

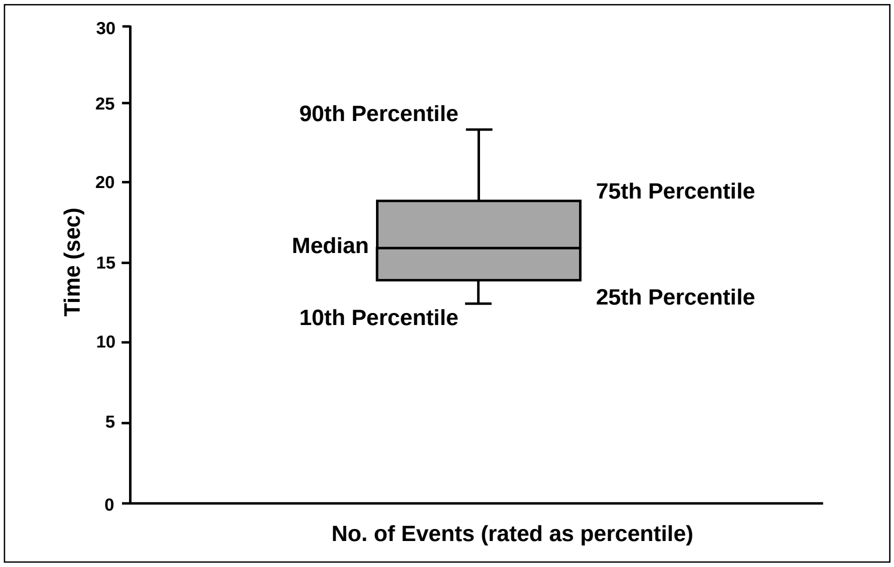


Fig. 5.—Box-and-whiskers plot (for interval data found in Table 1) shows median and percentiles as marked. Compare this graph with Figure 4 that expresses the same data with quartiles.

choose Figure 2. Conversely, if showing the typical time were your goal (say to choose an optimal imaging time), then the expression of data in Figure 3 would be appropriate. Neither choice is artificial; each emphasizes a different aspect of the data.

Having plotted our data and appreciated its distribution, we must determine three primary attributes: the center, the dispersion, and the symmetry of the data distribution.

Measurements of Central Tendency

The central tendency is the tendency of the observations to accumulate at a particular value or in a particular category. The three ways of describing this phenomenon are mean, median, and mode.

The most widely used measure of central tendency is the familiar mean, in which the calculation of the mean is simply adding all values in the data set and dividing the sum by the number of samples. This procedure yields a mean value of 17.2 sec for our time data. The mean is only applicable to ratio or interval scale data.

Another way to look at this data is to make a cumulative frequency diagram. We first convert the frequency histogram (Fig. 3) to a cumulative frequency table (Table 3) and then plot Table 3 as the final cumulative frequency diagram (Fig. 4). The conversion is started by listing the number of occurrences for each interval under the interval values in Table 3. Then we calculate the cumulative frequency for each interval as the total frequency of that interval, plus the frequency of all lower intervals. For example, the cumulative frequency for the interval 18–21 sec is the actual frequency (seven occurrences) plus the total frequency in all smaller intervals (*n* = 18) to yield 25. It is also possible to convert the raw frequency histogram to the cumulative frequency diagram in an entirely analogous way using the individual data values rather than the intervals.

The cumulative frequency diagram provides the investigator with an opportunity to visualize three important measures of the data: the first quartile (q), the median value (M, or the second quartile), and the third quartile (Q). The median is the middle value from the data set. Because there is an even number of observations (*n* = 30), we take the 15th and 16th values from a list of the data with increasing values (16 and 17 sec here) and take the average, which is 16.5 sec. The median divides the data into two equal parts (by the number of observations); the quartiles divide each of these halves into two or four parts total. The values of q, M, and Q can help to show whether the data are symmetric in the interquartile range, which happens if the M–q and Q–M ranges are approximately equal. This determination of interquartile ranges is our first introduction to measures that characterize the dispersion or spread of the observed data.

Imagine that the histogram illustrated in Figure 3 could be physically weighed instead of occupying some space in a plot. The mean can be conceptually thought of as dividing the histogram into two equal parts by weight, whereas the median is simply the middle measurement in the data set. The median also expresses less information than the mean because the median is based on the rank of the individual data values (not the actual values). When the data set has many values that are low or high compared with the average, the median is less sensitive to these values and may be a preferential way to describe the central tendency. Thus, the median is insensitive to the data extremes. In our example, this insensitivity could happen if we exchanged the highest value in our set (28 sec) with a larger data point (100 sec). Although the median value would remain the same (16.5 sec), the new mean is 19.6 sec. Thus, the median retains its ability to identify a value more consistent with the spirit of the data compared with the mean, which has been increased by the extreme value.

In addition to the quartile divisions mentioned previously, the distribution can also be divided into other parts, such as percentiles (or 100 parts). A representative example of this division is the use of lethal dose 50 (LD50) from pharmacologic studies. The LD50 is actually the dose at which 50% of the experimental animals died, or the 50th percentile of lethal doses, or the median lethal dose. Similarly, q (first quartile) is the 25th percentile and Q (third quartile) is the 75th percentile.

A useful way to depict this type of data is the box-and-whiskers plot (Fig. 5), which is effective in summarizing the properties of a data set. The bottom and top of the box are the 25th and 75th percentiles (which are q and Q in Fig. 4), the line in the box is the median value (M), and the "whiskers" (looking like error bars) extend to the 10th and 90th percentiles.

The mode is another term used to describe the central tendency of a data set. The mode is defined as the most frequently occurring measurement, which is 16 sec for our enhancement data. It is possible that the data set has more than one mode. Hence, it is possible to see the descriptor "bimodal" for a distribution of data having two modes or two peaks on a plot of the data.

TABLE 4 **Comparison of Vessels Revealed on Digital Subtraction Angiography (DSA) and MR Imaging Techniques**

Patient No.	DSA	Contrast-Enhanced Time-of-Flight Angiography	3D Time-of-Flight Angiography	Dynamic MR Angiography
1	++	+++	+	+++
2	(+)	++	++	++
3	+++	+++	+++	+++
4	++	++	(+)	++
5	+++	++	+++	+++

Note.—3D = three-dimensional, (+) = partial visibility, + = fair visibility, ++ = moderate visibility, +++ = excellent visibility.

Measurements of Dispersion

As seen in Figure 2, our enhancement maxima do not all occur at the same time and are spread over a substantial range (8–28 sec). We can exactly express this dispersion or nonuniformity in the data. The most commonly used measure of dispersion for a single sample of continuous data is the SD, and, like the mean, the SD takes all the data into account. The SD is a statistical measure that expresses the average amount by which all data values in the set deviate from the mean value: the smaller the differences, the smaller the deviations, and the smaller the SD (and vice versa). For our data set, the mean is 17.2 sec with an SD of 4.7 sec.

If we can assume that the data we collect is normally distributed, the SD has some useful interpretations. For example, 68% of all observations will lie within ± 1 SD of the mean value. Ninety-five percent of the data lies within ± 2 SDs, and 99.7% lies within ± 3 SDs of the mean. Hence, the SD is approximately one sixth of the total data range for a normal distribution.

The mean and SD of a normally distributed data set tell us about the internal structure or internal proportions. Another term that is often seen is the standard error. The SD of the means of many samples from the same population is called the standard error. The standard error depends on the sample SD, the number of samples, and the proportion of the population in the sample. These three statistical measures—mean, SD, and standard error—are used to determine whether two experimentally determined samples are from different populations. When we compare samples, we are applying a test of significance. "Statistically significant" may not equate to "interesting" or "important."

Ordinal Data

Tables are effective for the presentation of ordinal data. Table 4 illustrates an example of the reporting of vessel conspicuity for different visualization techniques: digital subtraction angiography, contrast-enhanced time-of-flight MR angiography, three-dimensional time-of-flight MR angiography, and dynamic MR angiography. The ordinal scale is partial visibility to excellent visibility in four steps represented in the table by "+" to "+++" in an intuitively obvious way.

Proportions and Rates

Proportions and rates are descriptive parameters for a population that can be estimated from a sample. Rate is the occurrence of a particular event in a sample and is given as a percentage. Table 5 shows an example in which the number of events (and rate as a percentile) is listed for four possible categories of neurologic outcome resulting from carotid artery stenting. "Proportion" is a de-

TABLE 5 **Complications Associated with Stenting of the Carotid Artery**

Neurologic Outcomes	Hemisphere	
	Ipsilateral (*n* = 156)	Contralateral (*n* = 88)
Major stroke	2 (1.2)	2 (2.2)
Minor stroke	18 (11.5)	13 (14.8)
Neurologic death	1 (0.6)	0 (0)
Nonneurologic death	3 (1.9)	5 (5.7)
Total events (%)	24 (15)	20 (23)

Note.—Data are numbers (%) of patients.

scriptor that is applicable to categoric data. A stacked bar chart permits visualization of the proportions of three measures in three different patient groups (Fig. 6). In radiology, we frequently use a common statistical test (chi-square) to determine whether the rate or proportion of observations is different in two or more populations.

Relationship Between Two Variables

At times, we take two simultaneous measurements of our study population for the purpose of determining whether a relationship exists. In some instances, the measurements are taken to establish a pattern in the data (e.g., body weight and X-ray attenuation) or to search for an easy-to-measure surrogate marker for a hard-to-measure value (e.g., to measure the amount of iodinated contrast agent in a solution using its optical absorbance).

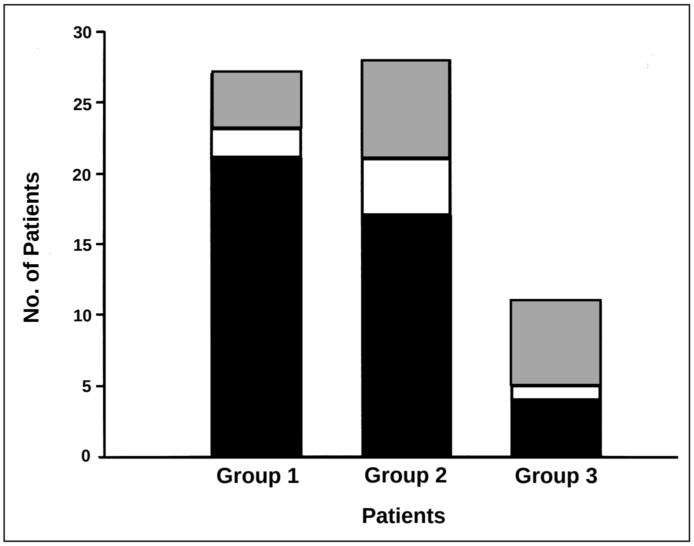


Fig. 6.—Bar chart shows proportion of patients in three treatment groups who were found with no change in size of prostate (*black bar*), enlargement (*white bar*), or decrease in size of prostate (*gray bar*). Note proportion of patients in each classification in each of three differently sized groups.

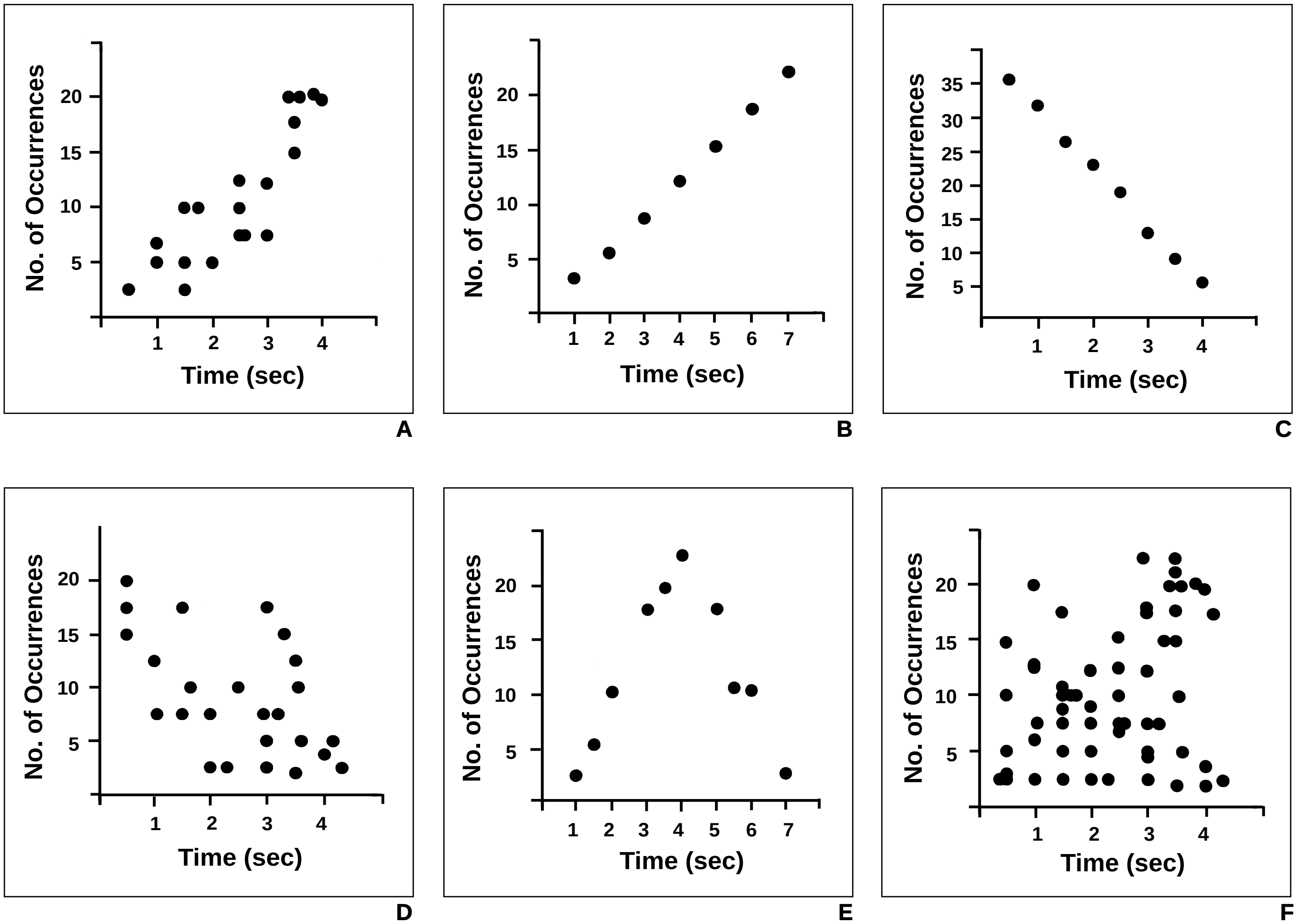


Fig. 7.—Scatterplots for six data sets show different data distributions.
A–E, Pearson's product moment correlation coefficients for data sets are as follows: 0.864 (**A**), 0.991 (**B**), –0.992 (**C**), –0.549 (**D**), 0.078 (**E**), and 0.247 (**F**).

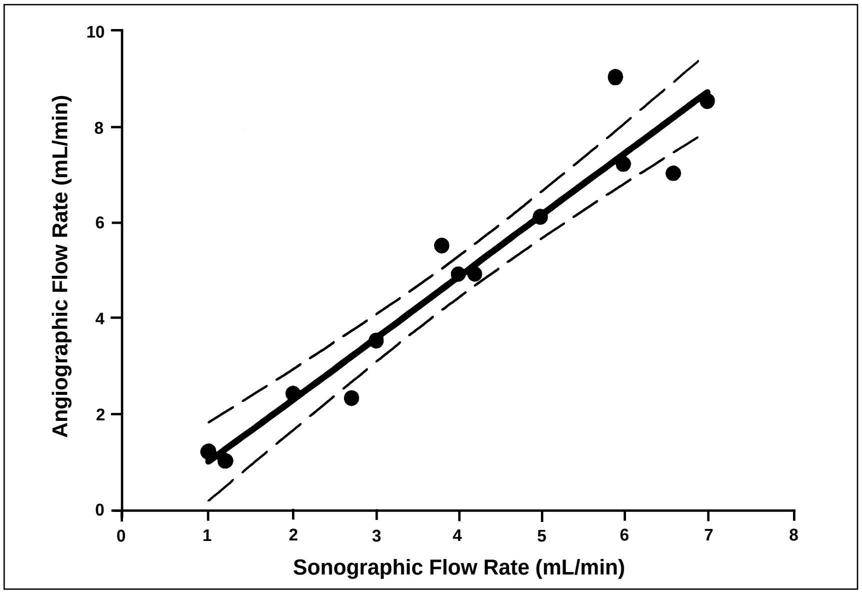


Fig. 8.—Graph shows hypothetic data set (●) with linear regression (*solid line*) and 95% confidence intervals (*dashed lines*) plotted. Note that confidence intervals permit appreciation of strength of regression. r^2 = 0.927, slope (m = 1.28), and *x*-intercept = –0.286.

A scatterplot is the first step in examining the relationship between two sets of measures. The correlation coefficient (*r*) measures how close the relationship between two measurements is to linearity. The maximal values for *r* are 1 or –1, and the two variables can be positively or negatively correlated. If the two variables show a nonlinear relationship (e.g., parabolic), then *r* equals zero, even though a strong relationship exists. The two calculations for correlation coefficient are Pearson's product moment correlation for normal data, and for ordinal data, Spearman's rank correlation.

When a correlation coefficient is used, three steps should be adhered to: first, plot the raw data in a scatterplot; second, observe whether a relationship exists between the variables; and third, if the data suggest a linear, but not a curvilinear relationship, then calculate *r*. The problem with correlation calculations is that correlation can be confused with causality, and caution should be used about such an interpretation. The possibility of an indirect relationship, via a third and unmeasured variable, should be eliminated. It is up to the scientist to prove that these third variables have no effect on the observed correlation. Another caution is that Pearson's correlation coefficient is only dependable when the two compared variables are normally distributed because an outlier point can dominate the correlation.

In interpreting the strength of a correlation coefficient, we found no common consensus on the scale descriptors. A useful published example of descriptors might be: 0.0–0.2, very weak or negligible; 0.2–0.4, weak or low; 0.4–0.7, moderate; 0.7–0.9, strong, high, or marked; 0.9–1.0, very strong or very high [3].

Plotting data sets in scatterplots (Fig. 7) permits us to visually evaluate the data, and we can predict the outcome of an analysis of the correlation coefficients. The data in Figure 7A would have a good correlation, which is supported by a Pearson's test yielding an *r* value of 0.864 (strong correlation). Figures 7B and 7C are obviously linear and have an *r* value of 0.99 and an *r* value of –0.99 (very strong correlation). Figure 7D is somewhat ambiguous. However, *r* is equal to –0.549 and thus a moderate correlation exists. The data in Figure 7E are clearly related, but because the relationship is nonlinear, *r* is equal to 0.078. Even Figure 7F has a higher correlation coefficient, an *r* value of 0.247, than Figure 7E. A look at the correlation values alone for these data sets would suggest that the data in Figure 7E had no relationship, whereas the data have an interesting one that is immediately visible in the scatterplot.

When the scatterplot of the data for two variables looks like a linear relationship exists, then it is tempting to try and describe the relationship as linear and calculate the relationship between them using linear regression. This approach compares a dependent variable (*y*) in relation to an independent variable (*x*), which yields the familiar $y = mx + b$, where *m* is the slope of line and *b* is the *y*-intercept (when $x = 0$). Our hypothetic example shows the plot of raw data, regression line, and 95% confidence limits (Fig. 8). The difference between correlation and regression is that in a correlation, neither variable can be fixed, whereas in regression, one measurement is a variable (*y*) and depends on the other (*x*). Often, the value of *x* is assumed to be fixed, is capable of observation without error, and is normally distributed. Should there be no logical argument to define one variable as dependent and the other as independent, then the solution is to use a calculation of correlation and avoid the concept of dependence altogether. The importance of confidence limits should not be underestimated, either here with regression [4], or elsewhere with statements of sensitivity and specificity [5] or proportions and rates. For example, if we claim no side effects from contrast injections in 20 patients (rate = 0%), the upper 95% confidence limit of the rate of occurrence is actually 19%.

TABLE 6 Sample Contingency Table

Diagnostic Test Result	Target Disorder		
	Present	Absent	Total
Positive	653 (A)	127 (B)	780 (A + B)
Negative	77 (C)	1400 (D)	1477 (C + D)
Total	730 (A + C)	1527 (B + D)	2257 (A + B + C + D)

Note.—Sample contingency table summarizes the number of patients with and without a target disorder that is positive or negative on a single diagnostic test.

Sensitivity and Specificity

Sensitivity and specificity are ratios fundamental to the radiology discipline. They relate the ability of an imaging technique to reveal disease when present (sensitivity) and to rule out disease when absent (specificity). The numbers are generated using the familiar 2 × 2 table, which we have seen previously in this series [6], for proportions used to compare diagnostic determination (presence or absence of disease) with a standard of reference. The better the latter (e.g., surgical confirmation), the more valuable and accurate the diagnostic measurement will be. Although the analysis of a 2 × 2 contingency table has been shown previously in this series of articles, we will use the example in Table 6 to calculate these values. Sensitivity is $a(a + c\ )$, which is equal to 653/730 or 89%; specificity is $d\ /\ (b + d)$, which is equal to 1400/1537 or 92%. Missing from most reports in the radiology literature is the confidence interval based on the binomial theorem [7]. There are a few key questions to consider when evaluating sensitivity and specificity values: Was there an independent and blind comparison with the standard of reference? Was the diagnostic test evaluated in a group of patients appropriate to the target population? Was the standard of reference applied regardless of the diagnostic test [7]? Both negative and positive predictive values can also be calculated from the 2 × 2 table, as well as prevalence, pre- and posttest odds, likelihood ratios, and posttest probability. Usually, these statistical measurements are portrayed in simple tables or in the text of an article. It is useful to show all 2 × 2 contingency tables because it is then possible for the reader to calculate all these values. Even when the 2 × 2 is expanded into a receiver operating characteristic analysis (to be described later in the series), the relevant measure (usually area under the curve) can be expressed in table format with the appropriate confidence intervals.

Summary

The purpose of this article was to define the different variables that radiologists routinely use to describe their data. Categoric and continuous data types were identified, and suitable graphs and tables were shown to depict the findings in an informative and succinct manner. Continuous data and measures of central tendency and dispersion were shown. The relationship between two variables was determined by the correlation coefficient with a consideration of the caveat that correlation should not be confused with causality. The familiar 2 × 2 contingency table and derived values were explored. Identifying variable types and choosing their appropriate displays should be a more straightforward task after studying these examples.

References

1. Kurtzke JF. On the evaluation of disability in multiple sclerosis. *Neurology* **1998**;50:1961–1970
2. Clarke GM. *Statistics and experimental design*. London: Edward Arnold, **1994**:7
3. Rowntree D. *Statistics without tears*. London: Penguin, **1991**:170
4. Glanz SA. *Primer of biostatistics*. New York: McGraw-Hill, **1992**:211
5. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. *BMJ* **1999**;318:1322–1323
6. Jarvik JG. The research framework. *AJR* **2001**; 176:873–878
7. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine*. New York: Churchill Livingston, **1997**:118–128

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series: Introduction, which appeared in the February 2001 issue; Framework, April 2001; Protocol, June 2001; Data Collection, October 2001; Population and Sample, November 2001; Statistically Engineering the Study for Success, July 2002; and Screening for Preclinical Disease: Test and Disease Characteristics, October 2002.

Fundamentals of Clinical Research for Radiologists

Stephen J. Karlik[1]

Visualizing Radiologic Data

It should come as no surprise to radiologists, who earn their living by analysis of visual information, that the analysis and presentation of scientific data should also have a significant visual component. Not only does the visual presentation enhance the clarity of the data, whether for presentation or for publication, but in fundamental ways it also assists in our understanding of it. In fact, modern data graphics should be considered instruments for reasoning about quantitative information. Sometimes the best way to understand, describe, or summarize numeric data is to look at a picture of it. In consideration of statistical graphics, one publication rises above all others: *The Visual Display of Quantitative Information,* by E. R. Tufte [1]. Far from being a cold and clinical tome, as the title suggests, this is an entertaining and approachable work from which we can take important concepts and apply them to the expression of radiologic data.

The display of data in graphs, charts, and diagrams has a specific aim: to discover any patterns in the data. Although this data display is particularly useful in continuous data (described in "Exploring and Summarizing Radiologic Data" in this series [2]), this article will use examples to illustrate those instances in which other graphic styles can help us conceptualize the phenomena underlying our observations. When data are prepared for publication, it is customary to choose the most relevant and smallest number of illustrations or radiographic images to describe the findings. Statistical graphs can assist in this process by revealing and concentrating large amounts of data into a manageable size for portrayal. Graphic excellence has certain properties: clarity, precision, efficiency, consideration of several variables simultaneously, and honesty in revealing the data [1].

Graphic Integrity

This article is an examination of a series of figures from the recent radiology literature, with careful attention to the basis of graphic integrity as outlined by Tufte [1]. Graphic integrity includes using the physical size of numbers or symbols in proportion to the actual values; showing data variation—not design variation; using clear and unambiguous labeling; not quoting data out of context; and avoiding having the number of graphic dimensions exceed the number of dimensions in the data [1]. Other key definitions and concepts brought up by Tufte are illustrated in the figures. Because these figures are reprinted to illustrate points about graphic design, none was changed to conform to the *American Journal of Roentgenology* style for figures.

One fundamental concept in judging graphic competence is that of "data ink," in which the data–ink ratio equals the ink used for data (data ink) divided by the total ink used in the graph [1]. Therefore, background grids, three-dimensional pictures, shading, and hyperactive bar fills are unproductive ink, diluting the data–ink ratio. For clarity, then, nondata ink should be erased. The overall principles to optimize the data–ink ratio include showing the data, maximizing the data–ink ratio, erasing nondata ink, and erasing redundant data ink [1]. Some of the radiologic examples in this article pertain to the issue of data ink.

Furthermore, there are annoying charts and graphs that substitute graphic variation for data variation. One type of colorfully named "chartjunk" [1] is the moiré optical effect

Received June 13, 2002; accepted after revision July 9, 2002.

Series editors: Craig A. Beam, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the ninth in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

[1] Department of Diagnostic Radiology and Nuclear Medicine, Rm. 2MR21, University of Western Ontario, London Health Sciences Center–University Campus, 339 Windermere Rd., London, Ontario N6A 5A5, Canada. Address correspondence to S. J. Karlik.

AJR 2003;180:607–619

0361–803X/03/1803–607

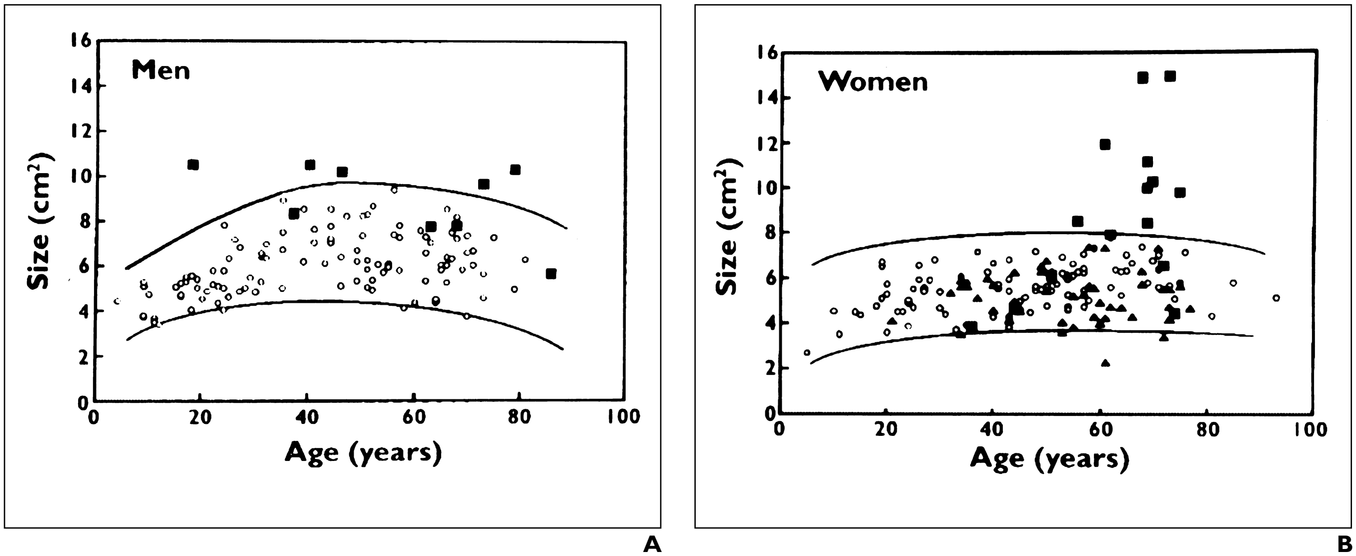


Fig. 1.—Example of presentation with high data–ink ratio. (Reprinted from [3])
A and **B**, Multivariable graphs show control MR imaging–determined parotid gland size (○) for male (**A**) and female (**B**) patients. Each patient data point represents parotid gland size, age, and patient condition. Parotid gland size increased in patients with hyperlipidemia (■) but not Sjögren's syndrome (▲). Mean values ± two standard deviations are plotted (containing 95% of data) to provide visualization of spread of control data versus patient values.

caused by closely spaced lines. You have seen this effect on the television screen (particularly in striped clothing), and now, with the promulgation of computerized graphing programs, it is becoming more common in research reports. Although background grids can assist in the reading of a complex data set, de-enhancing the grid to a lighter shade of gray may help to minimize the optical assault. Most of the data ink should be devoted to data variation. Following this premise enhances the efficiency of communication. In the design of statistical graphs, the ability to portray complexity, structure, and density of data should always be considered.

The Good, the Bad, and the Ugly

This article reviews 21 figures taken from the recent radiologic literature to examine how these graphic presentation issues have been dealt with. Figures 1 and 2 are examples of data-intense multivariable graphs in which a substantial amount of data is concentrated in a format that permits visualization of data variation in patients (Fig. 1) and temporal relationships (Fig. 2). Figure 1 has a high data–ink ratio and allows the reader to easily comprehend the control-versus-patient differences in the MR imaging determination of parotid gland size. Figure 2, although containing a background grid that dilutes the data–ink ratio somewhat, is effective in coordinating the temporal events associated with contrast enhancement.

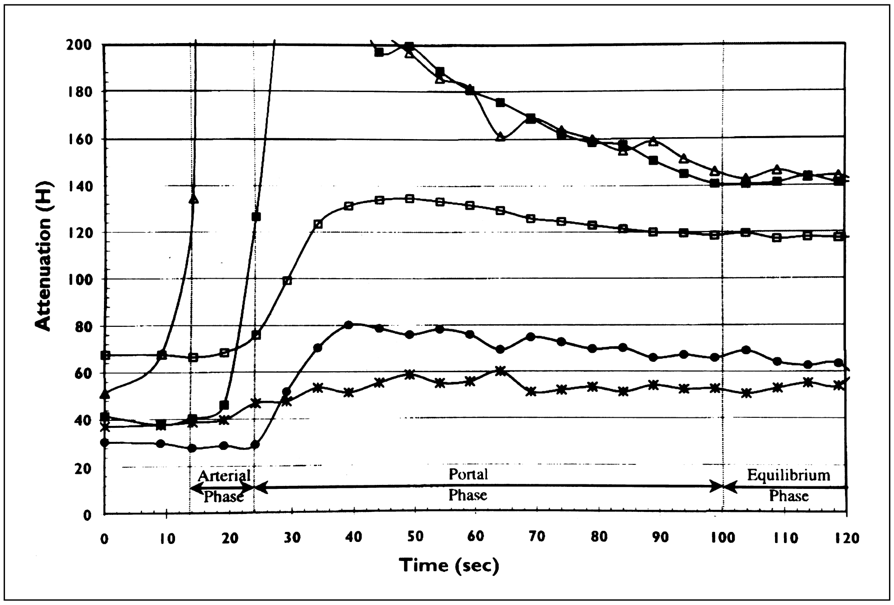


Fig. 2.—Example of graph with high data–ink ratio that portrays related data in one presentation. Multivariable graph depicts attenuation versus time for several tissues after contrast injection. Conspicuity (●) and attenuation of liver (□), tumor (✳), aorta (▲), and portal vein (■) are plotted. Phases of hepatic enhancement are also illustrated. (Reprinted from [4])

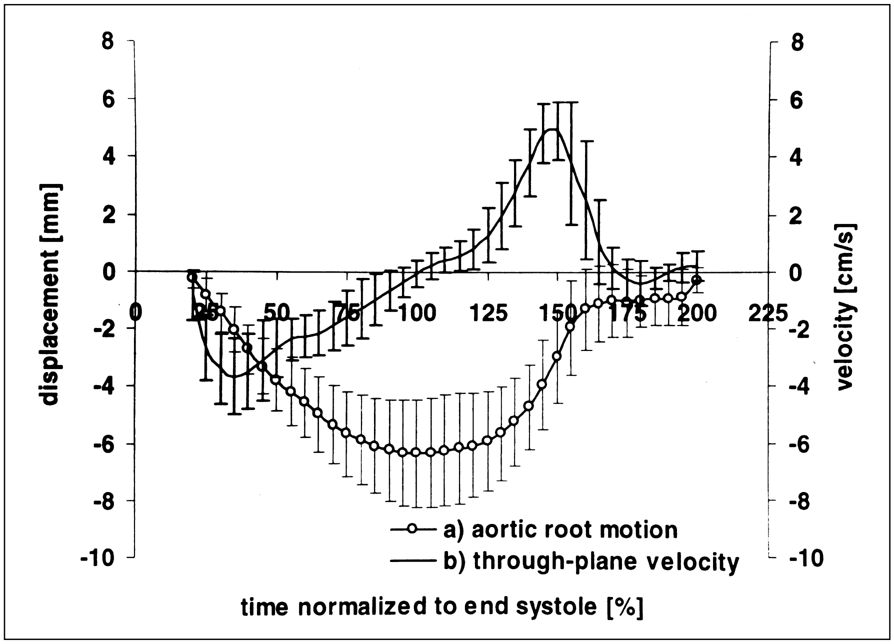


Fig. 3.—Example of figure that successfully illustrates temporal relationship between two dependent variables. Graph shows plotting relationship between two different but related phenomena using two different *y* axes: displacement (on left) and velocity (on right) for mean through-plane motion of prosthetic valve. This figure has high data–ink ratio, especially with error bars included. Choice for *x*-axis position is compromised, leading to some obscuring of data values and of *x*-axis tick labels. (Reprinted with permission from [5])

A previous article in this series discussed graphing two variables to show a relationship in their correlation [2]; the graph style shown in Figure 3 allows the comparison of two characteristics. Two phenomena that are different but related are plotted with displacement on the left *y*-axis and velocity on the right *y*-axis. The graphic is data intensive with a high data–ink ratio. Unfortunately, the choice of the *x*-axis position has obscured some of the *x*-axis tick labels and modestly confuses the interpretation of the data.

Figure 4 shows box-and-whiskers plots for contrast-to-noise ratios for a variety of different coronary vessel segments. No statistical differences exist, and the plot permits the reader to visualize that result. However, the plot does not give the number observed for each artery segment and contains additional lines of division that are nondata ink and could be erased.

Figure 5 shows an example of the ubiquitous receiver operating characteristic curve. In this instance, however, the straightforward curve with 10 data points is obscured in a sea of nondata ink, including the background grid, line of unity, extra axis tick marks, and the inserted legend, which is clearly not needed because only one data set is plotted on the graph. All these represent chartjunk and should be eliminated. Compare Figure 5

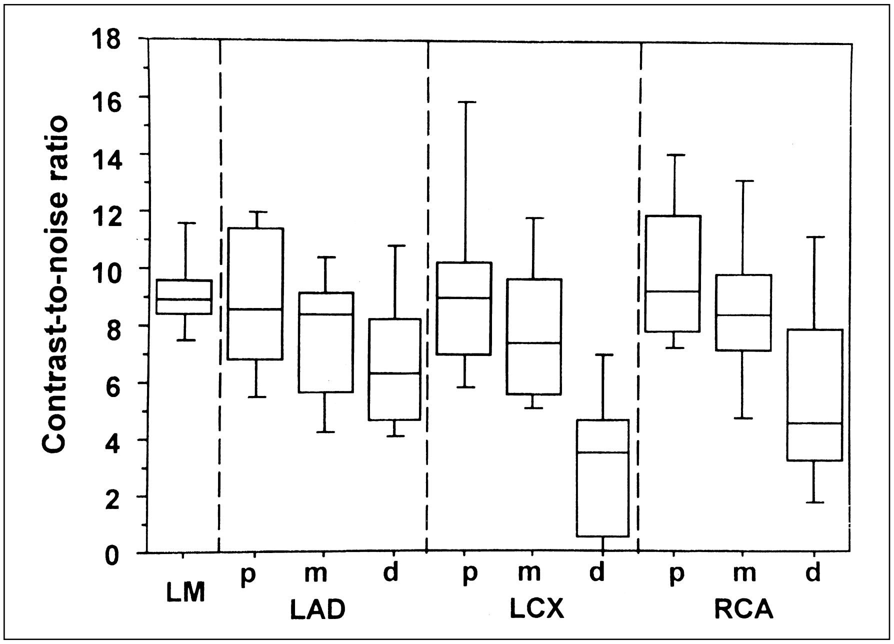


Fig. 4.—Box-and-whiskers plots. Graph shows contrast-to-noise ratio in electron-beam CT coronary angiography for different coronary vessel segments. Bottom and top edges of box are 25th and 75th percentiles, horizontal line represents the median, and error bars delimit extent of 10th and 90th percentiles. No statistical differences were observed, and this type of plot effectively portrays this data variability. LM = left main coronary artery, LAD = left anterior descending coronary artery, LCX = left circumflex coronary artery, RCA = right coronary artery, p = proximal segment, m = middle segment, d = distal segment. (Reprinted from [6])

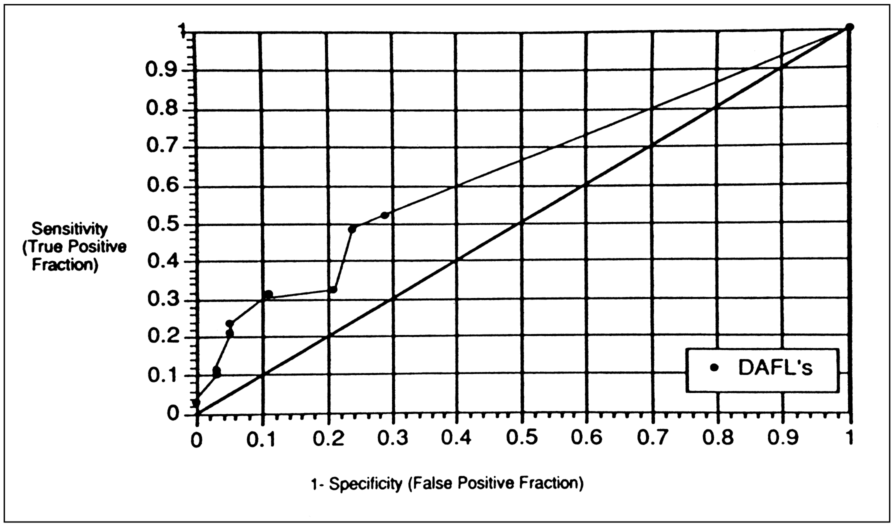


Fig. 5.—Example of poor data–ink ratio for receiver operating characteristic curve. Graph shows only 10 data points, which are obscured by tremendous amount of nondata ink, including background grid, tick marks, and line of unity. DAFL = differential air–fluid level. (Reprinted from [7])

with Figure 6, in which four curves are plotted and the data–ink ratio is high. It is clear from the curves that no differences were seen for the four display formats and three abnormalities (a–c). These data could be presented in tables because no significant differences were observed.

Figure 7 is our first example of moiré optical vibrations. This complex figure describes the calculated optimum treatment strategy in a two-way sensitivity analysis varying the relative risk of failure after stent placement. The graph shows a decrease in relative risk with the enlarging proportion of patients requiring stent placement after percutaneous transluminal angioplasty. However, no confidence intervals are shown, and the input confidence interval and proportion indicated by the arrows suggest that the three groups may not be differentiated.

Moiré effects (optical noise) are one design fault in Figure 8. Additional chartjunk is seen in the threefold repetition of the type of radiologists and the actual data values sitting atop each bar. An examination of the amount of data actually shown in the figure reveals very few data points considering the amount of ink used to represent them. Also, no significant differences exist in detectability between any measurements for any of the lesion types. A replot of the data values that decreases the repetition and clearly portrays the paucity of data (Fig. 9) still shows a lack of significance.

The bar charts shown in Figure 10 are also dominated by optical effects. The graphs depict the area under the receiver operating characteristic curve for 20 radiologists interpreting from four different displays. This figure occupies a considerable amount of visual real estate to show virtually no significant differences. Although minimal differences are indicated, no correction for multiple comparisons is indicated nor are confidence intervals shown.

The three panels of Figure 11 show mild moiré patterns and background grids. The graphs illustrate the decrease in number of vertebral disks seen with a decrease in radiation dose. However, no statistical tests were indicated. Normally, it is sufficient to plot only the upgoing section of the error bars on the top of each bar. However, in this instance, the error bars are actually the data range (the same as the range whiskers in the box-and-whiskers plot), so this is an unfamiliar hybrid plot.

Figure 12 shows the upward shift in receiver operating characteristic area values observed when a group of radiologists used computer-aided diagnosis. The choice of bar fills does not dominate the visual picture. The shift to higher receiver operating characteristic areas is clearly seen, and the variability of the distribution in performance is intact and interpretable.

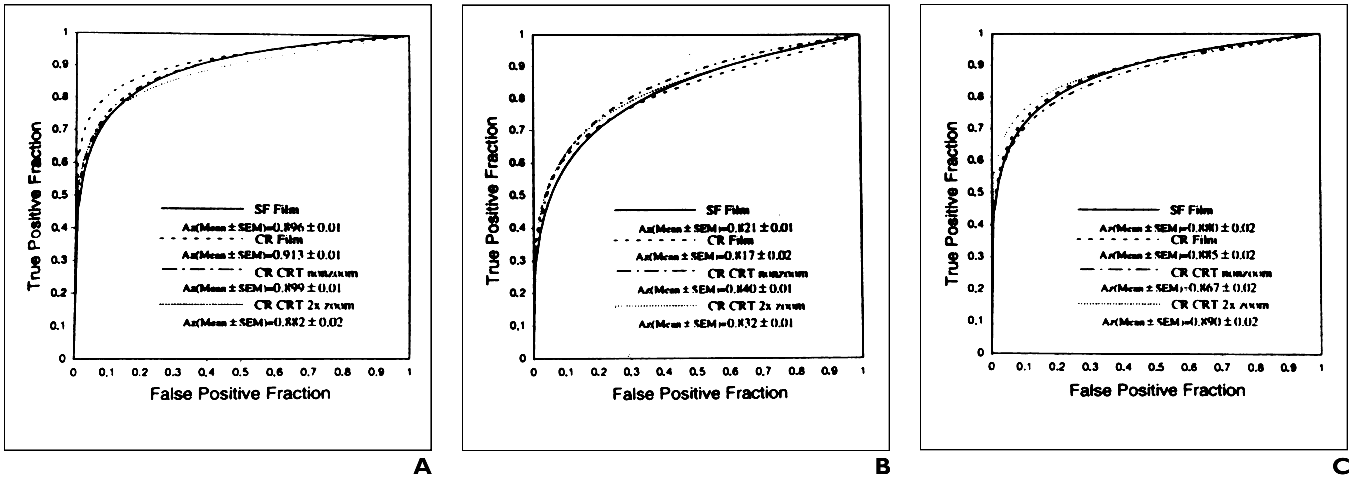


Fig. 6.—Example of data that could have been handled in table format. (Reprinted with permission from [8])

A–C, Graphs show findings for reticular (**A**), small nodular (**B**), and ground-glass (**C**) abnormalities in four display formats. Appropriate receiver operating characteristic curves are used, but curves are not significantly different for any abnormalities. Repetition is unproductive. In each graph, it is difficult to discern individual curves and their identification.

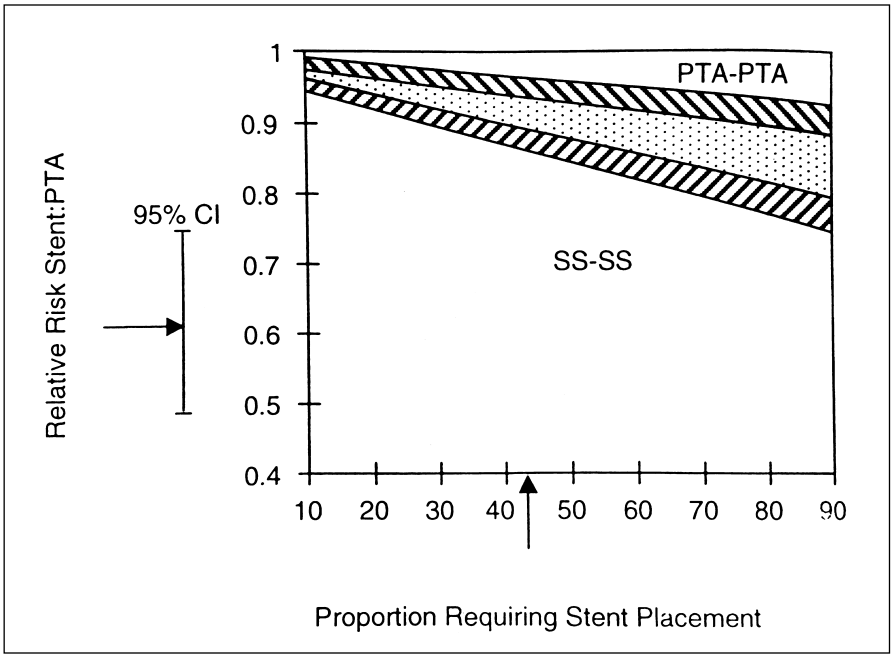


Fig. 7.—Figure in which data–ink ratio and optical vibrations (moiré effect) are poor. Graph shows complex theoretic analysis of optimal treatment strategy using two-way sensitivity analysis. PTA = percutaneous transluminal angioplasty, SS = selective stent placement, CI = confidence interval. (Reprinted with permission from [9])

The use of filled versus open bars is an effective method of delineation between groups in Figure 13. However, the graph does contain superfluous background grids, and design variation was chosen over data variation. One of the rules for graphic design suggested by Tufte [1] is that the graph's dimensions should not exceed the data dimension. Here we have a three-dimensional plot of only two-dimensional data. The graph design adds substantial visual ink without adding anything to the interpretation. However, unless the graphs are carefully considered, even one with copious data ink can be confusing.

Figure 14 shows the raw and mean values for a number of measures of FDG uptake for 10 patients. The reader cannot follow the actual values from each patient because too many overlapping symbols appear. Although the mean (the only filled symbol) is easy to pick out, the error bars add to the confusion. Clutter could have been avoided by offsetting the mean and standard deviation plots to the side of the raw data.

Figure 15 shows the alterations in proportion for a group of 20 radiologists interpreting images from two formats. They were given three types of images to view and asked which gave the best processing. No significant differences were reported. Although the bar graphs show interradiologist variability well, much ink is used to show an absence of significant changes between formats. Because all the bars

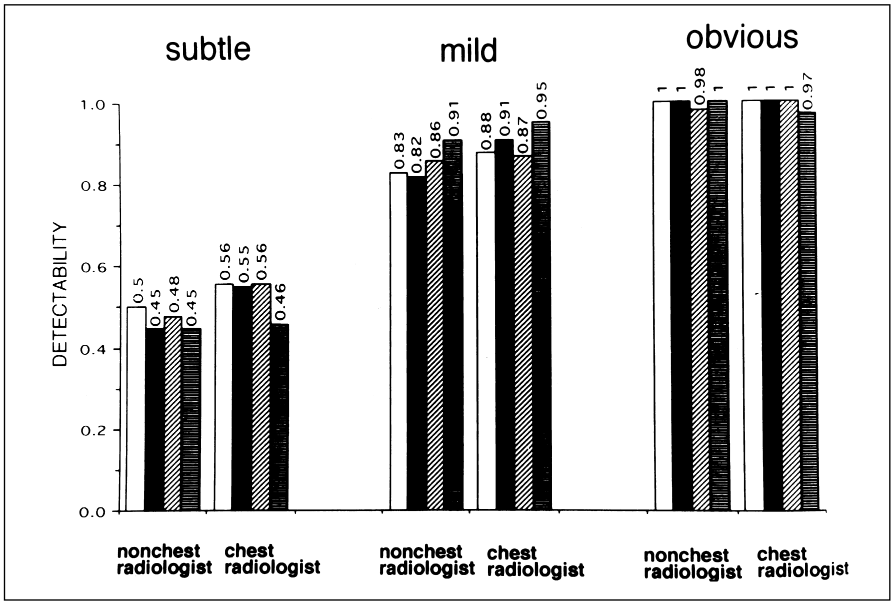


Fig. 8.—Example of figure that could have been simplified. Bar chart shows average detectability of lung abnormalities divided into severity for two groups of radiologists and four display methods. The presentation has two principal problems: moiré vibrations (optical noise) and redundancy, with the two groups of radiologists repeated for each degree of abnormality. Reprinted with permission from [8])

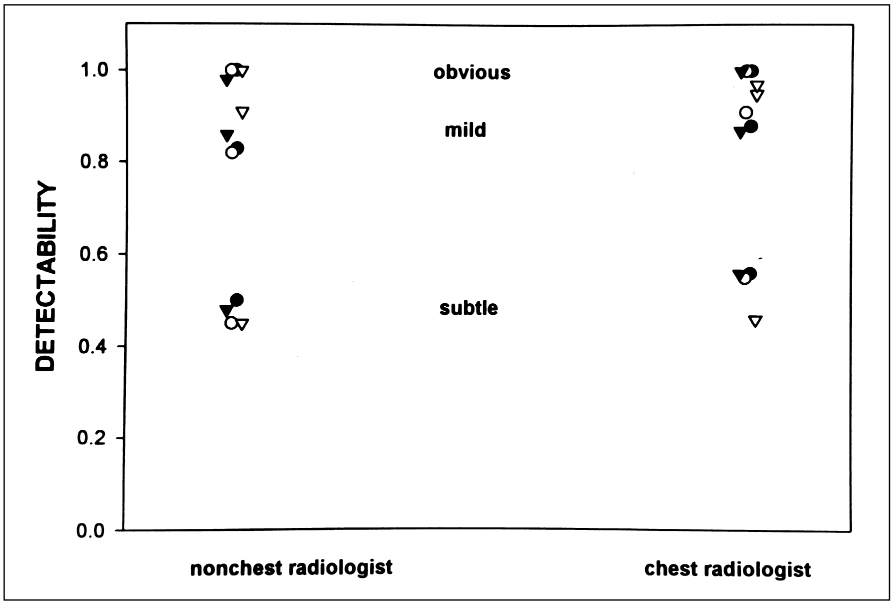


Fig. 9.—Example of another way data in Figure 8 might have been presented. Plot uses much less data ink without losing portrayal of any raw data. Different symbols are used to represent each radiologist.

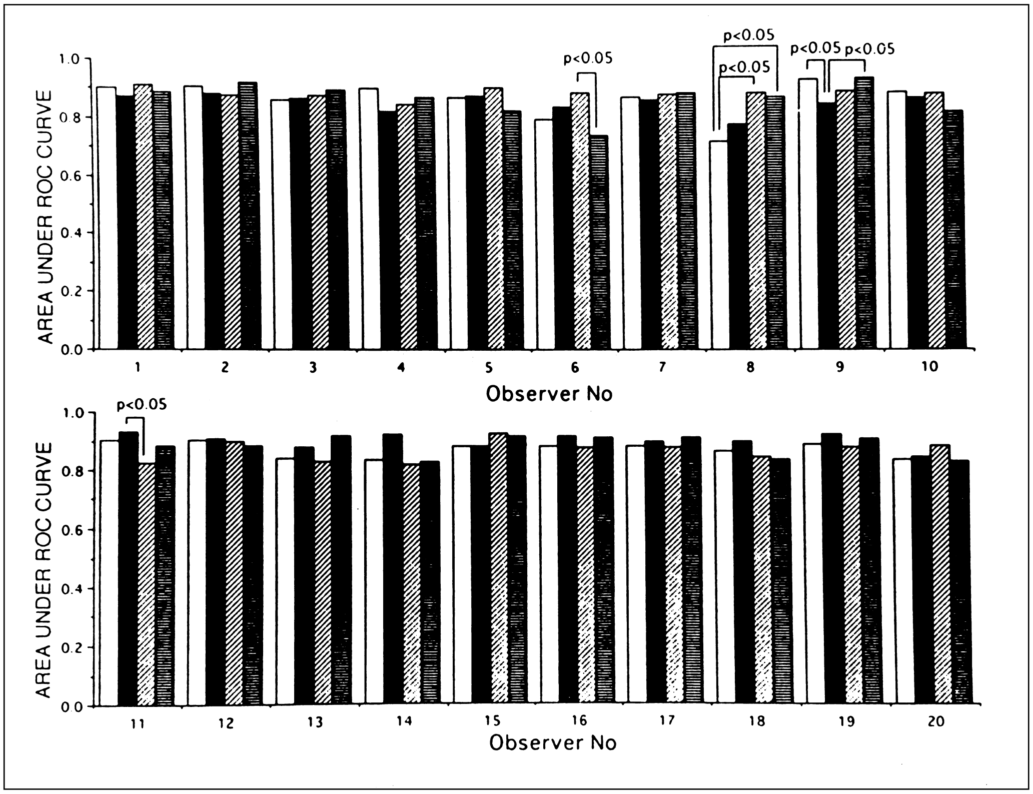


Fig. 10.—Example of bar charts dominated by moiré patterns. Illustration of all raw data for many areas from receiver operating characteristic analyses hides fact that multiple comparisons would require additional statistical tests. There is little value in occupying so much visual real estate for not much significant data. Reprinted with permission from [8])

add up to unity (one), the black infill for the third proportion is redundant.

Kaplan-Meier curves (Fig. 16) are rarely seen in radiology but are common in clinical studies. These curves are excellent for showing how rapidly a proportion of different populations reaches a predetermined clinical outcome (in this instance, stroke) for two populations divided by sonographic criteria on day 0. The left panel represents less than 50% stenosis and the right panel, greater than 50% stenosis. It is unfortunate that the two panels have different *y*-axis ranges. Visually, it appears that the patients with nonhypoechoic findings in the greater-than-50% group have about the same number of strokes as both groups in the less-than-50% panel on the left. They appear about equal, however, because of the change in scale between the panels.

Changing the axis to accentuate the differences between groups is also shown in Figure 17. Whereas the left panel has time points for 30, 60, and 90 sec, the center and right panels show the same data for one time point only, and the three scatterplots have different axis ranges. The inclusion of all the raw data is commendable, but no indication of statistical differences is shown. Using one graph could have eliminated repetition, and additional lines could have joined the same tumor at each time point to show whatever trends were found in the temporal evolution of the enhancement.

Figure 18 shows three-dimensional graphs for four spectroscopic measurements and clinical outcome in three groups of patients. Three-dimensional graphs are intuitively difficult to comprehend and these examples also show moiré effects. The use of three graph dimensions is appropriate to the three data dimensions: proportion, MR-spectroscopy measurement, and clinical outcome. The number of patients whose findings contribute to each of the bars is small, however, so this graph format overstates the value of the data. Also the lack of confidence intervals allows the graph to appear to tell a definitive story, whereas the variability of the data that would be associated with such low numbers is not illustrated. Similarly, the three-dimensional bar graphs for MR imaging findings in Figure 19 show an appropriate number of dimensions (three: grade, cohort, and age). No statistical analysis is indicated nor are confidence intervals shown. It appears that the three grades of the three panels increase with age in the whole cohort independent of the group subdivision. A considerable amount of visual real estate is used to illustrate data that have a common pattern. The findings from these three graphs could be summarized in a few sentences in the results section of the text.

An odd combination of two measurements in one graph is seen in Figure 20. The main panel represents the mean and 95% confidence intervals for the loss of cartilage thickness under pressure for 210 min. The inserted panel has a different time axis, although the scale is the same. Perhaps a better way to show these data would be to use the release point at 210 min as the zero point with times

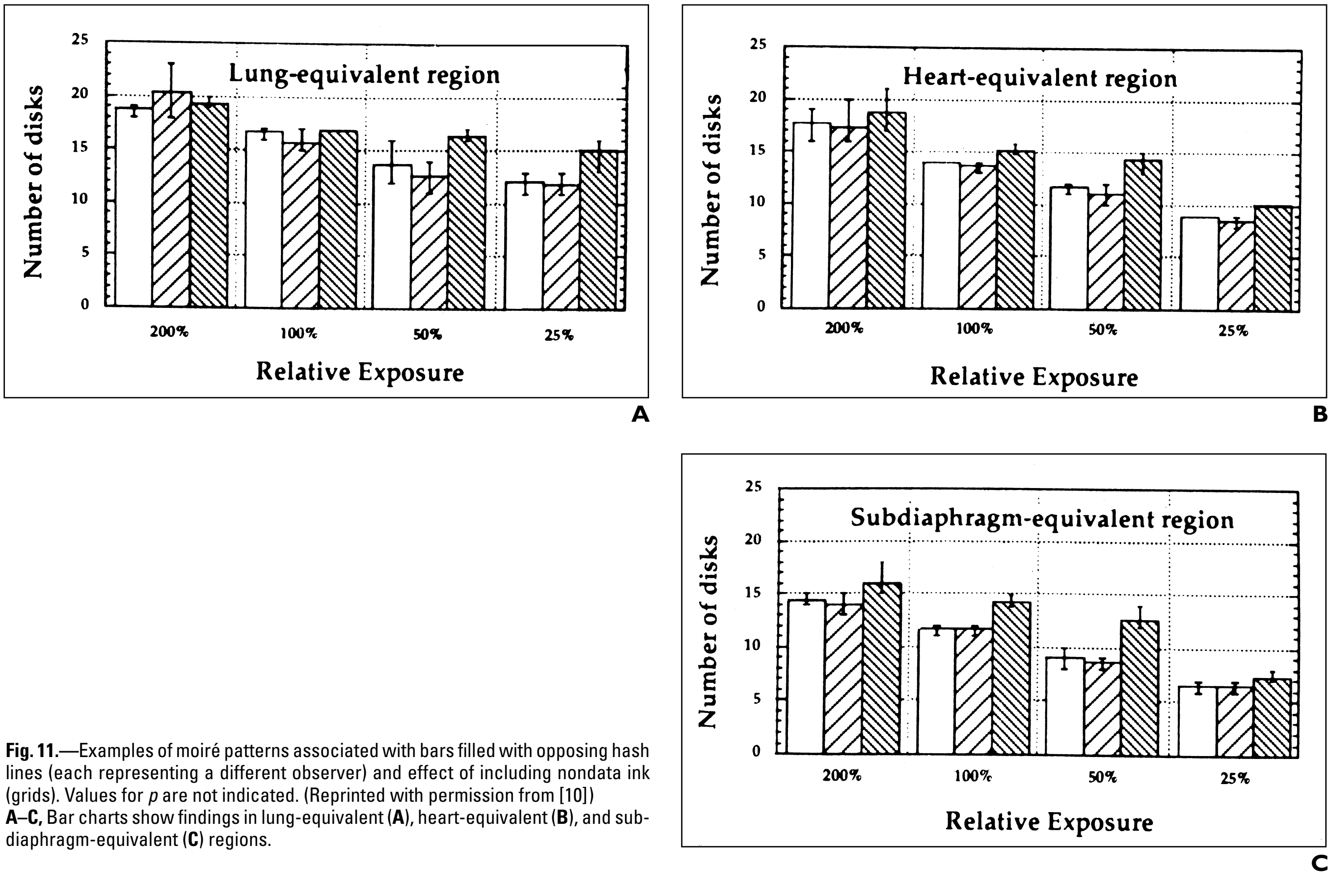


Fig. 11.—Examples of moiré patterns associated with bars filled with opposing hash lines (each representing a different observer) and effect of including nondata ink (grids). Values for *p* are not indicated. (Reprinted with permission from [10])
A–C, Bar charts show findings in lung-equivalent (**A**), heart-equivalent (**B**), and subdiaphragm-equivalent (**C**) regions.

negative before (during compression) and times positive afterward (during decompression). The two *y*-axes should be either the same or better coordinated.

The final figure of this review, Figure 21, has two panels showing the change in two phenomena as a function of time after angioplasty. The graphs have a good data–ink ratio, and actual measurements for 10 patients are illustrated. Although the overall patterns can be discerned, the mean values (dashed lines) are partially obscured, and the line indicating abnormal values is also a dashed line. The *y*-axis on panel b has been broken between 6 and 12, and the scale is smaller above the break, giving an emphasis to the lower values. The graphs also lack an indication of the reliability of the measurements and a statistical evaluation of the results.

The critique of the graphs in this article was designed to help the reader understand the principles of good data presentation, in which economy, clarity, and honesty are the essential guides.

Summary

Radiologists should apply to the selection and content of graphics conveying radiologic data the same skills they use in the selection of radiographic images for presentation or publication. This article has reviewed the fundamentals for visual display of quantitative information from radiologic studies. The truth about the data should be shown in an efficient manner and the chartjunk minimized. Clarity and honesty are paramount. Although meeting these criteria seems a valuable goal and an easy task to accomplish, these examples of graphics from the recent literature suggest that we need to scrutinize more carefully. Clarity of graphing leads to clarity of thinking and of presentation.

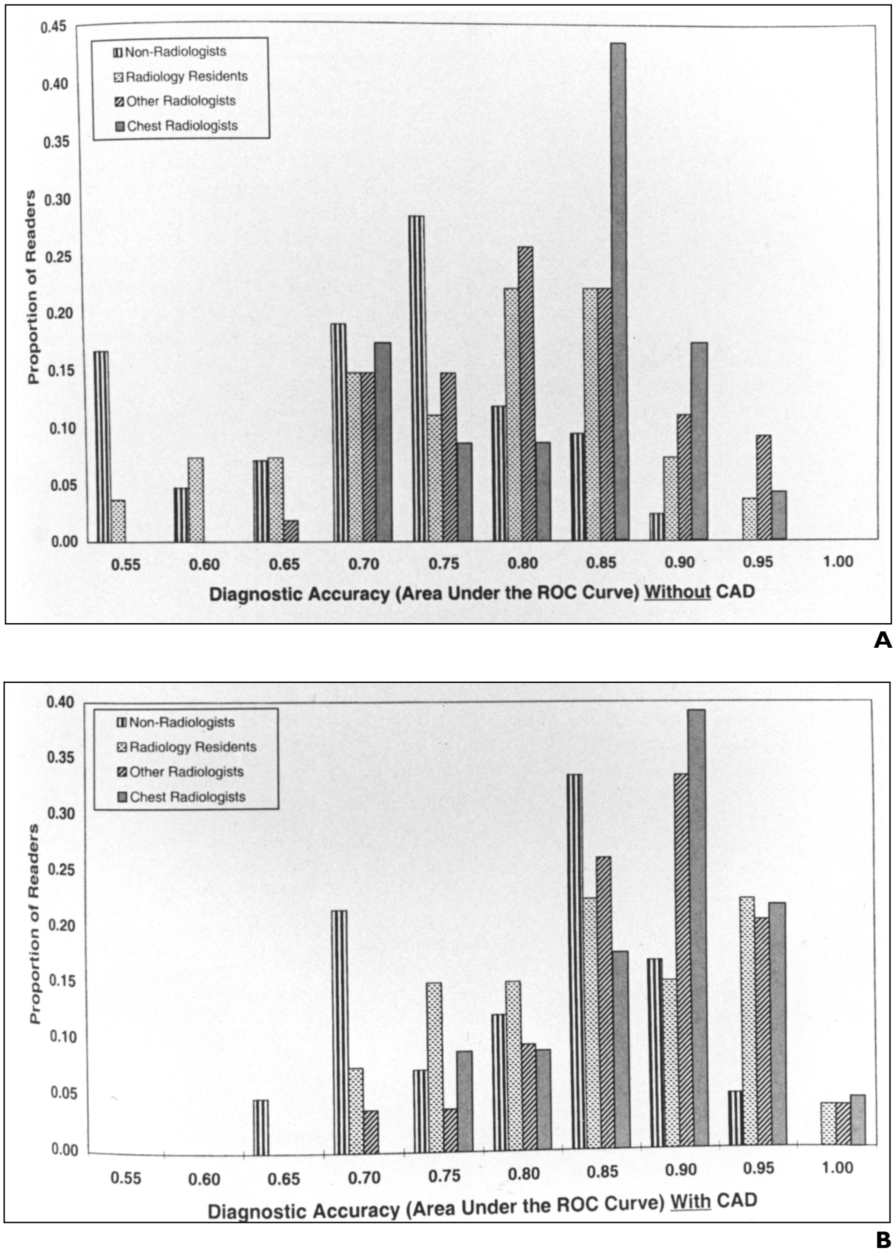


Fig. 12.—Example of figure that provides value-filled expression of improvement in diagnostic accuracy and leaves variability visible. (Reprinted with permission from [11])
A and **B,** Bar charts show diagnostic accuracy without (**A**) and with (**B**) computer-aided diagnosis (CAD). Bars have muted moiré effect and charts have more pleasing overall appearance compared with those of Figures 8, 10, and 11. Panel **B** shows that using CAD resulted in increase in diagnostic accuracy for all groups of radiologists.

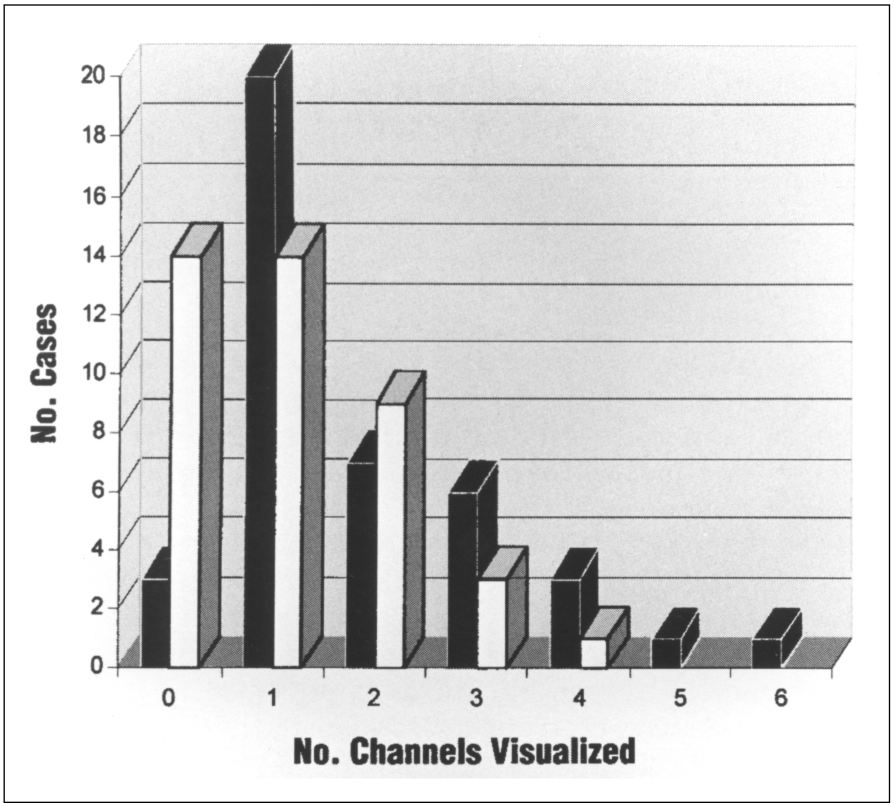


Fig. 13.—Example of visually effective use of filled versus open bars for comparing distribution of number of cases per channels visualized. Use of three-dimensional bars gives graphic variation but adds no value to depiction of data. Figure also has nondata ink in background. (Reprinted with permission from [12])

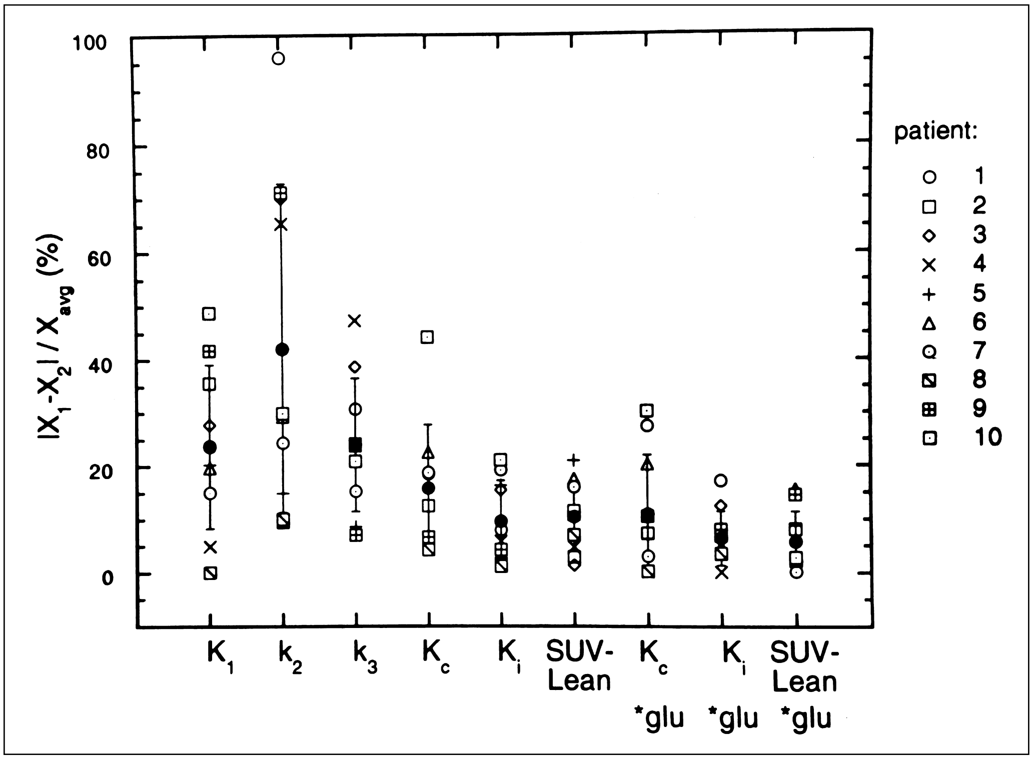


Fig. 14.—Example of complicated scatterplot. Figure depicts large amount of information for variety of FDG parameters for 10 patients. It is difficult to follow specific values for individual patients and to discern mean percentage differences (●). Error bars are confusing. (Reprinted with permission from [13])

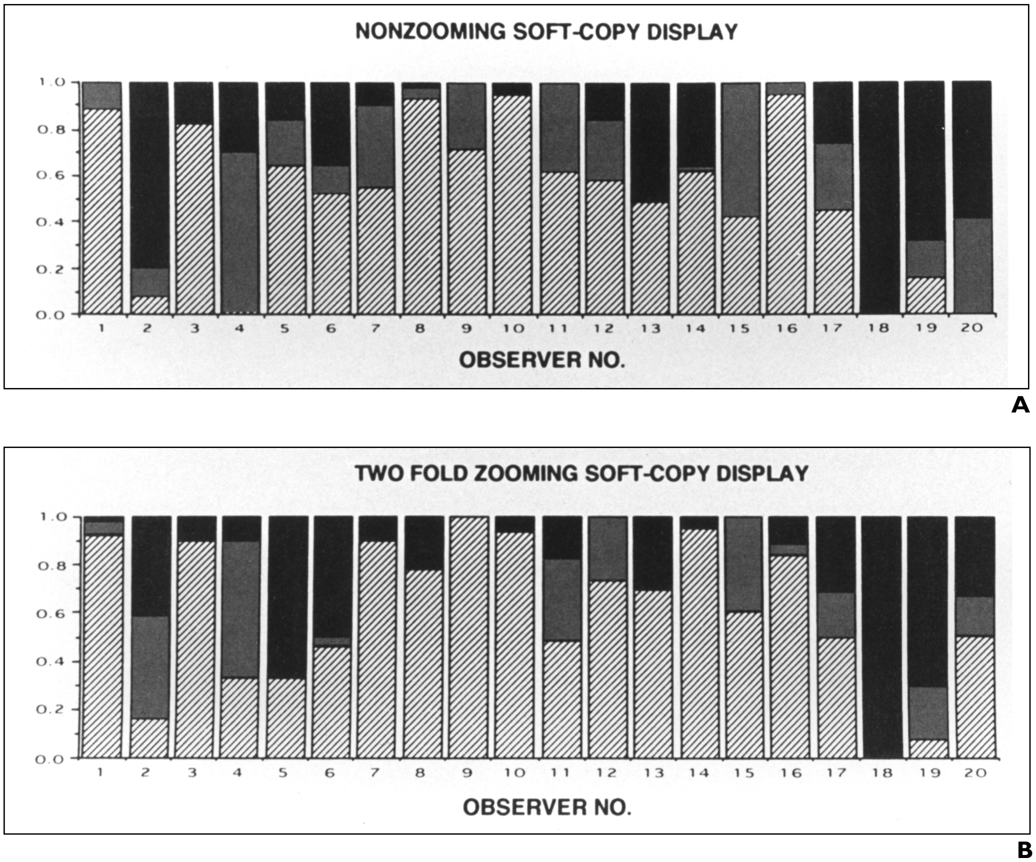


Fig. 15.—Example of interesting use of data ink to show proportions for two variables and 20 observers, with change in display parameter. No difference exists in discrimination between modalities; therefore, much ink is used to show no differences. (Reprinted with permission from [8])
A and **B,** Bar charts show differences in observer interpretation of nonzooming (**A**) and twofold zooming (**B**) soft-copy displays.

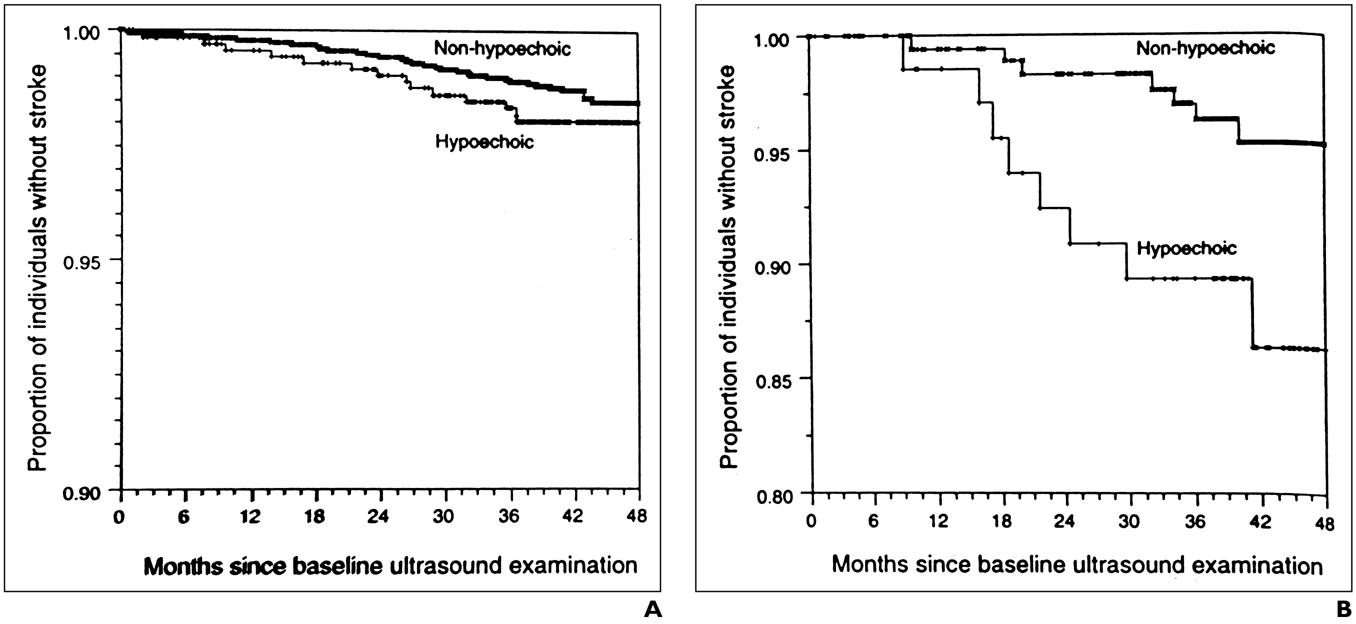


Fig. 16.—Example of Kaplan-Meier survival graphs. (Reprinted with permission from [14])
A and **B,** Graphs illustrate proportion of individuals who remain without stroke divided by degree of stenosis of less than 50% (**A**) and greater than 50% (**B**). Each group is further divided by nonhypoechoic and hypoechoic findings. Although patients with nonhypoechoic findings in **B** have higher occurrence of strokes than those of both groups in **A**, difference in *y*-axis range in **B** makes proportions appear nearly identical.

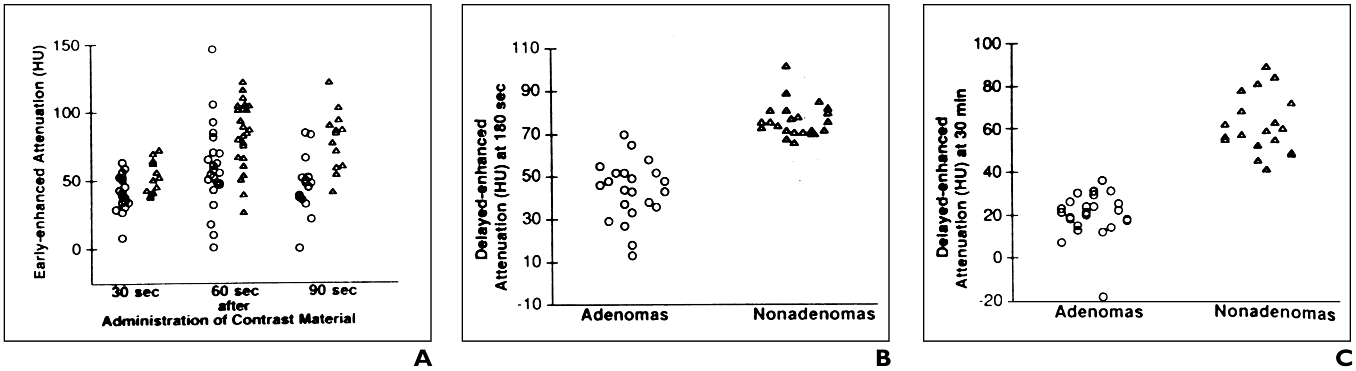


Fig. 17.—Three scatterplots showing attenuation of early-enhanced CT images of adenomas and nonadenomas at different times after injection of contrast material. No statistical differences were indicated. (Reprinted with permission from [15])

A–C, Scatterplots show data at different time intervals: 30, 60, and 90 sec (**A**); 180 sec only (**B**); and 30 min only (**C**). Because *y*-axis scales are changed for each part, this presentation visually suggests that discrimination between groups is noted at 30 min. Parts **B** and **C** should have also been plotted with attenuation versus all times of observation to reduce redundancy and nondata ink.

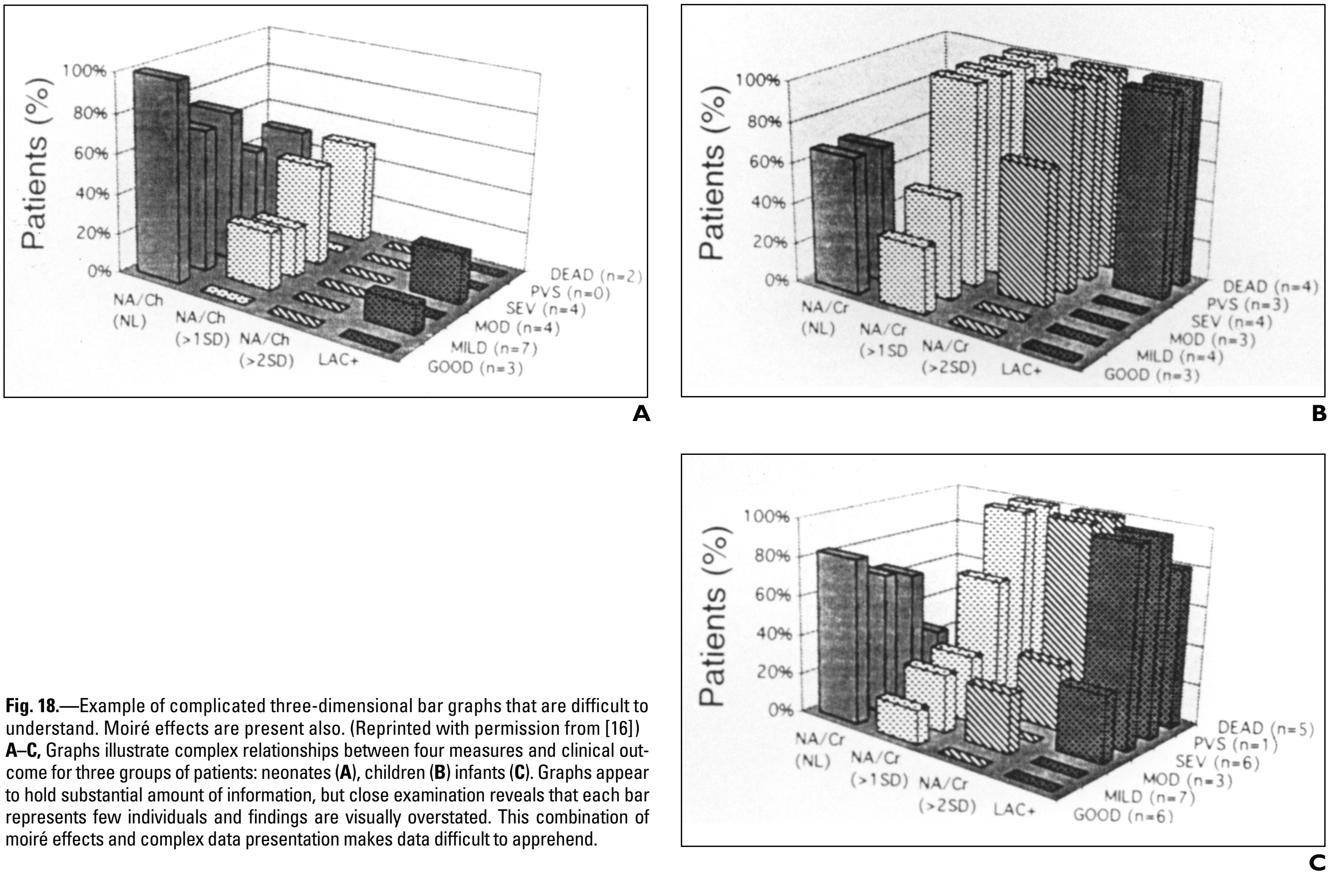


Fig. 18.—Example of complicated three-dimensional bar graphs that are difficult to understand. Moiré effects are present also. (Reprinted with permission from [16])

A–C, Graphs illustrate complex relationships between four measures and clinical outcome for three groups of patients: neonates (**A**), children (**B**) infants (**C**). Graphs appear to hold substantial amount of information, but close examination reveals that each bar represents few individuals and findings are visually overstated. This combination of moiré effects and complex data presentation makes data difficult to apprehend.

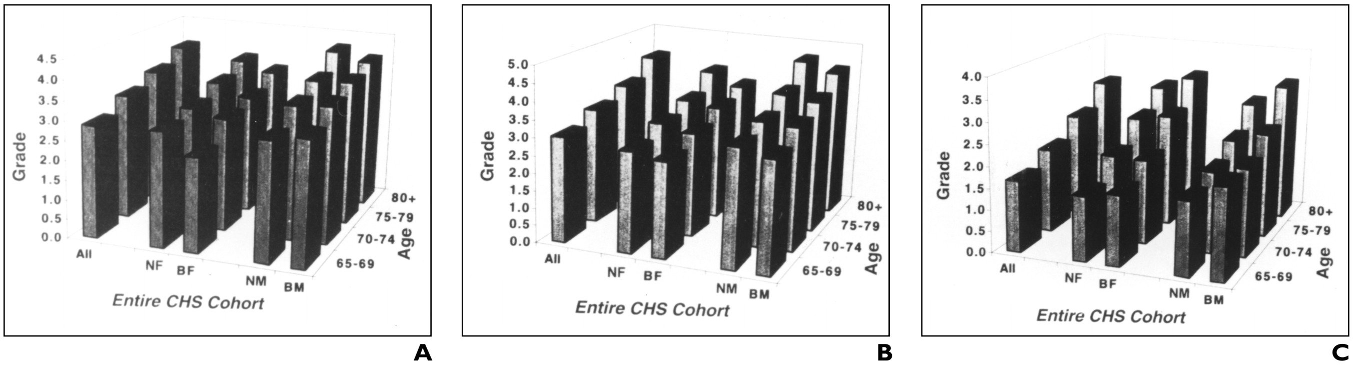


Fig. 19.—Example of material that could have been presented in text or table format because no significant differences were found and data content is minimal. (Reprinted with permission from [17])

A–C, Three-dimensional graphs show grade-scoring changes for subgroups sulcai (**A**), ventricular (**B**), and white matter (**C**) grades and ages. No error bars are shown, and numbers of subjects in each subgroup are not given. CHS = cardiovascular health study, NF = nonblack female, BF = black female, NM = nonblack male, BM = black male.

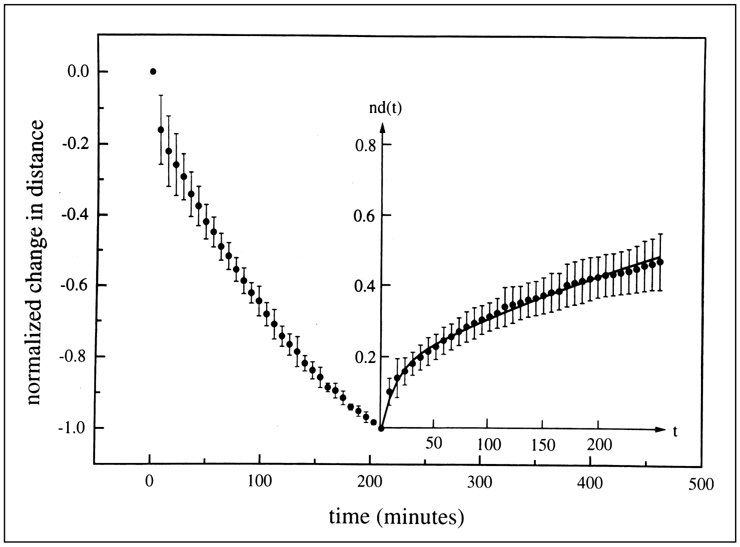


Fig. 20.—Unusual figure that inserts graph of completely different phenomenon within main (enclosing) graph. Although it is sometimes useful to have different plots using different axes in one figure, this combination is both confusing and potentially misleading. Minimum acceptable figure would have identical time axis, perhaps with release point at which time equals zero. (Reprinted with permission from [18])

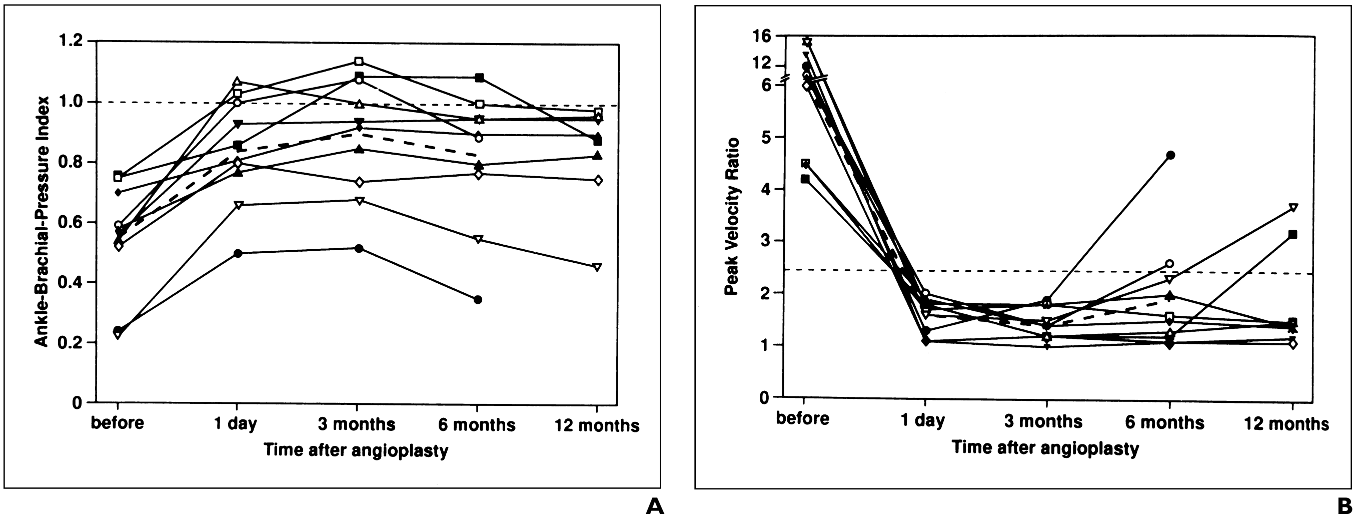


Fig. 21.—Examples of graphs in which changes in values for individual patients are almost impossible to follow. A large amount of data ink was used. (Reprinted with permission from [19])

A and **B,** Graphs illustrate changes before and after angioplasty in two vascular phenomena, ankle–brachial pressure (**A**) and peak velocity (**B**). Discerning mean values (*thick dashed lines*) is difficult. Limits for abnormal values (*thin dashed lines*) are useful. *Y*-axis scaling for part **B** is different below and above axis break, emphasizing lower values. No indication of reliability or statistical tests for measurements are provided, even for individual cases, so we cannot judge whether differences are significant.

References

1. Tufte ER. *The visual display of quantitative information.* Cheshire, CT: Graphics, **1983**:51–111
2. Karlik SJ. Exploring and summarizing radiologic data. *AJR* **2003**;180:47–54
3. Izumi M, Hida A, Takagi Y, Kawabe Y, Eguchi K, Takashi N. MR imaging of the salivary glands in Sicca syndrome: comparison of lipid profiles and imaging in patients with hyperlipidemia and patients with Sjögren's syndrome. *AJR* **2000**;175:829–834
4. Kuszyk BS, Bluemke DA, Choti MA, Horton KM, Magee CA, Fishman EK. Contrast-enhanced CT of small hypovascular hepatic tumors: effect of lesion enhancement on conspicuity in rabbits. *AJR* **2000**;174:471–475
5. Kozerke S, Hasenkam JM, Nygaard H, Paulsen PK, Pedersen EM, Boesiger P. Heart-motion-adapted MR velocity mapping of blood velocity distribution downstream of aortic valve prostheses: initial experience. *Radiology* **2001**;218: 548–555
6. Chernoff DM, Ritchie CJ, Higgins CB. Evaluation of electron beam CT coronary angiography in healthy subjects. *AJR* **1997**;169:93–99
7. Harlow CL, Stears RLG, Zeligman BE, Archer DG. Diagnosis of bowel obstruction on plain abdominal radiographs: significance of air–fluid levels at different heights in the same loop of bowel. *AJR* **1993**;161:291–295
8. Ishigaki T, Endo T, Ikeda M, et al. Subtle pulmonary disease: detection with computed radiography versus conventional chest radiography. *Radiology* **1996**;201:51–60
9. Bosch JL, Tetteroo E, Mali WP, Hunik MGM. Iliac artery occlusive disease: cost-effectiveness analysis of stent placement versus percutaneous transluminal angioplasty. *Radiology* **1998**;208: 641–648
10. Harrell GC, Floyd CE, Johnston GA, Ravin CE. Quality control phantom for digital chest radiography. *Radiology* **1997**;202:111–116
11. McMahon H, Engelmann R, Behlen FM, et al. Computer-aided diagnosis of pulmonary nodules: results of a large-scale observer test. *Radiology* **1999**;213:723–726
12. Goldfarb LR, Alazraki NP, Eshima D, Eshima LA, Herda SC, Halkar RK. Lymphoscintigraphic identification of sentinel lymph nodes: clinical evaluation of 0.22 mm filtration of Tc-99m sulfur colloid. *Radiology* **1998**;208:505–509
13. Minn H, Zasadny KR, Quint LE, Wall RL. Lung cancer: reproducibility of quantitative measurements for evaluating 2-[F-18]-Fluoro-2-deoxy-d-glucose uptake at PET. *Radiology* **1995**;196:167–173
14. Polak JF, Shemanski L, O'Leary DH, et al. Hypoechoic plaque at ultrasound of the carotid artery: an independent risk factor for incident stroke in adults age 65 years or older. *Radiology* **1998**;208: 649–654
15. Szolar DH, Kammerhuter F. Quantitative CT evaluation of adrenal gland masses: a step forward in the differentiation between adenomas and non-adenomas? *Radiology* **1997**;202:517–521
16. Holhouser BA, Ashwal S, Luy GY, et al. Proton MR spectroscopy after acute central nervous system injury: outcome prediction in neonates, infants and children. *Radiology* **1997**;202:487–496
17. Chang Yue N, Arnold AM, Lonsteth WT, et al. Sulcal, ventricular and white matter changes at MR imaging in the aging brain: data from the cardiovascular health study. *Radiology* **1997**; 202:33–37
18. Rubenstein JD, Kim JK, Henkelman RM. Effects of comparison and recovery on bovine articular cartilage: appearance on MR images. *Radiology* **1996**;201:843–850
19. Minar E, Pokrajac B, Ahmadi R, et al. Brachytherapy for prophylaxis of restenosis after long-segment femoropopliteal angioplasty: pilot study. *Radiology* **1998**;208:173–179

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003

Fundamentals of Clinical Research for Radiologists

Lawrence Joseph[1,2]
Caroline Reinhold[3]

Introduction to Probability Theory and Sampling Distributions

Statistical inference allows one to draw conclusions about the characteristics of a population on the basis of data collected from a sample of subjects from that population. Almost all the statistical inferences typically seen in the medical literature are based on probability models that connect summary statistics calculated using the observed data to estimates of parameter values in the population. This article will cover the basic principles behind probability theory and examine a few simple probability models that are commonly used, including the binomial, normal, and Poisson distributions. We will then see how sampling distributions are used as the basis for statistical inference and how they are related to simple probability models. Thus, this article forms the foundation for future articles in the series that will present the details of statistical inference in particular clinical situations.

Making medical decisions on the basis of findings from various radiologic diagnostic tools is an everyday occurrence in clinical practice. In radiologic research, one often needs to draw conclusions about the relative performance of one diagnostic tool compared with another for the detection of a given condition of interest. Both of these tasks depend, in large part, on probability theory and its applications. In diagnosis, we are interested in calculating the probability that the condition of interest is present on the basis of results of a radiologic test. This probability depends on how sensitive and specific that test is in diagnosing the condition and on the background rate of the condition in the population.

This calculation largely depends on a result from probability called Bayes' theorem. Similarly, all statistical inferences, whether comparisons of two proportions representing diagnostic accuracies from two instruments or inferences from a more complex model, are based on probabilistic reasoning. Therefore, a thorough understanding of the meaning and proper interpretation of statistical inferences, crucial to daily decision making in a radiology department, depends on an understanding of probability and probability models.

This article is composed of three main parts. We begin with an introduction to probability, including the definitions of probability, the different schools of thought about the interpretation of probabilities, and some simple examples. We continue by defining conditional probabilities and present Bayes' theorem, which is used to manipulate conditional probabilities. The most common simple probability models, including the binomial, normal, and Poisson distributions, are presented next, along with the types of situations in which we would be most likely to use them. Finally, sampling strategies are examined. Armed with these basics of probability and sampling, we conclude with a discussion of how the outcome of interest defines the model parameter on which to focus inferences and how the sampling distribution of the estimator of that parameter enables valid inferences from the data collected in the sample about the population at large.

Probability

Definitions of Probability

Last's *Dictionary of Epidemiology* [1] presents two main definitions for probability. The

Received July 9, 2002; accepted after revision August 1, 2002.

Series editors: Craig A. Beam, C. Craig Blackmore, Stephen Karlik, and Caroline Reinhold.

This is the tenth in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Reseach and Technology Assessment

[1] Department of Medicine, Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Ave., Montreal, Quebec, H3G 1A4, Canada. Address correspondence to L. Joseph.

[2] Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave. W., Montreal, Quebec, H3A 1A2, Canada.

[3] Department of Diagnostic Radiology, Montreal General Hospital, McGill University Health Centre, 1650 Cedar Ave., Montreal, Quebec, H3G 1A4, Canada.

AJR 2003;180:917–923

0361–803X/03/1804–917

first definition, which represents the view of the frequentist school of statistics, defines the probability of an event as the number of times the event occurs divided by the number of trials in which it could have occurred, *n*, as *n* approaches infinity. For example, the probability that a coin will come up heads is 0.5 because, assuming the coin is fair, as the number of trials (flips of the coin) gets larger and larger, the observed proportion will be, on average, closer and closer to 0.5. Similarly, the probability that an intervention for back pain is successful would be defined as the number of times it is observed to be successful in a large (theoretically infinite) number of trials in patients with back pain.

Although this definition has a certain logic, it has some problems. For example, what is the probability that team A will beat team B in their game tonight? Because this is a unique event that will not happen an infinite number of times, the definition cannot be applied. Nevertheless, we often hear statements such as "There is a 60% chance that team A will win tonight." Similarly, suppose that a new intervention for back pain has just been developed, and a radiologist is debating whether to apply it to his or her next patient. Surely the probability of success of the new intervention compared with the probability of success of the standard procedure for back pain will play a large role in the decision. However, no trials (and certainly not an infinite number of trials) as yet exist on which to define the probability. Although we can conceptualize an infinite number of trials that may occur in the future, this projection does not help in defining a probability for today's decision. Clearly, this definition is limited, not only because some events can happen only once, but also because one cannot observe an infinite number of like events.

The second definition, often referred to as the Bayesian school, defines the probability of any event occurring as the personal degree of belief that the event will occur. Therefore, if I personally believe that there is a 70% chance that team A will win tonight's game, then that is my probability for this event. In coin tossing, a Bayesian may assert, on the basis of the physics of the problem and perhaps a number of test flips, that the probability of a coin flip coming up heads should be close to 0.5. Similarly, on the basis of an assessment that may include both previously available data and subjective beliefs about the new technique, a radiologist may assert that the probability that a procedure will be successful is 85%.

The obvious objection to Bayesian probability statements is that they are subjective, and thus different radiologists may state different probabilities for the success rate of the new technique. In general, no single "correct" probability statement may be made about any event, because such statements reflect personal subjective beliefs. Supporters of the Bayesian viewpoint counter that the frequentist definition of probability is difficult to apply in practice and does not pertain to many important situations. Furthermore, the possible lack of agreement as to the correct probability for any given event can be viewed as an advantage, because it will correctly mirror the range of beliefs that may exist about any event that does not have a large amount of data from which to accurately estimate its probability. Hence, having a range of probabilities depending on the personal beliefs of a community of clinicians is a useful reflection of reality. As more data accumulate, Bayesian and frequentists probabilities tend to agree, each essentially converging to the mean of the data. When this occurs, similar inferences will be reached from either viewpoint.

Discussion of these two ways of defining probability may seem to be of little relevance to radiologists but, later in this series, it will become apparent that it has direct implications for the type of statistical analysis to be performed. Different definitions of probability lead to different schools of statistical inference and, most importantly, often to different conclusions based on the same set of data. Any given statistical problem can be approached from either a frequentist or a Bayesian viewpoint, and the choice often depends on the experience of the user more than it does on one or the other approach being more appropriate for a given situation. In general, Bayesian analyses are more informative and allow one to place results into the context of previous results in the area [2], whereas frequentist methods are often easier to carry out, especially with currently available commercial statistical packages. Although most analyses in medical journals currently follow the frequentist definition, the Bayesian school is increasingly present, and it will be important for readers of medical journals to understand both.

The lack of a single definition of probability may be disconcerting, but it is reassuring to know that whichever definition one chooses, the basic rules of probability are the same.

Rules of Probability

Four basic rules of probability exist. These rules are usually expressed more rigorously than is necessary for the purposes of this article, through the use of set theory and probability notation.

The first rule states that, by convention, all probabilities are numbers between 0 and 1. A probability of 0 indicates an impossible event, and a probability of 1 indicates an event certain to happen. Most events of interest have probabilities that fall between these extremes.

The second rule is that events are termed "disjoint" if they have no outcomes in common. For example, the event of a patient having cancer is disjoint from the event of the same patient not having cancer, because both cannot happen simultaneously. On the other hand, the event of cancer is not disjoint from the event that the patient has cancer with metastases because in both cases the outcome of cancer is present. If events are disjoint, then the probability that one or the other of these events occurs is given by the sum of the individual probabilities of these events. For example, in looking at an MR image of the liver, if the probability that the diagnosis is a hepatoma is 0.5 (meaning 50%) and the probability of a metastases is 0.3, then the probability of either hepatoma or metastases must be 0.8, or 80%.

The third rule is expressed as follows: If one could list the set of all possible disjoint events of an experiment, then the probability of one of these events happening is 1. For example, if a patient is diagnosed according to a 5-point scale in which 1 is defined as no disease; 2, as probably no disease; 3, as uncertain disease status; 4, as probably diseased; and 5, as definitely diseased, then the probability that one of these states is chosen is 1.

The fourth rule states that, if two events are independent (i.e., knowing the outcome of one provides no information concerning the likelihood that the other will occur), then the probability that both events will occur is given by the product of their individual probabilities. Thus, if the probability that findings on an MR image will result in a diagnosis of a malignant tumor is 0.1, and the probability that it will rain today is 0.3 (an independent event, presumably, from the results of the MR imaging), then the probability of a malignant tumor and rain today is $0.1 \times 0.3 = 0.03$, or 3%.

In summary, probabilities for events always follow these four rules, which are compatible with common sense. Such probability calculations can be useful clinically, for example, in deriving the probability of a certain diagnosis given one or more diagnostic test results. Many probability calculations used in clinical research involve conditional probabilities. These are explained next.

Conditional Probabilities and Bayes' Theorem

What is the probability that a given patient has endometrial cancer? Clearly, this depends on a number of factors, including age, the presence or absence of postmenopausal bleeding, and others. In addition, our assessment of this probability may drastically change between the time of the patient's initial clinic visit and the point at which diagnostic test results become known. Thus, the probability of endometrial cancer is conditional on other factors and is not a single constant number by itself. Such probabilities are known as conditional probabilities. Notationally, if unconditional probabilities can be denoted by Pr(cancer), then conditional probabilities can be denoted by Pr(cancer | diagnostic test is positive), read as "the probability of cancer given or conditional on a positive diagnostic test result," and, similarly, Pr(cancer | diagnostic test is negative), read as "the probability of cancer given a negative diagnostic test result." These probabilities are highly relevant to radiologic practice and clinical research in radiology.

Because they are a form of probability, conditional probabilities must follow all rules as outlined in the previous section. In addition, however, there is an important result that links conditional probabilities to unconditional probability statements. In general, if we denote one event by A, and a second event by B, then we can write

$$Pr\,(A \mid B) = \frac{Pr\,(A \text{ and } B)}{Pr(B)}.$$

In words, the probability that event A occurs, given that we already know that event B has occurred, denoted by $Pr\ (A \mid B)$, is given by dividing the unconditional probability that these two events occur together by the unconditional probability that B occurs. Of course, this formula can be algebraically manipulated, so that it must also be true that

$$Pr\,(A \text{ and } B) = Pr(B) \times Pr\,(A \mid B).$$

For example, suppose that in a clinic dedicated to evaluating patients with postmenopausal bleeding, endovaginal sonography is often used for the detection of endometrial cancer. Assume that the overall probability of a patient in the clinic having endometrial cancer is 10%. This probability is unconditional, that is, it is calculated from the overall prevalence in the clinic; before any test results are known. Furthermore, suppose that the sensitivity of endovaginal sonography for diagnosing endometrial cancer is 90%. If we let A represent the event that the patient has a positive endovaginal sonography, and let B represent the probability of endometrial cancer in this patient population, then we can summarize the above information as $Pr(B) = 0.1$ and $Pr(A \mid B) = 0.9$. By using the formula described, we can deduce that the probability that a patient in this clinic has both endometrial cancer and positive results on endovaginal sonography is $0.1 \times 0.9 = 0.09$ or 9%.

In typical clinical situations, we may know the background rate of the disease in question in the population referred to a particular clinic (which may differ from clinic to clinic), and we may have some idea of the sensitivity and specificity of the test. Notice that in the terms used, sensitivity and specificity may be considered conditional probabilities because they provide the probability of testing positive given a subject who truly has the condition of interest (i.e., $Pr\ [A \mid B]$, which is the sensitivity), and the probability of not testing positive given the absence of the condition of interest (i.e., the specificity, Pr [not A | not B]). What should a clinician conclude if a patient walks through the door with a "positive" test result in hand? In this case, one would like to know the probability of the patient's being truly positive for the condition, given that he or she has just had a test with positive findings. Of course, if the diagnostic test is a perfect gold standard, one can simply look at the test result and be confident of the conclusion.

However, most tests do not have perfect sensitivity and specificity, and thus a probability calculation is needed to find the probability of a true-positive, given the positive test result. In our notation, we know the prevalence of the condition in our population, $Pr(B)$, and we know the sensitivity and specificity of our test, given by $Pr\ (A \mid B)$ and Pr (not A | not B), but we want to know $Pr(B \mid A)$, which is opposite in terms of what is being conditioned on. How does one reverse the conditioning argument, in effect making statements about $Pr(B \mid A)$ when we only know $Pr\ (A \mid B)$? The answer is to use a general result from probability theory, called Bayes' theorem, which states

$$Pr\,(B \mid A) = \frac{Pr\,(B) \times Pr\,(A \mid B)}{Pr\,(B) \times Pr(A \mid B) + Pr(\text{not } B) \times Pr(A \mid \text{not } B)}.$$

Suppose that the background rate of endometrial cancer seen in patients referred to a particular radiology clinic is 10% and that a diagnostic test is applied that has $Pr(A \mid B)$ = 90% sensitivity and Pr(not A | not B) = 80% specificity. What is the probability that a patient with positive test results in fact has endometrial cancer? According to Bayes' theorem, we calculate

$$Pr\,(B \mid A) = \frac{Pr\,(B) \times Pr\,(A \mid B)}{Pr\,(B) \times Pr\,(A \mid B) + Pr\,(\text{not } B) \times Pr\,(A \mid \text{not B})}$$

$$= \frac{0.1 \times 0.9}{0.1 \times 0.9 + 0.9 \times 0.2}$$

$$= 0.33$$

or about 33%. In this case, even when a patient has a positive test result, the chances that the disease is present are less than 50%.

Similarly, what is the probability that a subject testing negative has endometrial cancer? Again using Bayes' theorem,

$$Pr\,(B \mid \text{not } A) = \frac{Pr\,(B) \times Pr\,(\text{not } A \mid B)}{Pr(B) \times Pr(\text{not} A \mid B) + Pr(\text{not} B) \times Pr(\text{not} A \mid \text{not} B)}$$

$$= \frac{0.1 \times 0.1}{0.1 \times 0.1 + 0.9 \times 0.8}$$

$$= 0.013.$$

Thus, starting from a background rate of 10% (pretest probability), the probability of cancer rises to 33% after a positive diagnosis and falls to approximately 1% after a negative test (posttest probabilities). Thus, Bayes' theorem allows us to update our probabilities after learning the test result, and it is thus of great usefulness to practicing radiologists. The next module in this series covers Bayes' theorem and diagnostic tests in more detail.

Probability Models

Rather than working out all problems involving probabilities by first principles using the basic probability rules as we have discussed, it is possible to use short cuts that have been devised for common situations, leading to probability functions and probability densities. Here we review three of the most common distributions: the binomial, the normal, and the Poisson. Which distribution to use depends on many situation-specific factors, but we provide some general guidelines for the appropriate use of each.

The Binomial Distribution

One of the most commonly used probability functions is the binomial. The binomial distribution allows one to calculate the probability of obtaining a given number of "successes" in a given number of trials, wherein the probability of a success on each trial is assumed to be p. In general, the formula for the binomial probability function is

$$Pr(\text{x successes in } n \text{ trials}) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x},$$

where $n!$ is read "n factorial" and is shorthand for

$$n\times(n-1)\times(n-2)\times(n-3)\times\ldots\times3\times2\times1.$$

For example, 5! = 5 × 4 × 3 × 2 × 1 = 120, and so on. By convention, 0! = 1. Suppose we wish to calculate the probability of $x = 8$ successful angioplasty procedures in $n = 10$ patients with unilateral renal artery stenosis, wherein the probability of a successful angioplasty each time is 70%. From the binomial formula, we can calculate

$$\frac{10!}{8!2!}0.7^8(1-0.7)^2 = 0.2335,$$

so that there is slightly less than a one-in-four chance of getting eight successful angioplasty procedures in 10 trials. Of course, these days such calculations are usually done by computer, but seeing the formula and calculating a probability using it at least once helps to avoid that "black box" feeling one can often get when using a computer and adds to the understanding of the basic principles behind statistical inference. Similarly, the probability of getting eight or more (that is, eight or nine or 10) successful angioplasty procedures is found by adding three probabilities of the type shown, using the second probability rule because these events are disjoint. As an exercise, one can check that this probability is 0.3829. See Figure 1 for a look at all probabilities for this problem, in which x varies from zero to 10 successes for $n = 10$ and $p = 0.7$.

The binomial distribution has a theoretic mean of $n \times p$, which is a nice intuitive result. For example, if one performs $n = 100$ trials, and on each trial the probability of success is $p = 0.4$ or 40%, then one would intuitively expect 100 × 0.4 = 40 successes. The variance, σ^2, of a binomial distribution is $n \times p \times (1 - p)$, so that in the example just given it would be 100 × 0.4 × 0.6 = 24. Thus, the SD is

$$\sqrt{\sigma^2} = \sigma = \sqrt{24} = 4.90,$$

roughly meaning that although on average one expects approximately 40 successes, one also expects each result to deviate from 40 by an average of approximately five successes.

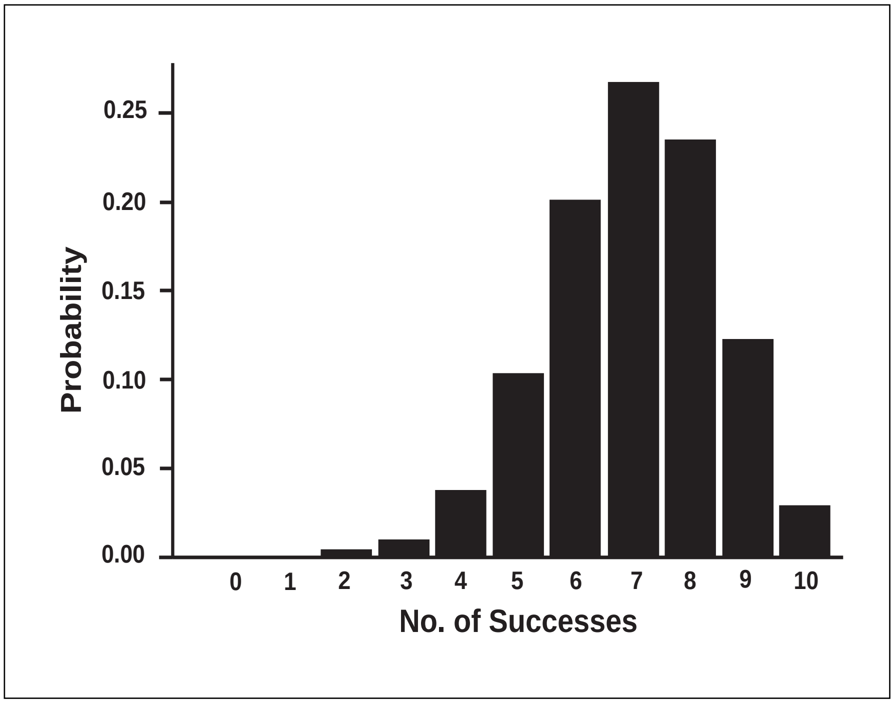


Fig. 1.—Graph shows binomial distribution with sample size of 10 and probability of success $p = 0.7$.

The binomial distribution can be used any time one has a series of independent trials (different patients in any trial can usually be considered as independent) wherein the probability of success remains the same for each patient. For example, suppose that one has a series of 100 patients, all with known endometrial cancer. If each patient is asked to undergo MR imaging, for example, and if the true sensitivity of this test is 80%, what is the probability that 80 of them will in fact test positive? By plugging $p = 0.8$, $n = 100$, and $x = 80$ into the binomial probability formula as discussed, one finds that this probability is 0.0993, or about 10%. (One would probably want to do this calculation on a computer because 100!, for example, would be a tedious calculation.)

Normal Distribution

Perhaps the most common distribution used in statistical practice is the normal distribution, the familiar bell-shaped curve, as seen in Figure 2. Many clinical measurements follow normal or approximately normal distributions (e.g., tumor sizes). Technically, the curve is traced out by the normal density function

$$\frac{1}{\sqrt{2\pi\sigma}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\},$$

where "exp" denotes the exponential function to the base $e = 2.71828$. The Greek letter μ is the mean of the normal distribution set to zero in the SD curve of Figure 2, and the SD is σ, set to 1 in the standard normal curve. Although Figure 2 presents the standard version of the normal curve ($\mu = 0$, $\sigma^2 = \sigma = 1$), more generally, the mean μ can be any real number and the SD can be any number greater than zero. Changing the mean shifts the curve depicted in Figure 2 to the left or right so that it remains centered at the mean, whereas changing the SD stretches or shrinks the curve around the mean, all while keeping its bell shape. Note that the mean (usual arithmetic average), median (middle value, i.e., point at which 50% of the area under the curve lies above and below), and mode (most likely value, i.e., highest point on the curve) of a normal distribution are always the same and equal to μ. Approximately 95% of the area under the curve falls within 2 SDs on either side of the mean, and approximately 68% of the area falls within 1 SD of the mean.

The normal density function has been used to represent the distribution of many measures in medicine. For example, tumor size, biparietal diameter, or bone mineral density in a given population may be said to follow a normal distribution with a given mean and SD. It is highly unlikely that any of these or other quantities exactly follow a normal distribution. For instance, none of these quantities can have negative numbers, whereas the range of the normal distribution always includes all negative (and all positive) numbers. Nevertheless, for appropriately chosen mean and SD, the probability of out-of-range numbers will be vanishingly small, so that this may be of little concern in practice. We may say, for example, that tumor size in a given population follows a normal distribution with a mean of 20 mm and an SD of 10 mm, so that the probability of a value less than zero is only approximately 2.5%. In the words of statistician George Box [3], "All models are wrong, but some are useful."

To calculate probabilities associated with the normal distribution, one must find the area under the normal curve. Because doing so is mathematically difficult, normal tables or a computer program are usually used. For example, the area under the standard normal curve between –1 and 2 is 0.8186, as calculated via normal tables or via a computer package for statistics.

The normal distribution is central to statistical inference for an additional reason. Consider taking a random sample of 500 patients visiting their family physicians for periodic health examinations. If the blood pressure of each patient were recorded and an average were taken, one could use this value as an estimate of the average in the population of all patients who might visit their family physicians for routine checkups. However, if the experiment were repeated, it would be unexpected for the second average of 500 patients to be identical to the first average, although one could expect it to be close.

How these averages vary from one sample to another is given by the central limit theorem, which in its simplest form is explained as follows. Suppose that a population characteristic has true (but possibly unknown) mean μ and standard deviation σ. The distribution of the sample average, $\bar{x}$, based on a sample of size n, approaches a normal distribution as the sample size grows large, with mean μ and SD $\sigma / \sqrt{n}$. As will be explained in future articles, the sample average, $\bar{x}$, is used to estimate the true (but unknown) population mean μ. The SD about a sample mean, $\sigma / \sqrt{n}$, is often called the standard error (SE).

This useful theorem has two immediate consequences. First, it accounts for the popularity of the normal distribution in statistical practice. Even if an underlying distribution in a population is nonnormal (e.g., if it is skewed or binomial), the distribution of the sample average from this population becomes close to normal if the sample size is large enough. Thus, statistical inferences can often be based on the normal distribution, even if the underlying population distribution is nonnormal. Second, the result connects the sample mean to the population mean, forming the basis for much of the statistical inference. In particular, notice that as the sample size n increases, the SD (SE) $\sigma / \sqrt{n}$ of the sample mean around the true mean decreases so that on average the sample mean $\bar{x}$ gets closer and closer to μ. We return to this important point later, but first look at our last distribution, the Poisson.

Poisson Distribution

Suppose that we would like to calculate probabilities relating to numbers of cancers over a given period of time in a given population. In principle, we can consider using a binomial distribution because we are talking about numbers of events in a given number of trials. However, the numbers of events may be enormous (number of persons in the population times the num-

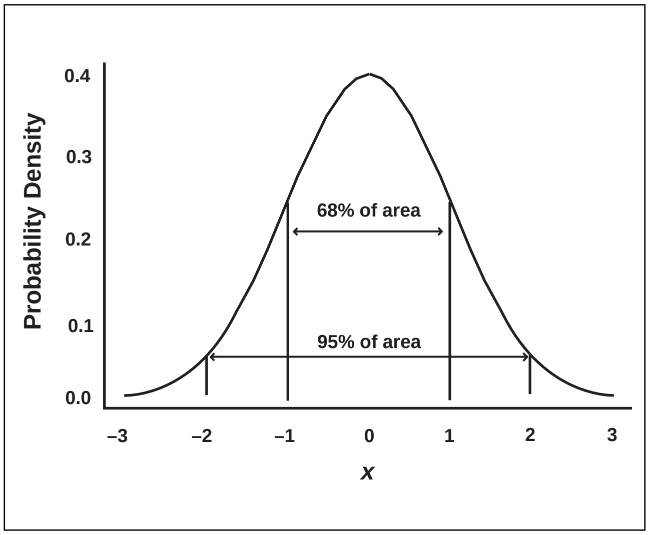


Fig. 2.—Graph shows the standard normal distribution with mean μ = 0 and SD σ = 1. Approximately 95% of area under curve falls within 2 SDs on either side of mean, and about 68% of area falls within 1 SD from mean.

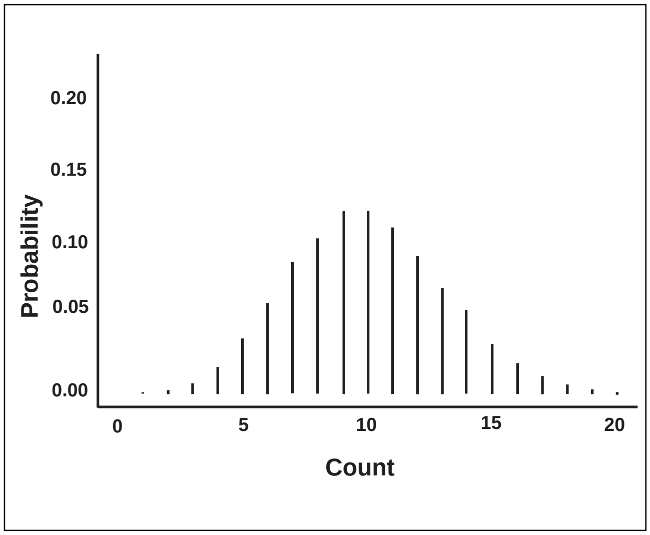


Fig. 3.—Chart shows the Poisson distribution with μ (mean = 10).

ber of time periods). Furthermore, we may not even be certain of the denominator but may have some idea of the rate (e.g., per year) of cancers in this population from previous data. In such cases in which we have counts of events through time rather than counts of successes in a given number of trials, we can consider using the Poisson distribution. More precisely, we make the following assumptions:

First, we assume that the probability of an event (e.g., a cancer) is proportional to the time of observation. We can notate this as Pr (cancer occurs in time t) = $\lambda \times t$, wherein λ is the rate parameter, indicating the event rate in units of events per time. Second, we assume that the time t is small enough that two events cannot occur in time t. For cancer in a population, t may be, for example, 1 min. The event rate λ is assumed to be constant through time (homogeneous Poisson process). Finally, we assume that events (cancers) occur independently.

If all of these assumptions are true, then we can derive the distribution of the number of counts in any given period of time. Let $\mu = \lambda \times t$ be the rate multiplied by time, which is the Poisson mean number of events in time t. Then the Poisson distribution is given by

$$Pr\,(x \text{ events occur in time } t) =$$

$$\frac{e^{-\mu}\,\mu^{x}}{x!\,,}$$

where e = 2.71828. . . , and x denotes factorial of x (the same as in the binomial distribution). Both the mean and the variance of the Poisson distribution are equal to μ. The graph of the Poisson distribution for $\mu = 10$ is given in Figure 3.

As an example of the use of the Poisson distribution, suppose that the incidence of a certain type of cancer in a given region is 250 cases per year. What is the probability that there will be exactly 135 cancer cases in the next 6 months? Let t =1 year, then μ = 250 cancers per year. We are interested, however, in t = 0.5, which means that μ = 125 cancers per 6-month period. Using the Poisson distribution, we can calculate

$$Pr\,(135 \text{ cancers} \mid \mu = 125) =$$

$$\frac{e^{-125}\,125^{135}}{135!}$$

$$= 0.0232.$$

Therefore, approximately 2.3% of a chance exists of observing 135 cancers in the next 6 months.

Summary

The binomial distribution is used for yes/no or success/fail dichotomous variables, the normal distribution is often used for probabilities concerning continuous variables, and the Poisson distribution is used for outcomes arising from counts. These three distributions, of course, are by no means the only ones available, but they are among the most commonly used in practice. Deciding whether they are appropriate in any given situation requires careful consideration of many factors and verification of the assumptions behind each distribution and its use.

This ends our brief tour of the world of probability and probability distributions. Armed with these basics, we are now ready to consider some simple statistical inferences.

Sampling Distributions

So far, we have seen the definitions of probability, the rules probabilities must follow, and three probability distributions. These ideas form the basis for statistical inferences, but how? The key is sampling distributions.

First, we must distinguish sampling distributions from probability distributions and population distributions, which can be explained through an example: Suppose we would like to measure the average tumor size on detection at MR imaging for a certain type of cancer. If we were able to collect the tumor size for all patients with this disease (i.e., a complete census) and create a histogram of these values, then these data would represent the population distribution. The mean of this distribution would represent the true average tumor size in this population.

It is rare, if not impossible, for anyone to perform a complete census, however. One will usually have the opportunity to observe only a subset of the subjects in the target population (i.e., a sample). Suppose that we are able to take a random sample of subjects from this population, of, for example, n = 100 patients. In each case, we observe the tumor size and record the average value. Suppose this average value is $\bar{x}$ = 20 mm, with a SD of σ = 10 mm. We can thus conclude that 20 mm, the average value in our sample, is a reasonable (unbiased) point estimate of the average tumor value in our population, but how accurate is it? How does this accuracy vary if we change the sample size to only 10 patients? What about if we increase it to 1000 patients?

The answer to these questions lies in the sampling distribution of the estimator, $\bar{x}$. First of all, what is the sampling distribution of $\bar{x}$? Suppose we were to take a second random sample of size 100 and record its mean. It would not likely be exactly 20 mm but perhaps be close to that value, for example, 18 mm. If we repeated this process for a third sample, we might get a mean of 21 mm, and so on. Now imagine the thought experiment in which we would repeat this process an infinite number of times and draw the histogram of these means of 100 subjects. The resulting histogram would represent the sampling distribution of $\bar{x}$ for this problem.

According to the central limit theorem, the sampling distribution of $\bar{x}$ is a normal distribution, with mean μ representing the true but unknown mean tumor size (available only if a complete census is taken), and with an SE $\sigma / \sqrt{n}$. Therefore, the SE in our example is $10 / \sqrt{100}$ = 1 mm. So the sampling distribution of $\bar{x}$ is normal, with unknown mean μ, and SE of 1. Although we do not know the mean of the sampling distribution, we do know, from our facts about the normal distribution, that 95% of all $\bar{x}$'s sampled in this experiment will be within ± 2 × 1 = 2 SEs from μ. Thus, although μ remains unknown, we do expect it to be near $\bar{x}$ in this sense. Chances are very good that $\bar{x}$ will be within 2 mm of μ, allowing statements called confidence intervals about μ that we will examine more closely in subsequent articles in this series. If we observed only 10 tumors rather than 100, our SE would have been $10 / \sqrt{10}$ = 3.2 mm, leading to less accuracy in estimating μ, whereas a sample size of 1000 would lead to an SE of 0.32, leading to increased accuracy compared with a size of 100.

To summarize, population distributions represent the spread of values of the variable of interest across individuals in the target population, whereas sampling distributions show how the estimate of the population mean varies from one sample to the next if the experiment were to be repeated and the mean calculated each time. The sampling distribution connects the estimator, here $\bar{x}$, to the parameter of interest, here μ, the mean tumor size in the population. Larger sample sizes lead to more accurate estimation.

Similar inferences can be made from observations that are dichotomous using the binomial distribution or for count data using the Poisson distribution. Again, these topics are relegated to a future article in this series.

Notice that we had to make various assumptions in the previous discussion—for example, that the distribution of tumor sizes in the population is approximately normal and, most importantly, that the subjects are representative of the population to whom we wish to make inferences. The easiest way to ensure representativeness is through random selection, but this may not be possible in some situations for practical reasons. For true random selection to occur, one must have a list of all members of the population and select subjects to form the study sample by random number generation or another random process. Lists of all members of the target population are rare, however, so that different mechanisms of subject selection are often necessary. Case series, or consecutive patients in a clinic, may or may not be representative, depending on the particularities of the selection process. Similarly, convenience samples—taking the subjects most easily available—are often not completely representative, because the very fact that subjects are easily available often tends to make them younger, less sick, and living near the clinic.

Because many outcomes of interest may differ between, for example, young and old or urban and rural patients, convenience samples and often case series are always suspect in terms of selection bias. In other words, although a tumor size of 20 mm may in fact be the average in your sample, this estimate is biased if patients with smaller or larger tumors are systematically left out. For example, subjects with preclinical symptoms may not visit your clinic, even if their tumors might have been detectable on MR imaging, resulting in 20 mm being an overestimate of the true average tumor size detectable on MR imaging in the clinic. Similarly, if patients with advanced disease do not visit the clinic because their tumors were clinically detected by other means, 20 mm may in fact be an underestimate of the true average. Selection bias should always be kept in mind when reading the medical literature.

Conclusion

This brief tour of probability, distributions, and the roots of statistical inferences barely scratches the surface. Many of these ideas will be amplified in future articles of this series. For the impatient, or those who want more detailed explanations of the concepts presented here, countless books explain basic statistical concepts—dozens with a focus on biostatistics. Among them are the works of Armitage and Berry [4], Colton [5], Rosenberg et al. [6], and Rosner [7].

References

1. Last J. *A dictionary of epidemiology,* 2nd ed. New York: Oxford University Press, **1988**:xx
2. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *JAMA* **1995**;273:871–875
3. Box G. *Statistics for experimenters: an introduction to design, data analysis, and model building.* New York: Wiley, **1978**
4. Armitage P, Berry G. *Statistical methods in medical research,* 3rd ed. Oxford: Blackwell Scientific Publications, **1994**
5. Colton T. *Statistics in medicine.* Boston: Little, Brown, **1974**
6. Rosenberg L, Joseph L, Barkun A. *Surgical arithmetic: epidemiological, statistical and outcome-based approach to surgical practice.* Austin, TX: Landes Biosciences, **2000**
7. Rosner B. *Fundamentals of biostatistics.* Belmont, CA: Duxbury, **1994**:105

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003

Observational Studies in Radiology

C. Craig Blackmore[1]
Peter Cummings[2]

The objectives of this paper are to describe the commonly used observational study designs—cohort and case-control studies—and to illustrate their use in radiology research. We will also discuss the strengths and limitations of observational studies and the basics of data analysis. Comprehensive discussions of observational studies can be found in several epidemiology textbooks [1–4].

An important goal in radiology research is to estimate any causal effect of radiology interventions on patient outcome [5–7]. For example, a number of investigators have studied the effect of mammography screening on mortality due to breast cancer [8–10]. A second goal of radiology research is to provide evidence to guide selection of optimal imaging strategies. Associations between clinical factors and diseases can form the basis of clinical prediction rules that guide development of imaging strategies [11, 12]. For example, mechanism of injury, such as a high-speed motor vehicle crash, is a predictor of cervical spine fracture that can be used to select between CT or radiography to evaluate the cervical spine of trauma patients [13, 14].

The best research design for the investigation of causal relationships is the randomized clinical trial. However, clinical trials require that the investigator control which subjects receive a given treatment or exposure under study. Many circumstances exist in which it is not ethical or feasible to perform a randomized trial. For example, we cannot study the influence of cervical spine imaging on outcome in major trauma patients by randomizing which trauma patients will have their cervical spines imaged and which will not. In general, it may not be appropriate to perform a randomized trial if the exposure cannot be manipulated, if manipulation of the exposure would be unethical, if the time from exposure to outcome is very long and more immediate results are desired, or if the outcome is rare, requiring a prohibitively large and expensive randomized clinical trial. Under these circumstances, observational studies may be the best alternatives. Observational studies, including cohort and case-control studies, are hypothesis-testing analytic studies that do not require manipulation of an exposure [15].

Cohort Studies

The most intuitively understood observational study is a cohort study, in which outcomes of subjects with and without a given exposure are compared. A well-known radiology cohort study is the comparison of high- and low-osmolar contrast media by Bettmann et al. [16]. In that study, the outcomes were adverse events that could be attributed to the contrast media. Outcomes were assessed prospectively, meaning that subjects were identified at the time of exposure (use of contrast material), and then followed up to see if the outcome (adverse reaction) occurred. Bettmann et al. found that use of low-osmolar contrast material was associated with a decreased rate of all adverse reactions. Cohort studies may also be retrospective; exposed and unexposed subjects are identified retrospectively after all outcomes of interest have occurred. Both exposure and outcome are then determined from medical records or some other data source.

In cohort studies, the rate of the outcome for each of the exposure cohorts is measured di-

Received May 24, 2004; accepted after revision June 2, 2004.

Supported in part by the Agency for Healthcare Research and Quality grant K08 HS11291-02.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 11th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment. Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Department of Radiology, Harborview Medical Center and Harborview Injury Prevention and Research Center, University of Washington, Box 359728, 325 Ninth Ave., Seattle, WA 98104. Address correspondence to C. C. Blackmore.

[2]Department of Epidemiology, Harborview Medical Center and Harborview Injury Prevention and Research Center, University of Washington, Seattle, WA.

AJR 2004;183:1203–1208

0361–803X/04/1835–1203

TABLE 1 Two-by-Two Table for a Cohort Study

Group	Outcome		Total
	Yes	No	
Exposed	*a*	*b*	[*a* +*b*]
Unexposed	*c*	*d*	[*c* + *d*]
Risk ratio	[*a* / (*a* + *b*)] / [*c* / (*c* + *d*)]		

TABLE 2 Cohort Study Comparing Reaction Rates Using Low-Osmolar Versus High-Osmolar Intraarterial Contrast Media (Risk Ratio = 0.71)

Group	Reaction	No Reaction	Total
Low-osmolar	[*a*] 942	[*b*] 8,482	[*a* +*b*] 9,424
High-osmolar	[*c*] 1,601	[*d*] 9,833	[*c* + *d*] 11,434
Risk ratio			
Formula	[*a* / (*a* + *b*)] / [*c* / (*c* + *d*)]		
Result	(942 / 9,424) / (1,601 / 11,434) = 0.71		

Note.—Derived from Bettmann et al. [16].

rectly. The groups are often compared using the risk ratio. Using notation from the 2 × 2 contingency table (Table 1), the risk ratio is computed as:

$$risk\ ratio = [a / (a + b)] / [c / (c + d)] = p_1 / p_2,$$

where p_1 is the probability of the outcome in subjects with the exposure and p_2 is the probability of outcome in subjects without the exposure.

The risk ratio provides an estimate of the strength of association between the exposure and outcome. Risk ratios may be greater than 1, indicating positive association between outcome and exposure, or less than 1, indicating that a given exposure is associated with a decreased risk of the outcome. Confidence intervals (CIs) for the risk ratio are described in detail elsewhere [1].

The study by Bettmann et al. [16] compared the intraarterial use of low-osmolar contrast material with intraarterial high-osmolar contrast material in diagnostic procedures. When compared with high-osmolar contrast material, low-osmolar contrast material was associated with a lower rate of adverse events, with an unadjusted risk ratio of 0.71 (95% CI, 0.67, 0.75) (Table 2) [16].

An advantage of a cohort study is that a single cohort may be used to study multiple outcomes. Bettmann et al. [16] investigated the rate of all adverse events after contrast administration. However, they were also able to investigate the rates of major reactions and minor reactions in the same patients. A disadvantage of a cohort study is that because subjects are selected on the basis of exposure; usually only a single exposure can be studied.

Case-Control Studies

In case-control studies, subjects are selected on the basis of their outcomes. Cases are those with the outcome being studied, and controls are subjects selected, often at random, from the population from which the cases arose. Exposure is then assessed for both the cases and the controls. Case-control studies may be used to study the impact of an imaging technique on patient outcome. For example, Moss et al. [8] used case-control methods to evaluate the impact of mammography screening on mortality due to breast cancer. Cases were subjects who died from breast cancer, and controls were age-matched women who survived in the Guilford and Stoke region of the United Kingdom. Women invited for breast cancer mammography screening as part of the Trial of Early Detection of Breast Cancer were considered to be exposed. Unexposed subjects were those not invited for screening. Being invited to screening was associated with decreased breast cancer–related mortality.

The analysis of case-control study data can also be illustrated using a 2 × 2 table (Table 3). However, the relevant measure of association is the odds ratio:

$$odds\ ratio = a \times d / b \times c = [p_1 / (1 - p_1)] / [p_2 / (1 - p_2)],$$

where p is the probability of the outcome, and p / 1 – p is the odds of the outcome. Like the risk ratio, the statistical significance of the odds ratio can be estimated using the chi-square statistic. Confidence intervals for odds ratios are described elsewhere [1].

TABLE 3 Two-by-Two Table for Case-Control Study

Group	Case	Control
Exposed	*a*	*b*
Unexposed	*c*	*d*
Odds ratio	(*a* / *c*) / (*b* / *d*) = *ad* / *bc*	

TABLE 4 Cohort Study of Mammography Screening and Mortality due to Breast Cancer (Risk Ratio = 0.74)

Group	Mortality due to Breast Cancer	Alive, or Death from Other Cause	Total
Offered screening	[*a*] 51	[*b*] 22,647	[*a* + *b*] 22,698
Not offered screening	[*c*] 147	[*d*] 48,324	[*c* + *d*] 48,471
Risk ratio			
Formula	[*a* / (*a* + *b*)] / [*c* / (*c* + *d*)]		
Result	(51 / 22,698) / (147 / 48,471) = 0.74		

Note.—Derived from Moss et al. [8].

TABLE 5 Case-Control Study of Mammography Screening and Mortality due to Breast Cancer (Odds Ratio = 0.75)

Group	Mortality due to Breast Cancer	Alive, or Death from Other Cause
Offered screening	[*a*] 51	[*b*] 312
Not offered screening	[*c*] 147	[*d*] 678
Odds ratio		
Formula	*ad* / *bc*	
Result	(51) (678) / (312) (147) = 0.75	

Note.—Derived from Moss et al. [8].

The case-control study design has several advantages. Case-control studies are a cost-efficient research design, particularly when the outcome under study is rare. Also, in a case-control study, multiple exposures may be studied from the same data. As an example, CT rather than radiography may be the more cost-effective imaging strategy in trauma patients with a high probability of cervical spine fracture [17]. Therefore, identification of subjects at high probability of fracture can aid appropriate selection of CT versus radiography. Blackmore et al. [14] performed a case-control study to identify factors that were associated with cervical spine fractures. Cases were those with a cervical spine fracture, and controls were randomly selected trauma patients without a cervical fracture. This single set of cases and controls was then used to simultaneously assess any association between cervical spine fracture (outcome) and a host of potential predictors (exposures), including mechanism of injury, presence of associated injuries such as head injury, and clinical findings such as neurologic deficits [14].

A disadvantage of case-control studies is that they yield the odds ratio rather than the risk ratio. The risk ratio from a cohort study has a more intuitive interpretation and is generally preferred, because the risk ratio directly compares the proportion of subjects with the outcome in the exposed group with the proportion of subjects with the outcome in the unexposed group. In case-control studies, the proportion of subjects with the outcome in the exposed and unexposed groups is generally not known, so the analysis is based on the odds of the outcome. Fortunately, when the study outcome is rare in the population from which the cases and controls are drawn, the odds ratio will provide a good approximation of the risk ratio. In cohort studies, the risk ratio is $[a / (a + b)] / [c / (c + d)]$ (Table 1). However, for rare outcomes, the contribution of the subjects with the outcome in the denominators becomes small; that is, a and c are small compared with b and d, respectively. The risk ratio becomes approximately $(a / b) / (c / d)$. This in turn reduces to $a \times d / b \times c$, which is equal to the odds ratio derived from the case-control study (Table 3).

The relationship between the odds ratio and the risk ratio, as well as a comparison between case-control and cohort studies, is shown in the breast cancer paper by Moss et al. [8]. In that paper, the authors report on both a case-control study and a cohort study that were performed simultaneously in the same population, in order to compare the two designs. Tables 4 and 5 illustrate the 2 × 2 tables for the two study designs. The risk ratio using the cohort data was [51 / (51 + 22,647)] / [147 / (147 + 48,324)], or 0.74 (95% CI, 0.54, 1.02). The odds ratio using the case-control approach for this study was approximately the same, (51) × (678) / (312) × (147), or 0.75 (95% CI, 0.52, 1.08). Note that the number of subjects with the outcome of death due to breast cancer was the same for both studies, but the number of subjects without the outcome differed. In the case-control study, several controls were selected for each case. In the cohort study, on the other hand, all the subjects with and without the exposure were included. As a result, there are thousands of subjects in the cohort study, but only several hundred in the case-control study. Because the outcome was rare, the cohort and case-control study results were nearly identical, but many fewer subjects were required under the case-control study design.

When the outcome or disease under study is common, the odds ratio may differ substantially from the risk ratio. For example, the theoretic data presented in Tables 6 and 7 compare two studies that yield an odds ratio of 2.0 for the outcome of death in subjects who received test A compared with those who received test B. When the outcome of death was rare (Table 6), the odds ratio and the risk ratio were both about 2.0. However, when death was common, the same odds ratio of 2.0 corresponds to a risk ratio of only 1.1. Thus, for common diseases or outcomes, the odds ratio may not approximate the risk ratio.

Subject Selection

In case-control studies, bias can arise if the selection of cases and controls is affected by exposure status other than through the influence of the exposure on outcome. Similarly, selection bias can arise in cohort studies if the outcome affects the selection of the exposed or unexposed subjects. A useful approach to avoid selection bias is to define

TABLE 6 **Two-by-Two Table for Cohort Study When the Outcome Is Rare (Odds Ratio = 2.00, Risk Ratio = 1.98)**

Group	Outcome (Death)		Total
	Yes	No	
Exposed (test A)	[*a*] 2	[*b*] 100	[*a* + *b*] 102
Unexposed (test B)	[*c*] 10	[*d*] 1,000	[*c* + *d*] 1,010
Odds ratio			
Formula	*ad* / *bc*		
Result	(2) (1,000) / (100) (10) = 2.00		
Risk ratio			
Formula	[*a* / (*a* + *b*)] / [*c* / (*c* + *d*)]		
Result	(2 / 102) / (10 / 1,010) = 1.98		

TABLE 7 **Two-by-Two Table for Cohort Study with Very Common Outcome (Odds Ratio = 2.00, Risk Ratio = 1.09)**

Group	Outcome (Death)		Total
	Yes	No	
Exposed (test A)	[*a*] 100	[*b*] 10	[*a* + *b*] 110
Unexposed (test B)	[*c*] 1,000	[*d*] 200	[*c* + *d*] 1,200
Risk ratio			
Formula	[*a* / (*a* + *b*)] / [*c* / (*c* + *d*)]		
Result	(100 / 110) / (1,000 / 1,200) = 1.09		
Odds ratio			
Formula	*ad* / *bc*		
Result	(100) (200) / (10) (1,000) = 2.00		

the target clinical population to which the results are expected to be applied. This population represents the ideal study population. Study subjects should be drawn from this target population when possible [4, 18, 19].

For example, a number of studies have been undertaken to define clinical risk factors for cervical spine fracture in order to help guide the care and evaluation of these patients in the emergency department. The target clinical population for these studies consisted of patients who were evaluated in the emergency department for possible cervical spine fracture. In the case-control study by Blackmore et al. [14] described earlier, case and control subjects (regardless of exposure) were selected from among those who presented to the emergency department, including patients who were discharged from the emergency department as well as those who were admitted to the hospital. Head injury was a strong risk factor for cervical spine fracture, with an odds ratio of 10.0 (95% CI, 5.2, 19.1) ($p < 0.0001$).

Other investigators have studied clinical predictors of cervical spine fracture but have used different subject selection criteria, with correspondingly different results [14, 20–22]. In a large cohort study by Williams et al. [22], exposed (i.e., head-injured) and unexposed (i.e., not head-injured) subjects were selected from an inpatient trauma registry. The rate of cervical spine fracture was similar in both groups (risk ratio = 1.1; 95% CI, 0.93, 1.3), suggesting no association between head injury and cervical spine fracture, and conflicting with the results from the study by Blackmore et al. [14].

The different results from these studies can be understood by applying both subject selection strategies to the case-control study data from the study by Blackmore et al. (Tables 8 and 9) [14]. When the controls for this study were selected from all emergency department trauma patients (the clinically relevant target population), the results revealed a strong association between head injury and cervical spine fracture (Table 8). Another approach would have been to select the subjects only from those admitted to the hospital (Table 9). However, admitted subjects had a much greater proportion of head injuries than did the group consisting of all emergency department subjects. This difference was expected, because patients with head injury were almost always admitted, whereas those without head injury were more likely to be discharged from the emergency department. However, the increased proportion of head-injured control subjects in the inpatient study led to an odds ratio of only 1.4 for cervical spine fracture among subjects with head injury when compared with those without head injury. The exposure, head injury, affected whether subjects were admitted and therefore affected whether subjects would be eligible for the study—leading to selection bias when only admitted patients were considered. Thus, to study predictors of cervical spine fracture in emergency department patients, it is most appropriate to select subjects from the target population, emergency department patients.

TABLE 8 Case-Control Study of Head Injury as a Predictor of Cervical Spine Fracture Using Emergency Department Trauma Patients as Cases and Controls (Odds Ratio = 10.0)

Group	Fracture	No Fracture
Head injury	[*a*] 52	[*b*] 13
No head injury	[*c*] 116	[*d*] 291
Odds ratio		
Formula	*ad* / *bc*	
Result	(52) (291) / (13) (116) = 10.0	

Note.—Derived from Blackmore et al. [14].

TABLE 9 Case-Control Study of Head Injury as a Predictor of Cervical Spine Fracture Using Admitted Trauma Patients as Cases and Controls (Odds Ratio = 1.4)

Group	Fracture	No Fracture
Head injury	[*a*] 52	[*b*] 11
No head injury	[*c*] 116	[*d*] 35
Odds ratio		
Formula	*ad* / *bc*	
Result	(52) (35) / (11) (116) = 1.4	

Note.—Derived from Blackmore et al. [14].

Confounding

In randomized clinical trials, the randomization process helps to ensure that, on average, the study groups are alike with respect to all known and unknown confounders [23]. In observational studies, on the other hand, confounding can occur if the groups being compared differ with respect to some factor that is associated with the outcome. For example, in the study of contrast agents by Bettmann et al. [16], subjects with a history of reaction to contrast material were more likely to receive low-osmolar contrast material than were subjects without a history of contrast reaction. Furthermore, those with a history of contrast reaction were more likely to have a new adverse reaction than were those without a history of reaction. Therefore, the group that received low-osmolar contrast material included more persons with a propensity to have a reaction than did the high-osmolar contrast group. Failure to account for history of reaction would bias the risk ratio estimate for adverse outcomes. Thus, a history of contrast reactions confounded the relationship between the type of contrast material and the outcome [16].

Several strategies may mitigate the bias induced by confounding variables. The first is to restrict the study to those subjects with only one level of the potential confounder. In this case, that could mean restricting the study to subjects without a history of contrast reactions. A second strategy is to stratify subjects on the basis of the confounder, create an estimate within each stratum, and then combine results across strata. For the contrast media example used by Bettmann et al. [16], separate analyses could be done for subjects with and without a history of previous contrast reaction. The relative risk estimates for the two strata could then be combined using Mantel-Haenszel techniques described later in this article [24]. Such stratification may be effective for a small number of potential confounders but can become impractical when multiple potential confounders must be considered. A third strategy is to adjust for potential confounders using regression methods. In the results reported for the study by Bettmann et al. [16], adjustment was made for potentially confounding variables in a regression model. The results showed that low-osmolar contrast material was associated with fewer reactions than high-osmolar contrast material was, after accounting for the effects of previous contrast reaction, asthma, steroid pretreatment, race, sex, and other potential confounders [16].

Finally, matching may be used to control for a potentially confounding variable. Matching in a cohort study involves selecting unexposed subjects who have equivalent values of a confounding variable as the exposed subjects. For the contrast media example, a matched cohort study could be designed whereby an unexposed (high-osmolar contrast material) subject was selected who had a history of contrast reaction for each exposed (low-osmolar contrast material) subject who had a previous contrast reaction, and an unexposed subject without contrast reaction selected for each exposed subject without a previous contrast reaction. This matching would control for potential confounding by past reaction to contrast material. Matching can also be used in case-control studies. However, if controls in a case-control study are selected on the basis of the presence of a potential confounder, then the frequency of the potential confounder will no longer be equal in the study controls and the underlying population. Matching in case-control studies can actually introduce bias unless it is accounted for in the analysis using stratification or regression.

Matching has several disadvantages. First, it is not possible to study the effects of the variable that was used for matching. Second, matching can be expensive and difficult. Third, matching may decrease the power of a study if some cases cannot be matched to appropriate controls. In general, matching should be used sparingly, or not at all.

Analysis

The basic analysis for an observational study of a binary exposure and binary outcome can be expressed in 2 × 2 tables. Measures of association—the relative risk for cohort studies and the odds ratio for case-control studies—are derived from the 2 × 2 table as described earlier. However, the 2 × 2 table allows consideration of only a single binary exposure and single binary outcome. Confounding variables may complicate the relationship between exposure and outcome.

The Mantel-Haenszel method allows consideration of one or more potentially confounding variables in assessment of the 2 × 2 table. Separate 2 × 2 tables are constructed for each level of the potentially confounding variable. The numerators and denominators for the odds ratios derived from each 2 × 2 table are then weighted on the basis of the total number of subjects in each and combined. The calculation of the Mantel-Haenszel odds ratio for a case-control study and a calculation of a Mantel-Haenszel version of the risk ratio for cohort study data are provided in Appendix 1 [2, 24]. Methods for determining variance estimates and confidence intervals for the Mantel-Haenszel estimators are explained in detail in standard epidemiology texts [1, 2].

TABLE 10 **Contrast Reaction Rates Using Low-Osmolar Versus High-Osmolar Intraarterial Contrast Media in Subjects with a History of Reaction (Risk Ratio = 0.64)**

Group	Reaction	No Reaction	Total
Low-osmolar	[*a*] 145	[*b*] 892	[*a* +*b*] 1,037
High-osmolar	[*c*] 75	[*d*] 268	[*c* + *d*] 343
Risk ratio			
Formula	[*a* / (*a* +*b*)] / [*c* / (*c* + *d*)]		
Result	(145 / 1,037) / (75 / 343) = 0.64		

Note.—Derived from Bettmann et al. [16].

TABLE 11 **Contrast Reaction Rates Using Low-Osmolar Versus High-Osmolar Intraarterial Contrast Media in Subjects Without a History of Reaction (Risk Ratio = 0.69)**

Group	Reaction	No Reaction	Total
Low-osmolar	[*a*] 797	[*b*] 7,599	[*a* +*b*] 8,396
High-osmolar	[*c*] 1,526	[*d*] 9,564	[*c* + *d*] 11,090
Risk ratio			
Formula	[*a* / (*a* +*b*)] / [*c* / (*c* + *d*)]		
Result	(797 / 8,396) / (1,526 / 11,090) = 0.69		

Note.—Derived from Bettmann et al. [16].

As an example, from the contrast study by Bettmann et al. [16], it is possible to use the Mantel-Haenszel risk ratio to account for any effect of previous contrast reaction on determination of the association between low-osmolar contrast media and any adverse reaction. Separate 2 × 2 tables for subjects with and without previous contrast reactions are shown in Tables 10 and 11. These tables are combined using the Mantel-Haenszel method to yield a Mantel-Haenszel risk ratio of 0.69, slightly lower than the crude estimate of risk ratio = 0.71.

Analyses involving multiple confounders, and analyses involving multiple exposures or outcomes, may be analyzed using regression techniques. Regression allows estimation of the odds ratio or risk ratio associated with a given variable after accounting for the effects of all other variables in the model [25, 26]. Logistic regression and other regression techniques will be discussed in future articles in this series.

Conclusion

Case-control and cohort study designs are valuable alternatives to randomized clinical trials. These study designs are particularly useful in determining the influence of a radiology intervention on patient outcome, and in determining clinical risk factors for disease, in order to aid determination of optimal imaging strategies. However, radiologists should be aware of the uses, limitations, and techniques of observational study designs.

References

1. Rothman K, Greenland S. *Modern epidemiology*, 2nd ed. Philadelphia, PA: Lippincott, 1998
2. Kelsey J, Whittemore A, Evans A, Thompson W. *Methods in observational epidemiology*. New York, NY: Oxford University Press, 1996
3. Weiss NS. *Clinical epidemiology: the study of outcome of illness*. New York, NY: Oxford, 1996
4. Schlesselman JJ. *Case-control studies*. New York, NY: Oxford University Press, 1982
5. Thornbury JR. Eugene W. Clinical efficacy of diagnostic imaging: love it or leave it. (Caldwell Lecture) *AJR* 1994;162:1–8
6. Blackmore CC, Black WB, Jarvik JG, Langlotz CP. A critical synopsis of the diagnostic and screening radiology outcomes literature. *Acad Radiol* 1999;6[suppl 1]:S8–S18
7. Hillman BJ. Outcomes research and cost-effectiveness analysis for diagnostic imaging. *Radiology* 1994;193:307–310
8. Moss SM, Summerley ME, Thomas BT, Ellman R, Chamberlain JO. A case-control evaluation of the effect of breast cancer screening in the United Kingdom trial of early detection of breast cancer. *J Epidemiol Community Health* 1992;46:362–364

9. Palli D, Del Turco MR, Buiatti E, Ciatto S, Crocetti E, Paci E. Time interval since last test in a breast cancer screening programme: a case-control study in Italy. *J Epidemiol Community Health* 1989;43:241–248
10. Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer* 1977;39: 2772–2782
11. Stiell IG, Greenberg GH, Wells GA, et al. Prospective validation of a decision rule for the use of radiography in acute knee injuries. *JAMA* 1996;275:611–615
12. Hoffman J, Mower W, Wolfson A, Todd K, Zucker M. Validity of a set of clinical criteria to rule out injury to the cervical spine in patients with blunt trauma. *N Engl J Med* 2000;343:94–99
13. Hanson JA, Blackmore CC, Mann FA, Wilson AJ. Cervical spine screening: a decision rule can identify high risk patients to undergo screening helical CT of the cervical spine. *AJR* 2000; 174:713–718
14. Blackmore CC, Emerson SS, Mann FA, Koepsell TD. Cervical spine imaging in patients with trauma: determination of fracture risk to optimize use. *Radiology* 1999;211:759–765
15. Rivara F, Cummings P, Koepsell T, Grossman D, Maier R. *Injury control: a guide to research and program evaluation.* New York, NY: Cambridge University Press, 2001
16. Bettmann MA, Heeren T, Greenfield A, Goudey C. Adverse events with radiographic contrast agents: results of the SCVIR contrast agent registry. *Radiology* 1997;203:611–620
17. Blackmore CC, Ramsey SD, Mann FA, Deyo RA. Cost-effectiveness of cervical spine CT in trauma patients. *Radiology* 1999;212:117–125
18. Eng J, Siegelman SS. Improving radiology research methods: what is being asked and who is being studied? *Radiology* 1997;205:651–655
19. Kazerooni E. Population and sample. *AJR* 2001;177:993–999
20. Cadoux CG, White JD, Hedberg MC. High-yield roentgenographic criteria for cervical spine injuries. *Ann Emerg Med* 1987;16:738–742
21. Sinclair D, Schwartz M, Gruss J, McLellan B. A retrospective review of the relationship between facial fractures, head injuries, and cervical spine injuries. *J Emerg Med* 1988;6:109–112
22. Williams J, Jehle D, Cottington E, Shufflebarger C. Head, facial, and clavicular trauma as a predictor of cervical spine injury. *Ann Emerg Med* 1992;21:70–73
23. Beam CA. Statistically engineering the study for success. *AJR* 2002;179:47–52
24. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies. *J Natl Cancer Inst* 1959;22:719–748
25. Hosmer D, Lemeshow S. A*pplied logistic regression*, 2nd ed. New York, NY: John Wiley and Sons, 2000
26. Kleinbaum DG, Kupper LL, Muller KE. *Applied regression analysis and other multivariate methods*. Belmont, NY: Duxbury, 1988

APPENDIX 1. Calculation of Mantel-Haenszel Odds Ratio and Risk Ratio Estimates

The Mantel-Haenszel odds ratio ($\mathrm{OR_{MH}}$) for a case-control study is derived as follows:

$$\mathrm{OR_{MH}} = \frac{a_1d_1/\,n_1 + a_2d_2\,/\,n_2 + ... + a_id_i\,/\,n_i}{b_1c_1/\,n_1 + b_2c_2\,/\,n_2 + ... + b_ic_i\,/\,n_i}$$

where each stratum of the confounding variable is denoted by the subscript 1 to i, and n_i is the total number of subjects ($a + b + c + d$) in stratum i [2, 24]. The $\mathrm{OR_{MH}}$ can also be expressed as the weighted sum of the stratum odds ratios.

$$\mathrm{OR_{MH}} = \frac{\sum (OR_i)(b_ic_i\,/\,n_i)}{\sum (b_ic_i\,/\,n_i)}$$

A Mantel-Haenszel version of the risk ratio ($\mathrm{RR_{MH}}$) can be calculated for cohort study data [2]:

$$\mathrm{RR_{MH}} = \frac{(a_1)(c_1 + d_1)\,/\,n_1 + (a_2)(c_2 + d_2)\,/\,n_2 + ... + (a_i)(c_i + d_i)\,/\,n_i}{(c_1)(a_1 + b_1)\,/\,n_1 + (c_2)(a_2 + b_2)\,/\,n_2 + ... + (c_i)(a_i + b_i)\,/\,n_i}$$

The $\mathrm{RR_{MH}}$ can also be expressed as the weighted sum of the stratum risk ratios:

$$\mathrm{RR_{MH}} = \frac{\sum (RR_i)\,\{[c_i(a_i + b_i)]\,/\,n_i\}}{\sum \{[c_i(a_i + b_i)]\,/\,n_i\}}$$

The reader’s attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003

Harald O. Stolberg[1]
Geoffrey Norman[2]
Isabelle Trop[3]

Randomized Controlled Trials

Preceding articles in this series have provided a great deal of information concerning research design and methodology, including research protocols, statistical analyses, and assessment of the clinical importance of radiologic research studies. Many methods of research design have already been presented, including descriptive studies (e.g., case reports, case series, and cross-sectional surveys), and some analytical designs (e.g., cohort and case-control studies).

Case-control and cohort studies are also called observational studies, which distinguishes them from interventional (experimental) studies because the decision to seek one treatment or another, or to be exposed to one risk or another, was made by someone other than the experimenter. Consequently, the researcher's role is one of observing the outcome of these exposures. By contrast, in experimental studies, the researcher (experimenter) controls the exposure. The most powerful type of experimental study is the randomized controlled trial. The basic principles of randomized controlled trials will be discussed in this article.

History of Randomized Controlled Trials

The history of clinical trials dates back to approximately 600 B.C. when Daniel of Judah [1] conducted what is probably the earliest recorded clinical trial. He compared the health effects of the vegetarian diet with those of a royal Babylonian diet over a 10-day period. The trial had obvious deficiencies by contemporary medical standards (allocation bias, ascertainment bias, and confounding by divine intervention), but the report has remained influential for more than two millennia [2].

The 19th century saw many major advances in clinical trials. In 1836, the editor of the *American Journal of Medical Sciences* wrote an introduction to an article that he considered "one of the most important medical works of the present century, marking the start of a new era of science," and stated that the article was "the first formal exposition of the results of the only true method of investigation in regard to the therapeutic value of remedial agents." The article that evoked such effusive praise was the French study on bloodletting in treatment of pneumonia by P. C. A. Louis [2, 3].

Credit for the modern randomized trial is usually given to Sir Austin Bradford Hill [4]. The Medical Research Council trials on streptomycin for pulmonary tuberculosis are rightly regarded as a landmark that ushered in a new era of medicine. Since Hill's pioneering achievement, the methodology of the randomized controlled trial has been increasingly accepted and the number of randomized controlled trials reported has grown exponentially. The Cochrane Library already lists more than 150,000 such trials, and they have become the underlying basis for what is currently called "evidence-based medicine" [5].

General Principles of Randomized Controlled Trials

The randomized controlled trial is one of the simplest but most powerful tools of research. In essence, the randomized controlled trial is a study in which people are allocated at random to receive one of several clinical interventions [2]. On most occasions, the term "intervention" refers to treatment, but it should be used in a much wider sense to include any clinical maneuver offered to study participants that may

Received June 14, 2004; accepted after revision July 2, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 12th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment; staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1] Department of Radiology, McMaster University Medical Centre, 1200 Main St. W, Hamilton, ON L8N 3Z5, Canada. Address correspondence to H. O. Stolberg.

[2] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON L8N 3Z5, Canada.

[3] Department of Radiology, Hôpital Saint-Luc, 1058 St. Denis St., Montreal, QC H2X 3J4, Canada.

AJR 2004;183:1539–1544

0361–803X/04/1836–1539

have an effect on their health status. Such clinical maneuvers include prevention strategies, screening programs, diagnostic tests, interventional procedures, the setting in which health care is provided, and educational models [2]. Randomized controlled trials in radiology can play a major role in the assessment of screening programs, diagnostic tests, and procedures in interventional radiology [6–13].

Randomized controlled trials are used to examine the effect of interventions on particular outcomes such as death or the recurrence of disease. Some consider randomized controlled trials to be the best of all research designs [14], or "the most powerful tool in modern clinical research" [15], mainly because the act of randomizing patients to receive or not receive the intervention ensures that, on average, all other possible causes are equal between the two groups. Thus, any significant differences between groups in the outcome event can be attributed to the intervention and not to some other unidentified factor. However, randomized controlled trials are not a panacea to answer all clinical questions; for example, the effect of a risk factor such as smoking cannot ethically be addressed with randomized controlled trials. Furthermore, in many situations randomized controlled trials are not feasible, necessary, appropriate, or even sufficient to help solve important problems [2]. Randomized controlled trials are not appropriate for cancer screening, a situation in which the outcome is rare and frequently occurs only after a long delay. Thus, although the test for appraising the ultimate value of a diagnostic test may be a large well-designed randomized controlled trial that has patient outcomes as the end point [16], the trial should presumably be performed after other smaller studies have examined the predictive value of the test against some accepted standard.

An excellent example of the controversies that can arise with randomized controlled trials is an overview of the publications on mammography screening. The most important references concern the article by Miettinen et al. [17] linking screening for breast cancer with mammography and an apparently substantial reduction in fatalities and the responses that it elicited [18–22].

Randomized controlled trials may not be appropriate for the assessment of interventions that have rare outcomes or effects that take a long time to develop. In such instances, other study designs such as case-control studies or cohort studies are more appropriate. In other cases, randomized controlled trials may not be feasible because of financial constraints or because of the expectation of low compliance or high drop-out rates.

Many randomized controlled trials involve large sample sizes because many treatments have relatively small effects. The size of the expected effect of the intervention is the main determinant of the sample size necessary to conduct a successful randomized controlled trial. Obtaining statistically significant differences between two samples is easy if large differences are expected. However, the smaller the expected effect of the intervention, the larger the sample size needed to be able to conclude, with enough power, that the differences are unlikely to be due to chance. For example, let us assume that we wish to study two groups of patients who will undergo different interventions, one of which is a new procedure. We expect a 10% decrease in the morbidity rate with the new procedure. To be able to detect this difference with a probability (power) of 80%, we need 80 patients in each treatment arm. If the expected difference in effect between the two groups increases to 20%, the number of patient required per arm decreases to 40. Conversely, if the difference between the groups is expected to be only 1%, the study population must increase to 8,000 per treatment arm. The sample size required to achieve power in a study is inversely proportional to the treatment effect squared [23]. Standard formulas are available to calculate the approximate sample size necessary when designing a randomized controlled trial [24–26].

Randomization: The Strength of the Randomized Controlled Trial

The randomization procedure gives the randomized controlled trial its strength. Random allocation means that all participants have the same chance of being assigned to each of the study groups [27]. The allocation, therefore, is not determined by the investigators, the clinicians, or the study participants [2]. The purpose of random allocation of participants is to assure that the characteristics of the participants are as likely to be similar as possible across groups at the start of the comparison (also called the baseline). If randomization is done properly, it reduces the risk of a serious imbalance in known and unknown factors that could influence the clinical course of the participants. No other study design allows investigators to balance these factors.

The investigators should follow two rules to ensure the success of the randomization procedure. They must first define the rules that will govern allocation and then follow those rules strictly throughout the entire study [2]. The crucial issue is that after the procedure for randomization is determined, it should not be modified at any point during the study. There are many adequate methods of randomization, but their common element is that no one should be able to determine ahead of time to which group a given patient will be assigned. Detailed discussion of randomization methods is beyond the scope of this article.

Numerous methods are also available to ensure that the sample of patients is balanced whenever a small predetermined number of patients have been enrolled. Unfortunately, the methods of allocation in studies described as randomized are poorly and infrequently reported [2, 28]. As a result, it is not possible to determine, on most occasions, whether the investigators used proper methods to generate random sequences of allocation [2].

Bias in Randomized Controlled Trials

The main appeal of the randomized controlled trial in health care derives from its potential for reducing allocation bias [2]. No other study design allows researchers to balance unknown prognostic factors at baseline. Random allocation does not, however, protect randomized controlled trials against other types of bias. During the past 10 years, randomized controlled trials have been the subject rather than the tool of important, albeit isolated, research efforts usually designed to generate empiric evidence to improve the design, reporting, dissemination, and use of randomized controlled trials in health care [28]. Such studies have shown that randomized controlled trials are vulnerable to multiple types of bias at all stages of their workspan. A detailed discussion of bias in randomized controlled trials was offered by Jadad [2].

In summary, randomized controlled trials are quantitative, comparative, controlled experiments in which a group of investigators studies two or more interventions by administering them to groups of individuals who have been randomly assigned to receive each intervention. Alternatively, each individual might receive a series of interventions in random order (crossover design) if the outcome can be uniquely associated with each intervention, through, for example, use of a "washout" period. This step ensures that the effects from one test are not carried over to the next one and subsequently affect the independent evaluation of the second test administered. Apart from random allocation to comparison groups, the elements of a randomized controlled trial are no different from those of any other type of prospective, comparative, quantitative study.

Types of Randomized Controlled Trials

As Jadad observed in his 1998 book *Randomised Controlled Trials* [2]:

> Over the years, multiple terms have been used to describe different types of randomized controlled trials. This terminology has evolved to the point of becoming real jargon. This jargon is not easy to understand for those who are starting their careers as clinicians or researchers because there is no single source with clear and simple definitions of all these terms.

The best classification of frequently used terms was offered by Jadad [2], and we have based our article on his work.

According to Jadad, randomized controlled trials can be classified as to the aspects of intervention that investigators want to explore, the way in which the participants are exposed to the intervention, the number of participants included in the study, whether the investigators and participants know which intervention is being assessed, and whether the preference of nonrandomized individuals and participants has been taken into account in the design of the study. In the context of this article, we can offer only a brief discussion of each of the different types of randomized controlled trials.

Randomized Controlled Trials Classified According to the Different Aspects of Interventions Evaluated

Randomized controlled trials used to evaluate different interventions include explanatory or pragmatic trials; efficacy or equivalence trials; and phase 1, 2, 3, and 4 trials.

Explanatory or pragmatic trials.—Explanatory trials are designed to answer a simple question: Does the intervention work? If it does, then the trial attempts to establish how it works. Pragmatic trials, on the other hand, are designed not only to determine whether the intervention works but also to describe all the consequences of the intervention and its use under circumstances corresponding to daily practice. Although both explanatory and pragmatic approaches are reasonable, and even complementary, it is important to understand that they represent extremes of a spectrum, and most randomized controlled trials combine elements of both.

Efficacy or effectiveness trials.—Randomized controlled trials are also often described in terms of whether they evaluate the efficacy or effectiveness of an intervention. Efficacy refers to interventions carried out under ideal circumstances, whereas effectiveness evaluates the effects of an intervention under circumstances similar to those found in daily practice.

Phase 1, 2, 3, and 4 trials.—These terms describe the different types of trials used for the introduction of a new intervention, traditionally a new drug, but could also encompass trials used for the evaluation of a new embolization material or type of prosthesis, for example. Phase 1 studies are usually conducted after the safety of the new intervention has been documented in animal research, and their purpose is to document the safety of the intervention in humans. Phase 1 studies are usually performed on healthy volunteers. Once the intervention passes phase 1, phase 2 begins. Typically, the intervention is given to a small group of real patients, and the purpose of this study is to evaluate the efficacy of different modes of administration of the intervention to patients. Phase 2 studies focus on efficacy while still providing information on safety. Phase 3 studies are typically effectiveness trials, which are performed after a given procedure has been shown to be safe with a reasonable chance of improving patients' conditions. Most phase 3 trials are randomized controlled trials. Phase 4 studies are equivalent to postmarketing studies of the intervention; they are performed to identify and monitor possible adverse events not yet documented.

Randomized Controlled Trials Classified According to Participants' Exposure and Response to the Intervention

These types of randomized controlled trials include parallel, crossover, and factorial designs.

Parallel design.—Most randomized controlled trials have parallel designs in which each group of participants is exposed to only one of the study interventions.

Crossover design.— Crossover design refers to a study in which each of the participants is given all of the study interventions in successive periods. The order in which the participants receive each of the study interventions is determined at random. This design, obviously, is appropriate only for chronic conditions that are fairly stable over time and for interventions that last a short time within the patient and that do not interfere with one another. Otherwise, false conclusions about the effectiveness of an intervention could be drawn [29].

Factorial design.—A randomized controlled trial has a factorial design when two or more experimental interventions are not only evaluated separately but also in combination and against a control [2]. For example, a 2 × 2 factorial design generates four sets of data to analyze: data on patients who received none of the interventions, patients who received treatment A, patients who received treatment B, and patients who received both A and B. More complex factorial designs, involving multiple factors, are occasionally used. The strength of this design is that it provides more information than parallel designs. In addition to the effects of each treatment, factorial design allows evaluation of the interaction that may exist between two treatments. Because randomized controlled trials are generally expensive to conduct, the more answers that can be obtained, the better.

Randomized Controlled Trials Classified According to the Number of Participants

Randomized controlled trials can be performed in one or many centers and can include from one to thousands of participants, and they can have fixed or variable (sequential) numbers of participants.

"N-*of-one trials.*"—Randomized controlled trials with only one participant are called "*n*-of-one trials" or "individual patient trials." Randomized controlled trials with a simple design that involve thousands of patients and limited data collection are called "megatrials." [30, 31]. Usually, megatrials require the participation of many investigators from multiple centers and from different countries [2].

Sequential trials.—A sequential trial is a study with parallel design in which the number of participants is not specified by the investigators beforehand. Instead, the investigators continue recruiting participants until a clear benefit of one of the interventions is observed or until they become convinced that there are no important differences between the interventions [27]. This element applies to the comparison of some diagnostic interventions and some procedures in interventional radiology. Strict rules govern when trials can be stopped on the basis of cumulative results, and important statistical considerations come into play.

Fixed trials.—Alternatively, in a fixed trial, the investigators establish deductively the number of participants (sample size) that will be studied. This number can be decided arbitrarily or can be calculated using statistical methods. The latter is a more commonly used method. Even in a fixed trial, the design of the trial usually specifies whether there will be one or more interim analyses of data. If a clear benefit of one intervention over the other can be shown with statistical significance before all participants are recruited, it may not be ethical to pursue the trial, and it may be prematurely terminated.

Randomized Controlled Trials Classified According to the Level of Blinding

In addition to randomization, the investigators can incorporate other methodologic strategies to reduce the risk of other biases. These strategies are known as "blinding." The purpose of blinding is to reduce the risk of ascertainment and observation bias. An open randomized controlled trial is one in which everybody involved in the trial knows which intervention is given to each participant. Many radiology studies are open randomized controlled trials because blinding is not feasible or ethical. One cannot, for example, perform an interventional procedure with its associated risks without revealing to the patient and the treating physician to which group the patient has been randomized. A single-blinded randomized controlled trial is one in which a group of individuals involved in the trial (usually patients) does not know which intervention is given to each participant. A double-blinded randomized controlled trial, on the other hand, is one in which two groups of individuals involved in the trial (usually patients and treating physicians) do not know which intervention is given to each participant. Beyond this, triple-blinded (blinding of patients, treating physicians, and study investigators) and quadruple-blinded randomized controlled trials (blinding of patients, treating physicians, study investigators, and statisticians) have been described but are rarely used.

Randomized Controlled Trials Classified According to Nonrandomized Participant Preferences

Eligible individuals may refuse to participate in a randomized controlled trial. Other eligible individuals may decide to participate in a randomized controlled trial but have a clear preference for one of the study interventions. At least three types of randomized controlled trials take into account the preferences of eligible individuals as to whether or not they take part in the trial. These are called preference trials because they include at least one group in which the participants are allowed to choose their preferred treatment from among several options offered [32, 33]. Such trials can have a Zelen design, comprehensive cohort design, or Wennberg's design [33–36]. For a detailed discussion of these designs of randomized controlled trials, the reader is directed to the excellent detailed discussion offered by Jadad [2].

The Ethics of Randomized Controlled Trials

Despite the claims of some enthusiasts for randomized controlled trials, many important aspects of health care cannot be subjected to a randomized trial for practical and ethical reasons. A randomized controlled trial is the best way of evaluating the effectiveness of an intervention, but before a randomized controlled trial can be conducted, there must be equipoise—genuine doubt about whether one course of action is better than another [16]. Equipoise then refers to that state of knowledge in which no evidence exists that shows that any intervention in the trial is better than another and that any intervention is better than those in the trial. It is not ethical to build a trial in which, before enrollment, evidence suggests that patients in one arm of the study are more likely to benefit from enrollment than patients in the other arm. Equipoise thus refers to the fine balance that exists between being hopeful a new treatment will improve a condition and having enough evidence to know that it does (or does not). Randomized controlled trials can be planned only in areas of uncertainty and can be carried out only as long as the uncertainty remains. Ethical concerns that are unique to randomized controlled trials as well as other research designs will be addressed in subsequent articles in this series. Hellman and Hellman [37] offered a good discussion on this subject.

Reporting of Randomized Controlled Trials

The Quality of Randomized Controlled Trial Reporting

Awareness concerning the quality of reporting randomized controlled trials and the limitations of the research methods of randomized controlled trials is growing. A major barrier hindering the assessment of trial quality is that, in most cases, we must rely on the information contained in the written report. A trial with a biased design, if well reported, could be judged to be of high quality, whereas a well-designed but poorly reported trial could be judged to be of low quality.

Recently, efforts have been made to improve the quality of randomized controlled trials. In 1996, a group of epidemiologists, biostatisticians, and journal editors published "CONSORT (Consolidated Standards of Reporting Trials)" [38], a statement that resulted from an extensive collaborative process to improve the standards of written reports of randomized controlled trials. The CONSORT statement was revised in 2001 [39]. It was designed to assist the reporting of randomized controlled trials with two groups and those with parallel designs. Some modifications will be required to report crossover trials and those with more than two groups [40]. Although the CONSORT statement was not evaluated before its publication, it was expected that it would lead to an improvement in the quality of reporting of randomized controlled trials, at least in the journals that endorse it [41].

Recently, however, Chan et al. [42] pointed out that the interpretation of the results of randomized controlled trials has emphasized statistical significance rather than clinical importance:

> The lack of emphasis on clinical importance has led to frequent misconceptions and disagreements regarding the interpretation of the results of clinical trials and a tendency to equate statistical significance with clinical importance. In some instances, statistically significant results may not be clinically important and, conversely, statistically insignificant results do not completely rule out the possibility of clinically important effects.

Limitations of the Research Methods Used in Randomized Controlled Trials

The evaluation of the methodologic quality of randomized controlled trials is central to the appraisal of individual trials, the conduct of unbiased systematic reviews, and the performance of evidence-based health care. However, important methodologic details may be omitted from published reports, and the quality of reporting is, therefore, often used as a proxy measure for methodologic quality. High-quality reporting may hide important differences in methodologic quality, and well-conducted trials may be reported badly [43]. As Devereaux et al. [41] observed, "[h]ealth care providers depend upon authors and editors to report essential methodological factors in randomized controlled trials (RCTs) to allow determination of trial validity (i.e., likelihood that the trials' results are unbiased)."

The most important limitations of research methods include the following:

Insufficient power.—A survey of 71 randomized controlled trials showed that most of these trials were too small (i.e., had insufficient power to detect important clinical differences) and that the authors of these trials seemed unaware of these facts [44].

Poor reporting of randomization—A study of 206 randomized controlled trials showed that randomization, one of the main design features necessary to prevent bias in randomized controlled trials, was poorly reported [45].

Other limitations.—Additional limitations identified by Chalmers [46] were inadequate randomization, failure to blind the assessors to the outcomes, and failure to follow up all patients in the trials.

Intent to Treat

A method to correct for differential dropout rates between patients from one arm of the study and another is to analyze data by the intent to treat—that is, data are analyzed in the way patients were randomized, regardless of whether or not they received the intended intervention. The intent to treat correction is a form of protection against bias and strengthens the conclusions of a study. A detailed discussion of the assessment of the quality of randomized controlled trials was offered by Jadad [2].

In the appraisal of randomized controlled trials, a clear distinction should be made between the quality of the reporting and the quality of methodology of the trials [43].

Recent Randomized Controlled Trials in Radiology

In recent years, randomized controlled trials have become increasingly popular in radiology research. In 1997, for instance, there were only a few good randomized studies in diagnostic imaging, such as the one by Jarvik et al. [47]. Since 2000, the number of good randomized controlled trials has significantly increased in both diagnostic and interventional radiology. Examples of randomized controlled trials in diagnostic imaging include the works of Gottlieb et al. [48] and Kaiser et al. [49]. Examples of interventional randomized controlled trials are the studies by Pinto et al. [50] and Lencioni et al. [51].

Randomized controlled trials are equally important in screening for disease. Our initial experience with breast screening was unfortunate, and controversy over this issue continues to this day [52, 53]. On the other hand, positive developments have occurred, such as the work of the American College of Radiology Imaging Network. Writing for this group, Berg [54] has offered a commentary on the rationale for a trial of screening breast sonography.

Radiologists have a great deal to learn about randomized controlled trials. Academic radiologists who perform research and radiologists who translate research results into practice should be familiar with the different types of these trials, including those conducted for diagnostic tests and interventional procedures. Radiologists also must be aware of the limitations and problems associated with the methodologic quality and reporting of the trials. It is our hope that this article proves to be a valuable source of information about randomized controlled trials.

Acknowledgments

We thank Alejandro Jadad for his support and Monika Ferrier for her patience and support in keeping us on track and for preparing the manuscript.

References

1. Book of Daniel 1:1–21
2. Jadad AR. *Randomised controlled trials: a user's guide*. London, England: BMJ Books, 1998
3. Louis PCA. Research into the effects of bloodletting in some inflammatory diseases and on the influence of tartarized antimony and vesication in pneumonitis. *Am J Med Sci* 1836;18:102–111
4. Hill AB. The clinical trial. *N Engl J Med* 1952; 247:113–119
5. Cochrane Library Web site. Available at: www.update-software.com/cochrane. Accessed September 10, 2004
6. Bree RL, Kazerooni EA, Katz SJ. Effect of mandatory radiology consultation on inpatient imaging use. *JAMA* 1996;276:1595–1598
7. DeVore GR. The routine antenatal diagnostic imaging with ultrasound study: another perspective. *Obstet Gynecol* 1994;84:622–626
8. Fontana RS, Sanderson DR, Woolner LB, et al. Screening for lung cancer: a critique of the Mayo Lung Project. *Cancer* 1991;67[suppl 4]:1155–1164
9. [No authors listed]. Impact of follow-up testing on survival and health-related quality of life in breast cancer patients: a multicenter randomized controlled trial—the GIVIO Investigators. *JAMA* 1994;271:1587–1592
10. Jarvik JG, Maravilla KR, Haynor DR, Levitz M, Deyo RA. Rapid MR imaging versus plain radiography in patients with low back pain: initial results of a randomized study. *Radiology* 1997;204:447–454
11. Kinnison ML, Powe NR, Steinberg EP. Results of randomized controlled trials of low-versus high-osmolality contrast media. *Radiolog*y 1989;170: 381–389
12. Rosselli M, Palli D, Cariddi A, Ciatto S, Pacini P, Distante V. Intensive diagnostic follow-up after treatment of primary breast cancer. *JAMA* 1994; 271:1593–1597
13. Swingler GH, Hussey GD, Zwarenstein M. Randomised controlled trial of clinical outcome after chest radiograph in ambulatory acute lower-respiration infection in children. *Lancet* 1998;351: 404–408
14. Cochrane Library Web site. Available at: www.update-software.com/abstracts/ ab001877.htm. Accessed September 10, 2004
15. Nystrom L, Rutqvist LE, Wall S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. *Lancet* 1993;341: 973–978
16. Duffy SW. Interpretation of the breast screening trials: a commentary on the recent paper by Gotzsche and Olsen. *Breast* 2001;10:209–212
17. Miettinen OS, Henschke CI, Pasmantier MW, Smith JP, Libby DM, Yankelevitz DF. Mammographic screening: no reliable supporting evidence? *Lancet* 2002;359:404–405
18. Tabar L, Vitak B, Chen HHT, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001; 91:1724–1731
19. Hoey J. Does mammography save lives? *CMAJ* 2002;166:1187–1188
20. Norman GR, Streiner DL. *Biostatistics: the bare essentials*, 2nd ed. Hamilton, ON, Canada: B. C. Decker, 2000
21. Silerman WA. Gnosis and random allotment. *Control Clin Trials* 1981;2:161–164
22. Gray JAM. *Evidence-based health care*. Edinburgh, Scotland: Churchill Livingstone, 1997
23. Rosner B. *Fundamentals of biostatistics*, 5th ed. Duxbury, England: Thomson Learning, 2000
24. Norman GR, Streiner DL. *PDQ statistics*, 2nd ed. St. Louis, MO: Mosby, 1997
25. Altman DG, Machin D, Bagant TN, Gardner MJ. *Statistics with confidence*, 2nd ed. London, England: BMJ Books, 2000
26. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;22:122–124
27. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall, 1991
28. Jadad AR, Rennie D. The randomized controlled

trial gets a middle-aged checkup. *JAMA* 1998; 279:319–320

29. Louis TA, Lavori PW, Bailar JC III, Polansky M. Crossover and self-controlled designs in clinical research. In: Bailar JC III, Mosteller F. eds. *Medical uses of statistics*, 2nd ed. Boston, MA: New England Medical Journal Publications, 1992:83–104
30. Woods KL. Megatrials and management of acute myocardial infarction. *Lancet* 1995;346:611–614
31. Charlton BG. Megatrials: methodological issues and clinical implications. *Coll Phys Lond* 1995; 29:96–100
32. Till JE, Sutherland HJ, Meslin EM. Is there a role for performance assessments in research on quality of life in oncology? *Quality Life Res* 1992; 1:31–40
33. Silverman WA, Altman DG. Patient preferences and randomized trials. *Lancet* 1996;347:171–174
34. Zelen M. A new design for randomized clinical trials. *N Engl J Med* 1979;300:1242–1245
35. Olschewski M, Scheurlen H. Comprehensive Cohort Study: an alternative to randomized consent design in a breast preservation trial. *Methods Inf Med* 1985;24:131–134
36. Brewin CR, Bradley C. Patient preferences and randomized clinical trials. *BMJ* 1989;299:684–685
37. Hellman S, Hellman DS. Of mice but not men: problems of the randomized trial. *N Engl J Med* 1991;324:1585–1592
38. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276:7–9
39. Moher D, Schulz KF, Altman DG, CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001;285:1987–1991
40. Altman DG. Better reporting of randomized controlled trials: the CONSORT statement. *BMJ* 1996;313:570–571
41. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Guyatt GH. The reporting of methodological factors in randomized controlled trials and the association with a journal policy to promote adherence to the Consolidated Standards of Reporting Trials (CONSORT) checklist. *Control Clin Trials* 2002;23:380–388
42. Chan KBY, Man-Son-Hing M, Molnar FJ, Laupacis A. How well is the clinical importance of study results reported? an assessment of randomized controlled trials. *CMAJ* 2001;165:1197–1202
43. Huwiler-Müntener K, Jüni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA* 2002;287:2801–2804
44. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of Beta, the type 2 error, and sample size in design and interpretation of randomized controlled trials. *N Engl J Med* 1978;299:690–694
45. Schulz KF, Chalmers I, Hayes RJ, Altman DJ. Empirical evidence of bias: dimensions of the methodologic quality associated with estimates of treatment efforts in controlled trials. *JAMA* 1995;273:408–412
46. Chalmers I. Applying overviews and meta-analysis at the bedside: discussion. *J Clin Epidemiol* 1995;48:67–70
47. Jarvik JG, Maravilla KR, Haynor DR, Levitz M, Deyo RA. Rapid MR imaging versus plain radiography in patients with low back pain: initial results of a randomized study. *Radiology* 1997;204: 447–454
48. Gottlieb RH, Voci SL, Syed L, et al. Randomized prospective study comparing routine versus selective use of sonography of the complete calf in patients with suspected deep venous thrombosis. *AJR* 2003;180:241–245
49. Kaiser S, Frenckner B, Jorulf HK. Suspected appendicitis in children: US and CT–a prospective randomized study. *Radiology* 2002;223:633–638
50. Pinto I, Chimeno P, Romo A, et al. Uterine fibroids: uterine artery embolization versus abdominal hysterectomy for treatment—a prospective, randomized, and controlled clinical trial. *Radiology* 2003;226:425–431
51. Lencioni RA, Allgaier HP, Cioni D, et al. Small hepatocellular carcinoma in cirrhosis: randomized comparison of radio-frequency thermal ablation versus percutaneous ethanol injection. *Radiology* 2003;228:235–240
52. Dean PB. Gotzsche's quixotic antiscreening campaign: nonscientific and contrary to Cochrane principles. *JACR* 2004;1:3–7
53. Gotzsche PC. The debate on breast cancer screening with mammography is important. *JACR* 2004;1:8–14
54. Berg WA. Rationale for a trial of screening breast ultrasound: American College of Radiology Imaging Network (ACRIN) 6666. *AJR* 2003;180: 1225–1228

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004

Fundamentals of Clinical Research for Radiologists

Susan Weinstein[1]
Nancy A. Obuchowski[2]
Michael L. Lieber[2]

Clinical Evaluation of Diagnostic Tests

The evaluation of the accuracy of diagnostic tests and the appropriate interpretation of test results are the focus of much of radiology and its research. In this article, we first will review the basic definitions of diagnostic test accuracy, including a brief introduction to receiver operating characteristic (ROC) curves. Then we will evaluate how diagnostic tests can be used to address clinical questions such as "Should this patient undergo this diagnostic test?" and, after ordering the test and seeing the test result, "What is the likelihood that this patient has the disease?" We will finish with a discussion of some important concepts for designing research studies that estimate or compare diagnostic test accuracy.

Defining Diagnostic Test Accuracy

Sensitivity and Specificity

There are two basic measures of the inherent accuracy of a diagnostic test: sensitivity and specificity. They are equally important, and one should never be reported without the other. Sensitivity is the probability of a positive test result (that is, the test indicates the presence of disease) for a patient with the disease. Specificity, on the other hand, is the probability of a negative test result (that is, the test does not indicate the presence of disease) for a patient without the disease. We use the term "disease" here loosely to mean the condition (e.g., breast cancer, deep venous thrombosis, intracranial aneurysm) that the diagnostic test is supposed to detect. We calculate the test's specificity based on patients without this condition, but these patients often have other diseases.

Table 1 summarizes the definitions of sensitivity and specificity [1]. The table rows give the results of the diagnostic test, as either positive for the disease of interest or negative for the disease of interest. The columns indicate the true disease status, as either disease present or disease absent. True-positives (TPs) are those patients with the disease who test positive. True-negatives (TNs) are those without the disease who test negative. False-negatives (FNs) are those with the disease but the test falsely indicates the disease is *not* present. False-positives (FPs) are those without the disease but the test falsely indicates the presence of disease. Sensitivity, then, is the probability of a TP among patients with the disease (TPs + FNs). Specificity is the probability of a TN among patients without the disease (TNs + FPs).

Consider the following example. Carpenter et al. [2] evaluated the diagnostic accuracy of MR venography (MRV) to detect deep venous thrombosis (DVT). They performed MRV in a group of 85 patients who presented with clinical symptoms of DVT. The patients also underwent contrast venography, which is an invasive procedure considered to provide an unequivocal diagnosis for DVT (the so-called "gold standard test" or "standard of reference"). Of a total of 101 venous systems evaluated, 27 had DVT by contrast venography. All 27 cases were detected on MRV; thus, the sensitivity of MRV was 27/27, or 100%. Of 74 venous systems without DVT, as confirmed by contrast venography, three tested positive on MRV (that is, three FPs). The specificity of MRV was 71/74, or 96% specificity (Table 2).

Combining Multiple Tests

Few diagnostic tests are both highly sensitive and highly specific. For this reason, patients sometimes are diagnosed using two or more tests. These tests may be performed ei-

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 13th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Department of Radiology, University of Pennsylvania Medical Center, Philadelphia, PA 19104. Address correspondence to S. Weinstein.

[2]Departments of Biostatistics and Epidemiology and Radiology, The Cleveland Clinic Foundation, Cleveland, OH 44195.

AJR 2005;184:14–19

0361–803X/05/18405–14

TABLE 1 Defining Sensitivity and Specificity

Test	Disease	
	Present	Absent
+	True-positive (TP)	False-positive (FP)
–	False-negative (FN)	True-negative (TN)

Note.—Sensitivity = TPs/(TPs + FNs), specificity = TNs/(TNs + FPs).

TABLE 2 Sensitivity and Specificity of MRV in 101 Venous Systems

MRV	Deep Venous Thrombosis	
	Present	Absent
+	27	3
–	0	71

Note.—MRV = magnetic resonance venography.

ther in parallel (i.e., at the same time and interpreted together) or in series (i.e., the results of the first test determine whether the second test is performed at all) [3]. The latter has the advantage of avoiding unnecessary tests, but the disadvantage of potentially delaying treatment for diseased patients by lengthening the diagnostic testing period.

Tests can be interpreted in parallel in two ways. The first, called “the OR rule,” yields a positive diagnosis if either test (let’s assume there are two tests) is positive and a negative diagnosis if both tests are negative. That is, if test A and test B are both negative, then the combined result is negative, but if either or both are positive, then the combined result is positive.

The second rule, called “the AND rule,” yields a positive diagnosis only if both tests are positive and a negative diagnosis if either test is negative. That is, if test A and test B are both positive, then the combined result is positive, but if either or both are negative, then the combined result is negative.

Let us denote the sensitivities of the two tests by SE_a and SE_b, and their specificities by SP_a and SP_b. To calculate the sensitivity of the combined test in parallel using the OR rule, the formula is: $SE_a + SE_b - (SE_a \times SE_b)$. Specificity under the OR rule is simply $SP_a \times SP_b$. Conversely, to calculate sensitivity using the AND rule, the formula is: $SE_a \times SE_b$, while specificity under the AND rule is $SP_a + SP_b - (SP_a \times SP_b)$.

Under the OR rule, the sensitivity of the combined result is higher than that of either test alone, but the combined specificity is lower than that of either test. With the AND rule, this is reversed: The specificity of the combined result is higher than either test alone, but the combined sensitivity is lower than that of either test.

Serial testing is an alternative to parallel testing that is particularly cost-efficient when screening for rare conditions and often is used when the second test is expensive and/or risky. Under the OR rule, if the first test is positive, the diagnosis is positive; otherwise, the second test is performed. If the second test is positive after a negative first test, then the diagnosis also is positive; otherwise, the diagnosis is negative. The OR rule, then, leads to a higher overall sensitivity than either test by itself. With the AND rule, if the first test is positive, the second test is performed. If the second test is positive, the diagnosis is positive; otherwise, the diagnosis is negative. The AND rule, then, leads to a higher overall specificity than either test by itself.

To calculate sensitivity of the combined test using serial testing with the OR rule, the formula is: $SE_a + (1 - SE_a) \times SE_b$. Specificity under the OR rule is simply $SP_a \times SP_b$. Conversely, to calculate sensitivity using the AND rule, the formula is: $SE_a \times SE_b$, while specificity under the AND rule is $SP_a + (1 - SP_a) \times SP_b$.

ROC Curves

While some tests provide dichotomous results (that is, positive or negative), other tests yield results that are numeric values (for example, attenuation of a lesion on CT) or ordered categories (for example, BI-RADS scoring used in mammography). Consider CT attenuation as a diagnostic test for distinguishing papillary renal cell carcinomas from other types of renal masses [4]. In Table 3, the ratio of tumor enhancement to normal kidney enhancement (T–K ratio) of 10 masses is listed.

How do we calculate the basic measures of accuracy, that is, sensitivity and specificity, for T–K ratio as a diagnostic test for papillary masses? We shall consider each unique T–K ratio value as a “cutoff,” or “decision threshold” and calculate the sensitivity and specificity associated with each cutoff. Masses with T–K ratio values greater than or equal to the cutoff are called “negative” for papillary lesions and masses with T–K ratio values less than the cutoff are called “positive” for papillary lesions. In Table 3, the third and fourth columns give the calculated sensitivity and specificity, respectively, using the T–K ratio value in column 2 as the cutoff. Note that as the value of the cutoff increases, the specificity decreases while the sensitivity increases.

TABLE 3 T–K Ratio Values of 5 Papillary and 5 Nonpapillary Renal Masses

Cell Type	T–K Ratio	Sensitivity	Specificity	FPR
PRCC	0.05	0.0	1.0	0.0
PRCC	0.11	0.2	1.0	0.0
Other	0.20	0.4	1.0	0.0
PRCC	0.22	0.4	0.8	0.2
PRCC	0.25	0.6	0.8	0.2
Other	0.29	0.8	0.8	0.2
Other	0.38	0.8	0.6	0.4
PRCC	0.43	0.8	0.4	0.6
Other	0.56	1.0	0.4	0.6
Other	0.66	1.0	0.2	0.8

Note.—PRCC = papillary renal cell carcinoma, FPR = false-positive rate, or 1 – specificity.

In Figure 1, we have plotted the 10 pairs of sensitivity and specificity calculated in Table 3. The *y*-axis is the sensitivity and the *x*-axis is 1 minus the specificity, or the false-positive rate (FPR). Connecting these points with line segments, we have constructed an ROC curve [5]. A test with an ROC curve that lies near the “chance diagonal line” in Figure 1 has no ability, beyond mere guessing, at distinguishing between patients with and without the disease. In contrast, a test with an ROC curve that passes near the upper left corner (that is, near 100% sensitivity and 0% FPR [100% specificity]) is nearly perfect at distinguishing disease from no disease. T–K ratio has moderate accuracy, with its ROC curve falling between these two extremes.

Suppose now that an investigator proposes the ratio of the attenuation of the mass to the attenuation of the abdominal aorta (T–A ratio) as a new diagnostic test for papillary lesions. This investigator, however, arbitrarily chooses a single cutoff and reports only the sensitivity and specificity at that cutoff. Figure 2 illustrates this single point (labeled A) in relation to the ROC curve for T–K ratio. We might be tempted to conclude that T–K ratio is superior to T–A ratio because point A falls below the ROC curve for T–K ratio. There

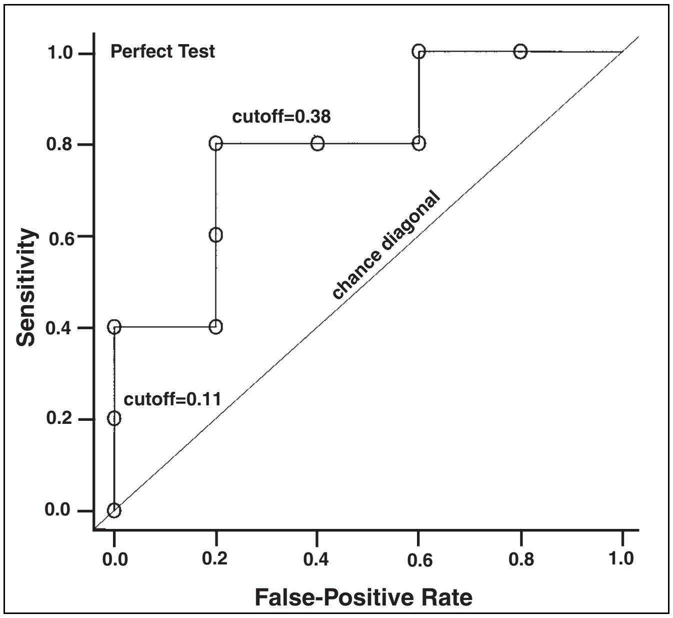


Fig. 1.—10 pairs of sensitivity and specificity as calculated in Table 3. The *y*-axis is the sensitivity and the *x*-axis is 1 minus the specificity, or the false-positive rate (FPR). Receiver operating characteristic (ROC) curve is created by connecting points with line segments.

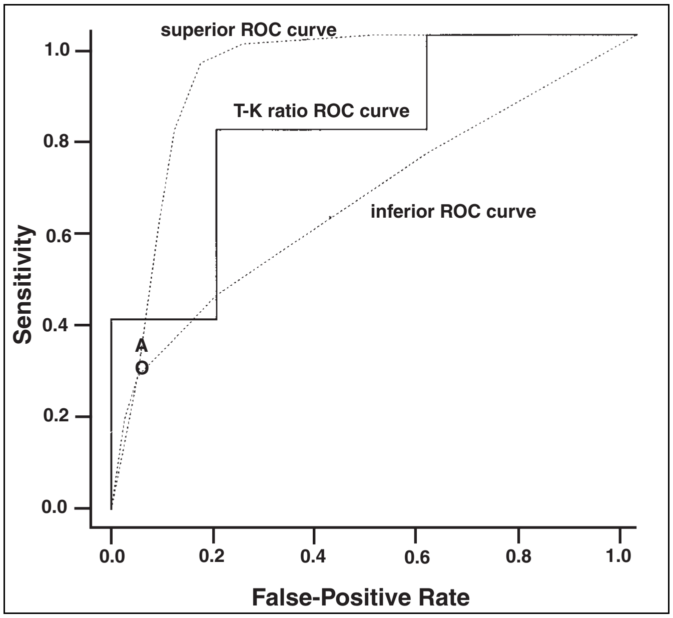


Fig. 2.—Single cutoff point (labeled A) in relation to the receiver operating characteristic (ROC) curve for T–K (tumor enhancement to normal kidney enhancement) ratio.

are, however, an infinite number of ROC curves that could pass through point A, two of which are depicted by dashed curves in Figure 2. Some of these ROC curves could be superior to the ROC curve for T–K ratio for most FPRs and others inferior. Based on the single sensitivity and specificity reported by the investigator, we cannot determine if the T–A ratio is superior or inferior in relation to the T–K ratio. However, if we had been given the ROC curves of both the T–A and T–K ratio, then we could compare these two diagnostic tests and determine, for any range of FPRs, which test is preferred.

This example illustrates the importance of ROC curves and why they have become the state-of-the-art method for describing the diagnostic accuracy of a test. In a future module in this series Obuchowski [6] provides a detailed account of ROC curves, including constructing smooth ROC curves, estimating various summary measures of accuracy derived from them, finding the optimal cutoff on the ROC curve for a particular clinical application, and identifying available software.

Interpretation of Diagnostic Tests

Calculating the Positive and Negative Predictive Values

Clinicians are faced each day with the challenge of deciding appropriate management for patients, based at least in part on the results of less than perfect diagnostic tests. These clinicians need answers to the following questions. “What is the likelihood that this patient has the disease when the test result is positive?” and “What is the likelihood that this patient does *not* have the disease when the test result is negative?” The answers to these questions are known as the positive and negative predictive values, respectively. We illustrate these with the following example.

The lemon sign has been described as an important indicator of spina bifida. Nyberg et al. [7] describe the sensitivity and specificity of the lemon sign in the detection of spina bifida in a high-risk population (elevated material serum α-fetoprotein level, suspected hydrocephalus or neural tube defect, or family history of neural tube defect). A portion of their data is summarized in Table 4.

Spina bifida occurred in 6.1% (14/229) of the sample, that is, sample prevalence was 6.1%. The lemon sign was seen in 92.9% (13/14) of the fetuses with spina bifida (92.9% sensitivity), and was absent in 98.6% (212/215) of the fetuses without spina bifida (98.6% specificity).

We also can calculate the positive and negative predictive values of the lemon sign from the available data. The positive predictive value (PPV) is the probability that the fetus has spina bifida when the lemon sign is present. The PPV is calculated as follows:

$$\mathrm{PPV} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FP}) = \\ 13 / (13 + 3) \times 100\% = 81.3\% \qquad (1)$$

The PPV differs from sensitivity. While the PPV tells us the probability of a fetus with spina bifida following detection of the lemon sign (that probability is 0.813, or 81.3%), the sensitivity tells us the probability that the lemon sign will be present among fetuses with spina bifida (probability is 0.929, or 92.9%). PPV helps the clinician decide how to treat the patient after the diagnostic test comes back positive. Sensitivity, on the other hand, is a property of the diagnostic test and helps the clinician decide which test to use.

The corollary to the PPV is the negative predictive value (NPV), that is, the probability that spina bifida will not be present when the lemon sign is absent. The NPV is calculated as follows:

$$\mathrm{NPV} = \mathrm{TN} / (\mathrm{TN} + \mathrm{FN}) = \\ 212 / (212 + 1) \times 100\% = 99.5\% \qquad (2)$$

If the lemon sign is absent, there is a 99.5% chance that the fetus will not have spina bifida. The NPV is different from the test’s specificity. Specificity tells us the probability that the lemon sign will be absent among fetuses without spina bifida (that probability is 0.986, or 98.6%).

TABLE 4 Lemon Sign Versus Spinal Cord Defect in Fetuses Prior to 24 Weeks

Lemon Sign	Spina Bifida	No Spina Bifida	Total
+	13	3	16
–	1	212	213
Total	14	215	229

Note.—SE = 92.9%, SP = 98.6%, PPV = 81.3%, NPV = 99.99%, prevalence = 6.1%.

TABLE 5 The PPV of the Lemon Sign in the General Population

Lemon Sign	Spina Bifida	No Spina Bifida	Total
+	9	140	149
–	1	9,850	9,851
Total	10	9,990	10,000

Note.—SE = 90.0%, SP = 98.6%, PPV = 6.0%, NPV = 99.99%, prevalence = 0.1%.

The PPV and NPV can also be calculated from Bayes' theorem. Bayes' theorem allows us to compute the PPV and NPV from estimates of the test's sensitivity and specificity, and the probability of the disease before the test is applied. The latter is referred to as the pretest probability and is based on the patient's previous medical history, previous and recent exposures, current signs and symptoms, and results of other screening and diagnostic tests performed. When this information is unknown or when calculating the PPV or NPV for a population, the prevalence of the disease in the population is used as the pretest probability. The PPV and NPV, then, are called posttest probabilities (also, revised or posterior probabilities), and represent the probability of the disease after the test result is known.

Let p denote the pretest probability of disease, and SE and SP the sensitivity and specificity of the diagnostic test. Recalling the expression for a conditional probability (see module 10 [8]),

$$\begin{gathered}\mathrm{PPV} = \mathrm{P}(\text{disease}|+\text{test}) = \\ [\mathrm{SE} \times p] \,/\, [\mathrm{SE} \times p + (1 - \mathrm{SP}) \times (1 - p)]\end{gathered} \qquad (3)$$

$$\begin{gathered}\mathrm{NPV} = \mathrm{P}(\text{no disease}|-\text{test}) = \\ [\mathrm{SP} \times (1 - p)] \,/\, [\mathrm{SP} \times (1 - p) + (1 - \mathrm{SE}) \times p]\end{gathered} \qquad (4)$$

Thus, the posttest probability of disease for any patient can be calculated if one knows the accuracy of the test and the patient's pretest probability of disease.

The PPV and NPV can vary markedly, depending on the patient's pretest probability, or prevalence of disease in the population. In the Nyberg et al. [7] study the prevalence rate of spina bifida in their high risk sample was 6.1%. In the general population, however, the prevalence of spina bifida is much less, about 0.1%. Filly [9] studied the predictive ability of the lemon sign in the general population. He assumed that the sensitivity of the lemon sign was 90.0% and the specificity was 98.6% (very similar to that in Nyberg's small study, 92.9% and 98.6%, respectively). In a sample of 10,000 fetuses from a low-risk population (see Table 5), Filly showed that the positive predictive value is only 6%. This is in contrast to the PPV of 81.3% in the Nyberg study. The drastic difference in PPVs is due to the different prevalence rates of spina bifida in the two samples, 6.1% in Nyberg's and 0.1% in Filly's. Thus, while a high-risk fetus with a lemon sign may have an 81% chance of having spina bifida, "a low risk fetus with a lemon sign has a 94% chance of being *perfectly normal*" [9]. This example illustrates the importance of reporting the pretest probability or prevalence rate of disease whenever one presents a PPV or NPV.

Rationale for Ordering a Diagnostic Test

The previous section described how clinicians can use the results of a diagnostic test to plan a patient's management. Let's back up a bit in the clinical decision-making process and look at the rationale for ordering a diagnostic test.

In the simplest scenario (ignoring monetary costs, insurance reimbursement rates, etc.), there are three pieces of information that a clinician needs to determine whether a diagnostic test should or should not be ordered:

1. From the patient's previous medical history, previous and recent exposures, current signs and symptoms, and results of other screening and diagnostic tests performed, what is the probability that this patient has the disease (that is, the pretest probability)?

2. How accurate (sensitivity and specificity) is the diagnostic test being considered?

3. Could the results of this test affect the patient's management?

In the previous section, we saw how the pretest probability and the test's sensitivity and specificity fit into Bayes' theorem to tell us the posttest probability of disease. We also saw, even for a very accurate test, how the PPV can be quite low when the pretest probability is low. The clinician ordering a test needs to consider how the patient will be managed if the test result is negative versus if the test result is positive. If the probability of disease will still be low after a positive test, then the test may have no impact on the patient's management.

An example is screening for intracranial aneurysms in the general population. The prevalence of aneurysms is low, maybe 1%, in the general population. Even though magnetic resonance angiography (MRA) may have excellent accuracy, say 95% sensitivity and specificity, the PPV is still quite low, 0.16 (16%) from equation 3. Considering the nontrivial risks of invasive catheter angiography (which is the usual presurgical tool) [10], the clinician may decide that even after a positive MRA, the patient should not undergo catheter angiography. In this scenario, the clinician may decide not to order the MRA, given that its result, either positive or negative, will not impact the patient's management.

Designing Studies to Estimate and Compare Tests' Diagnostic Accuracy

As with all new medical devices, treatments, and procedures, the efficacy of diagnostic tests must be assessed in clinical studies. In the second module of this series Jarvik [11] described six levels of diagnostic efficacy. Here, we will focus on the second level, which is the stage at which investigators assess the diagnostic *accuracy* of a test.

Phases in the Assessment of Diagnostic Test Accuracy

There typically are three phases to the assessment of a diagnostic test's accuracy [3]. The first is the *exploratory phase*. It usually is the first clinical study performed to assess the efficacy of a new diagnostic test. These tend to be small, inexpensive studies, typically involving 10 to 50 patients with and without the disease of interest. The patients selected for the study samples often are cases with classical overt disease (for example, symptomatic lung cancer) and healthy volunteer controls. If the test results of these two populations do not differ, then it is not worth pursuing the diagnostic test further.

The second phase is the *challenge phase*. Here, we recognize that a diagnostic test's sensitivity and specificity can vary with the extent and stage of the disease, and the comorbidities present. Thus, in this phase we select patients with subtle, or early disease, and

TABLE 6 Common Features of Diagnostic Test Accuracy Studies

Feature	Explanation
Two samples of patients	One sample of patients with and one sample without the disease are needed to estimate both sensitivity and specificity.
Well-defined patient samples	Regardless of the sampling scheme used to obtain patients for the study, the characteristics of the study patients (e.g., age, gender, comorbidities, stage of disease) should be reported.
Well-defined diagnostic test	The diagnostic test must be clearly defined and applied in the same fashion to all study patients.
Gold standard/reference standard	The true disease status of each study patient must be determined by a test or procedure that is infallible, or nearly so.
Sample of interpreters	If the test relies on a trained observer to interpret it, then two or more such observers are needed to independently interpret the test [15].
Blinded interpretations	The gold standard should be conducted and interpreted blinded to the results of the diagnostic test, and the diagnostic test should be performed and interpreted blinded to the results of the gold standard.
Standard reporting of findings	The results of the study should be reported following published guidelines for the reporting of diagnostic test accuracy [16].

with comorbidities that could interfere with the diagnostic test [12]. For example, in a study to assess the ability of MRI to detect lung cancer, the study patients might include those with small nodules (3 cm), and patients with nodules and interstitial disease. The controls might have diseases in the same anatomic location as the disease of interest, for example, interstitial disease but no nodules. These studies often include competing diagnostic tests to compare their accuracies with the test under evaluation. ROC curves are most often used to assess and compare the tests. If the diagnostic test shows good accuracy, then it can be considered for the third phase of assessment.

The third phase is the *advanced phase*. These studies often are multicenter studies involving large numbers of patients (100 or more). The patient sample should be representative of the target clinical population. For example, instead of selecting patients with known lung cancer and controls without cancer, we might recruit patients presenting to their primary care physician with a persistent cough or bloody sputum. Further testing and follow-up will determine which patients have lung cancer and which do not.

It is from this third phase where we obtain reliable estimates of a test's accuracy for the target clinical population. Estimates of accuracy from the exploratory phase usually are too optimistic because the "sickest of the sick" are compared with the "wellest of the well" [13]. In contrast, estimates of accuracy from the challenge phase often are too low because the patients are exceptionally difficult to diagnose.

Common Features of Diagnostic Test Accuracy Studies

The studies in the three phases differ in terms of their objectives, sampling of patients, and sample sizes. There are, however, some common features to all studies of diagnostic test accuracy, as summarized in Table 6. We elaborate here on a few important issues.

Studies of diagnostic test accuracy require both subjects with and without the disease of interest. If one of these populations is not represented in the study, then either sensitivity or specificity cannot be calculated. We stress that reporting one without reference to the other is uninformative and often misleading. The number of patients needed for diagnostic accuracy studies depends on the phase of the study, the clinical setting in which the test will be applied (for example, screening or diagnostic), and certain characteristics of the patients and test itself (for example, does the test require interpretation by human observers?). Statistical methods are available for determining the appropriate sample size for diagnostic accuracy studies [3, 14].

Studies of diagnostic test accuracy require a test or procedure for determining the true disease status of each patient. This test or procedure is called the "gold standard" (or "standard of reference," "reference standard," particularly when there is no perfect gold standard). The gold, or reference, standard must be conducted and interpreted blinded to the diagnostic test results to avoid bias. Common standards of reference in radiology studies are surgery, pathology results, and clinical follow-up. For example, in the study of Carpenter et al. [2] of the accuracy of MR venography for detecting deep venous thrombosis, contrast venography was used as the reference standard. Sometimes a study uses more than one type of reference standard. For example, in a study to assess the accuracy of mammography, patients with a suspicious lesion on mammography might undergo core biopsy and/or surgery, whereas patients with a negative mammogram would need to be followed for 2 years either to confirm that the patient was cancer free or to detect missed cancers on follow-up screenings. Note that when using different reference standards for patients with positive and negative test results, it is important that all the reference standards are infallible, or nearly so. One form of workup bias occurs when patients with one test result undergo a less rigorous reference standard than patients with a different test result [3].

Determining the appropriate reference standard for a study often is the most difficult part of designing a diagnostic accuracy study. Reference standards should be infallible, or nearly so. This is difficult, however, because even pathology is not infallible, as it is an interpretive field relying on subjective assessment from human observers with varying skill levels. One such example is the reader variability in pathologic interpretation of borderline intraductal breast carcinoma versus atypical ductal carcinoma. While some pathologists may interpret the lesion as intraductal cancer, others may interpret the same lesion as atypical ductal hyperplasia. While often we have to accept that a reference standard is not perfect, it is important that it be nearly infallible. If the reference standard is not nearly infallible, then *imperfect gold standard bias* can lead to unreliable and misleading estimates of accuracy. Zhou et al. [3] discuss in detail imperfect gold standard bias and possible solutions.

In other situations, no reference standard is available (for example, epilepsy) or it is unethical to subject patients to the reference standard because it poses a risk (for example, an invasive test such as catheter angiography). In these situations, we at least can correlate the test results to other tests' findings and to clinical outcome, even if we cannot report the test's sensitivity and specificity.

It is *never* an option to omit from the calculation of sensitivity and specificity those patients without a diagnosis confirmed by a reference standard. Such studies yield erroneous estimates of test accuracy due to a form of workup bias called *verification bias* [17, 18].

This is one of the most common types of bias in radiology studies [19] and is counterintuitive. Investigators often believe they are getting more reliable estimates of accuracy by excluding cases where the reference standard was not performed. If, however, the diagnostic test results were used in the decision of whether to perform the reference standard procedure, then verification bias most likely is present. For example, if the results of MR venography are used to determine which patients will undergo contrast venography, and if patients who did not undergo contrast venography are excluded from the calculations of the test's accuracy, then verification bias exists. Zhou et al. [3] discuss verification bias from a statistical standpoint and offer a variety of solutions.

Summary

We conclude with a summary of five key points in the clinical evaluation of diagnostic tests:

1. Sensitivity and specificity always should be reported together.

2. ROC curves allow a comprehensive assessment and comparison of diagnostic test accuracy.

3. PPV and NPV cannot be interpreted correctly without knowing the prevalence of disease in the study sample.

4. Patients who did not undergo the reference standard procedure should never be omitted from studies of diagnostic test accuracy.

5. Published guidelines should be followed when reporting the findings from studies of diagnostic test accuracy.

References

1. Gehlbach SH. Interpretation: sensitivity, specificity, and predictive value. In: Gehlbach SH, ed. *Interpreting the medical literature.* New York: McGraw-Hill, 1993:129–139
2. Carpenter JP, Holland GA, Baum RA, Owen RS, Carpenter JT, Cope C. Magnetic resonance venography for the detection of deep venous thrombosis: comparison with contrast venography and duplex Doppler ultrasonography. *J Vasc Surg* 1993;18:734–741
3. Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine.* New York: Wiley & Sons, 2002
4. Herts BR, Coll DM, Novick AC, Obuchowski N, Linnell G, Wirth SL, Baker ME. Enhancement characteristics of papillary renal neoplasms revealed on triphasic helical CT of the kidneys. *AJR* 2002;178:367–372
5. Metz CE. ROC methodology in radiological imaging. *Invest Radiol* 1986;21:720–733
6. Obuchowski NA. Receiver operating characteristic (ROC) analysis. *AJR* 2005(in press)
7. Nyberg DA, Mack LA, Hirsch J, Mahony BS. Abnormalities of fetal cranial contour in sonographic detection of spina bifida: evaluation of the "lemon" sign. *Radiology* 1988;167:387–392
8. Joseph L, Reinhold C. Introduction to probability theory and sampling distributions. *AJR* 2003;180:917–923
9. Filly RA. The "lemon" sign: a clinical perspective. *Radiology* 1988;167:573–575
10. Levey AS, Pauker SG, Kassirer JP, et al. Occult intracranial aneurysms in polycystic kidney disease: when is cerebral arteriography indicated? *N Engl J Med* 1983;308:986–994
11. Jarvik JG. The research framework. *AJR* 2001;176:873–877
12. Ransohoff DJ, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926–930
13. Sox Jr HC, Blatt MA, Higgins MC, Marton KI. *Medical decision making***.** Boston: Butterworths-Heinemann, 1988
14. Beam CA. Strategies for improving power in diagnostic radiology research. *AJR* 1992;159:631–637
15. Obuchowski NA. How many observers in clinical studies of medical imaging? *AJR* 2004;182:867–869
16. Bossuyt PM, Reitsma JB, Bruns DE, et al. Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Acad Radiol* 2003;10:664–669
17. Begg CB, McNeil BJ. Assessment of radiologic tests, control of bias, and other design considerations. *Radiology* 1988;167:565–569
18. Black WC. How to evaluate the radiology literature. *AJR* 1990;154:17–22
19. Reid MC, Lachs MS, Feinstein AR. Use of methodologic standards in diagnostic test research: getting better but still not good. *JAMA* 1995;274:645–651

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004

Fundamentals of Clinical Research for Radiologists

Nancy A. Obuchowski[1]

ROC Analysis

In this module we describe the standard methods for characterizing and comparing the accuracy of diagnostic and screening tests. We motivate the use of the receiver operating characteristic (ROC) curve, provide definitions and interpretations for the common measures of accuracy derived from the ROC curve (e.g., the area under the ROC curve), and present recent examples of ROC studies in the radiology literature. We describe the basic statistical methods for fitting ROC curves, comparing them, and determining sample size for studies using ROC curves. We briefly describe the MRMC (multiple-reader, multiple-case) ROC paradigm. We direct the interested reader to available software for analyzing ROC studies and to literature on more advanced statistical methods of ROC analysis.

Why ROC?

In module 13 [1], we defined the basic measures of accuracy: sensitivity (the probability the diagnostic test is positive for disease for a patient who truly has the disease) and specificity (the probability the diagnostic test is negative for disease for a patient who truly does not have the disease). These measures require a decision rule (or positivity threshold) for classifying the test results as either positive or negative. For example, in mammography the BI-RADS (Breast Imaging Reporting and Data System) scoring system is used to classify mammograms as normal, benign, probably benign, suspicious, or malignant. One positivity threshold is classifying probably benign, suspicious, and malignant findings as positive (and classifying normal and benign findings as negative). Another positivity threshold is classifying suspicious and malignant findings as positive. Each threshold leads to different estimates of sensitivity and specificity. Here, the second threshold would have higher specificity than the first but lower sensitivity. Also, note that trained mammographers use the scoring system differently. Even the same mammographer may use the scoring system differently on different reviewing occasions (e.g., classifying the same mammogram as probably benign on one interpretation and as suspicious on another), leading to different estimates of sensitivity and specificity even with the same threshold.

Which decision threshold should be used to classify test results? How will the choice of a decision threshold affect comparisons between two diagnostic tests or between two radiologists? These are critical questions when computing sensitivity and specificity, yet the choice for the decision threshold is often arbitrary.

ROC curves, although constructed from sensitivity and specificity, do not depend on the decision threshold. In an ROC curve, every possible decision threshold is considered. An ROC curve is a plot of a test's false-positive rate (FPR), or 1 – specificity (plotted on the horizontal axis), versus its sensitivity (plotted on the veritical axis). Each point on the curve represents the sensitivity and FPR at a different decision threshold. The plotted (FPR, sensitivity) coordinates are connected with line segments to construct an empiric ROC curve. Figure 1 illustrates an empiric ROC curve constructed from the fictitious mammography data in Table 1. The empiric ROC curve has four points corresponding to the four decision thresholds described in Table 1.

An ROC curve begins at the (0, 0) coordinate, corresponding to the strictest decision threshold whereby all test results are negative for disease (Fig. 1). The ROC curve ends at the (1, 1) coordinate, corresponding to the

Received October 28, 2004; accepted after revision November 3, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 14th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology.* The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, 9500 Euclid Ave., Cleveland, OH. Address correspondence to N. Obuchowski.

AJR 2005;184:364–372

0361–803X/05/1842–364

most lenient decision threshold whereby all test results are positive for disease. An empiric ROC curve has $h - 1$ additional coordinates, where h is the number of unique test results in the sample. In Table 1 there are 200 test results, one for each of the 200 patients in the sample, but there are only five unique results: normal, benign, probably benign, suspicious, and malignant. Thus, $h = 5$, and there are four coordinates plotted in Figure 1 corresponding to the four decision thresholds described in Table 1.

The line connecting the (0, 0) and (1, 1) coordinates is called the "chance diagonal" and represents the ROC curve of a diagnostic test with no ability to distinguish patients with versus those without disease. An ROC curve that lies above the chance diagonal, such as the ROC curve for our fictitious mammography example, has some diagnostic ability. The further away an ROC curve is from the chance diagonal, and therefore, the closer to the upper left-hand corner, the better discriminating power and diagnostic accuracy the test has.

In characterizing the accuracy of a diagnostic (or screening) test, the ROC curve of the test provides much more information about how the test performs than just a single estimate of the test's sensitivity and specificity [1, 2]. Given a test's ROC curve, a clinician can examine the trade-offs in sensitivity versus specificity for various decision thresholds. Based on the relative costs of false-positive and false-negative errors and the pretest probability of disease, the clinician can choose the optimal decision threshold for each patient. This idea is discussed in more detail in a later section of this article. Often, patient management is more complex than is allowed with a decision threshold that classifies the test results into positive or negative. For example, in mammography suspicious and malignant findings are usually followed up with biopsy, probably benign findings usually result in a follow-up mammogram in 3–6 months, and normal and benign findings are considered negative.

When comparing two or more diagnostic tests, ROC curves are often the only valid method of comparison. Figure 2 illustrates two scenarios in which an investigator, comparing two diagnostic tests, could be misled by relying on only a single sensitivity–specificity pair. Consider Figure 2A. Suppose a more expensive or risky test (represented by ROC curve Y) was reported to have the following accuracy: sensitivity = 0.40, specificity = 0.90 (labeled as coordinate 1 in Fig. 2A); a less expensive or less risky test (represented by ROC curve X) was reported to have the following accuracy: sensitivity = 0.80, specificity = 0.65 (labeled as coordinate 2 in Fig. 2A). If the investigator is looking for the test with better specificity, then he or she may choose the more expensive, risky test, not realizing that a simple change in the decision threshold of the less expensive, cheaper test could provide the desired specificity at an even higher sensitivity (coordinate 3 in Fig. 2A).

Now consider Figure 2B. The ROC curve for test Z is superior to that of test X for a narrow range of FPRs (0.0–0.08); otherwise, diagnostic test X has superior accuracy. A comparison of the tests' sensitivities at low FPRs would be misleading unless the diagnostic tests are useful only at these low FPRs.

TABLE 1 **Construction of Receiver Operating Characteristic Curve Based on Fictitious Mammography Data**

Mammography Results (BI-RADs Score)	Pathology/Follow-Up Results		Decision Rules 1–4	
	Not Malignant	Malignant	FPR	Sensitivity
Normal	65	5	(1) 35/100	95/100
Benign	10	15	(2) 25/100	80/100
Probably benign	15	10	(3) 10/100	70/100
Suspicious	7	60	(4) 3/100	10/100
Malignant	3	10		
Total	100	100		

Note.—Decision rule 1 classifies normal mammography findings as negative; all others are positive. Decision rule 2 classifies normal and benign mammography findings as negative; all others are positive. Decision rule 3 classifies normal, benign, and probably benign findings as negative; all others are positive. Decision rule 4 classifies normal, benign, probably benign, and suspicious findings as negative; malignant is the only finding classified as positive. BI-RADS = Breast Imaging Reporting and Data System, FPR = false-positive rate.

To compare two or more diagnostic tests, it is convenient to summarize the tests' accuracies with a single summary measure. Several such summary measures are used in the literature. One is Youden's index, defined as sensitivity + specificity − 1 [2]. Note, however, that Youden's index is affected by the choice of the decision threshold used to define sensitivity and specificity. Thus, different decision thresholds yield different values of the Youden's index for the same diagnostic test.

Another summary measure commonly used is the probability of a correct diagnosis, often referred to simply as "accuracy" in the literature. It can be shown that the probability of a correct diagnosis is equivalent to

$$\text{probability (correct diagnosis)} = \text{PREV}_S \times \text{sensitivity} + (1 - \text{PREV}_S) \times \text{specificity}, \quad (1)$$

where PREV_S is the prevalence of disease in the sample. That is, this summary measure of accuracy is affected not only by the choice of the decision threshold but also by the prevalence of disease in the study sample [2]. Thus, even slight changes in the prevalence of disease in the population of patients being tested can lead to different values of "accuracy" for the same test.

Summary measures of accuracy derived from the ROC curve describe the inherent accuracy of a diagnostic test because they are not affected by the choice of the decision threshold and they are not affected by the prevalence of disease in the study sample. Thus, these summary measures are preferable to Youden's index and the probability of a correct diagnosis [2]. The most popular summary measure of accuracy is the area under the ROC curve, often denoted as "AUC" for area under curve. It ranges in value from 0.5 (chance) to 1.0 (perfect discrimination or accuracy). The chance diagonal in Figure 1 has an AUC of 0.5. In Figure 2A the areas under both ROC curves are the same, 0.841. There are three interpretations for the AUC: the average sensitivity over all false-positive rates; the average specificity over all sensitivities [3]; and the probability that, when presented with a randomly chosen patient with disease and a randomly chosen patient without disease, the results of the diagnostic test will rank the patient with disease as having higher suspicion for disease than the patient without disease [4].

The AUC is often too global a summary measure. Instead, for a particular clinical application, a decision threshold is chosen so that the diagnostic test will have a low FPR (e.g., FPR < 0.10) or a high sensitivity (e.g., sensitivity > 0.80). In these circumstances, the accuracy of the test at the specified FPRs (or specified sensitivities) is a more meaningful summary measure than the area under the entire ROC curve. The partial area under the ROC curve, PAUC (e.g., the PAUC where FPR < 0.10, or the PAUC where sensitivity > 0.80), is then an appropriate summary measure of the diagnostic test's accuracy. In Figure 2B, the PAUCs for the two tests where the FPR is between 0.0 and 0.20 are the same, 0.112. For interpretation purposes, the PAUC is often divided by its maximum value, given by the range (i.e., maximum–minimum) of the FPRs (or false-negative rates [FNRs]) [5]. The PAUC divided by its maximum value is called the partial area index and takes on values between 0.5 and 1.0, as does the AUC. It is interpreted as the average sensitivity for the FPRs examined (or average specificity for the FNRs examined). In our example, the range of the FPRs of interest is 0.20–0.0 = 0.20; thus, the average sensitivity for FPRs less than 0.20 for diagnostic tests X and Z in Figure 2B is 0.56.

Although the ROC curve has many advantages in characterizing the accuracy of a diagnostic test, it also has some limitations. One criticism is that the ROC curve extends beyond the clinically relevant area of potential clinical interpretation. Of course, the PAUC was developed to address this criticism. Another criticism is that it is possible for a diagnostic test with perfect discrimination between diseased and nondiseased patients to have an AUC of 0.5. Hilden [6] describes this unusual situation and offers solutions. When comparing two diagnostic tests' accuracies, the tests' ROC curves can cross, as in Figure 2. A comparison of these tests based only on their AUCs can be misleading. Again, the PAUC attempts to address this limitation. Last, some [6, 7] criticize the ROC curve, and especially the AUC, for not incorporating the pretest probability of disease and the costs of misdiagnoses.

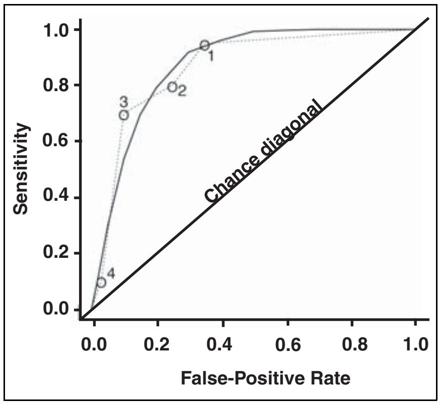


Fig. 1.—Empiric and fitted (or "smooth") receiver operating characteristic (ROC) curves constructed from mammography data in Table 1. Four labeled points on empiric curve (*dotted line*) correspond to four decision thresholds used to estimate sensitivity and specificity. Area under curve (AUC) for empiric ROC curve is 0.863 and for fitted curve (*solid line*) is 0.876.

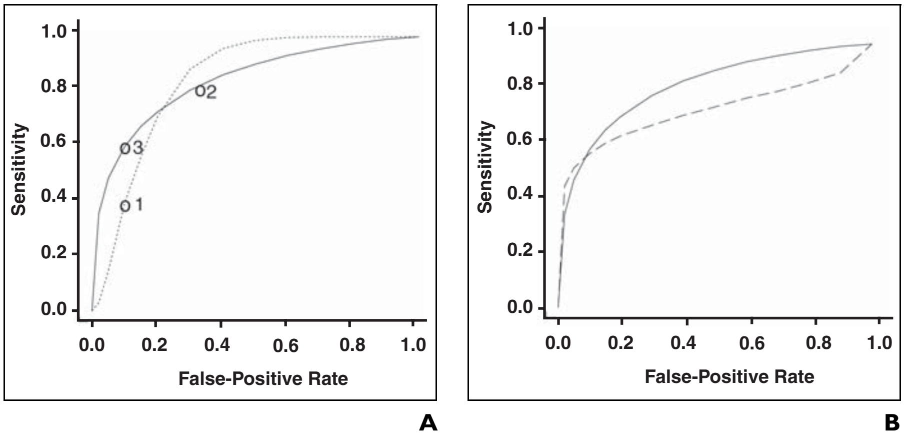


Fig. 2.—Two examples illustrate advantages of receiver operating characteristic (ROC) curves (see text for explanation) and comparing summary measures of accuracy.
A, ROC curve Y (*dotted line*) has same area under curve (AUC) as ROC curve X (*solid line*), but lower partial area under curve (PAUC) when false-positive rate (FPR) is ≤ 0.20, and higher PAUC when false-positive rate > 0.20.
B, ROC curve Z (*dashed line*) has same PAUC as curve X (*solid line*) when FPR ≤ 0.20 but lower AUC.

The ROC Study

Weinstein et al. [1] describe the common features of a study of the accuracy of a diagnostic test. These include samples from both patients with and those without the disease of interest and a reference standard for determining whether positive test results are true-positives or false-positives, and whether negative test results are true-negatives or false-negatives. They also discuss the need to blind reviewers who are interpreting test images and other relevant biases common to these types of studies.

In ROC studies we also require that the test results, or the interpretations of the test images, be assigned a numeric value or rank. These numeric measurements or ranks are the basis for defining the decision thresholds that yield the estimates of sensitivity and specificity that are plotted to form the ROC curve. Some diagnostic tests yield an objective measurement (e.g., attenuation value of a lesion). The decision thresholds for constructing the ROC curve are

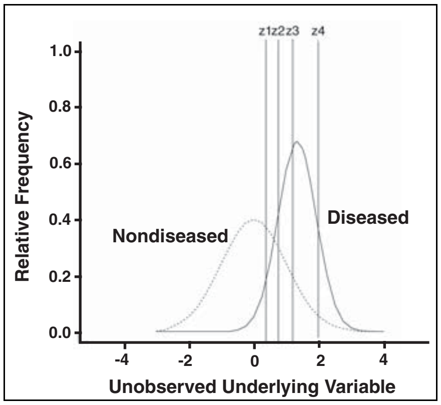


Fig. 3.—Unobserved binormal distribution that was assumed to underlie test results in Table 1. Distribution for nondiseased patients was arbitrarily centered at 0 with SD of 1 (i.e., $\mu_0 = 0$ and $\sigma_0 = 1$). Binormal parameters were estimated to be A = 2.27 and B = 1.70. Thus, distribution for diseased patients is centered at $\mu_1 = 1.335$ with SD of $\sigma_1 = 0.588$. Four cutoffs, z1, z2, z3, and z4, correspond to four decision thresholds in Table 1. If underlying test value is less than z1, then mammographer assigns test result of "normal." If the underlying test value is less than z2 but greater than z1, then mammographer assigns test result of "benign," and so forth.

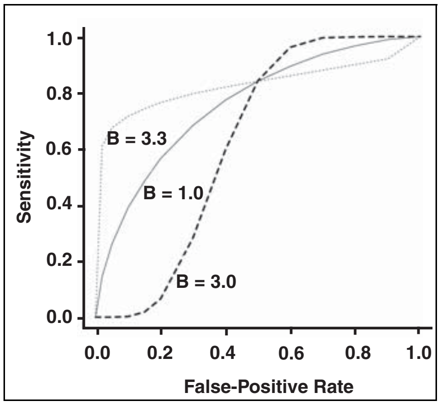
based on increasing the values of the attenuation coefficient. Other diagnostic tests must be interpreted by a trained observer, often a radiologist, and so the interpretation is subjective. Two general scales are often used in radiology for observers to assign a value to their subjective interpretation of an image. One scale is the 5-point rank scale: 1 = definitely normal, 2 = probably normal, 3 = possibly abnormal or equivocal, 4 = probably abnormal, and 5 = definitely abnormal.

The other popular scale is the 0–100% confidence scale, where 0% implies that the observer is completely confident in the absence of the disease of interest, and 100% implies that the observer is completely confident in the presence of the disease of interest. The two scales have strengths and weaknesses [2, 8], but both are reasonably well suited to radiology research. In mammography a rating scale already exists, the BI-RADS score, which can be used to form decision thresholds from least to most suspicion for the presence of breast cancer.

When the diagnostic test requires a subjective interpretation by a trained reviewer, the reviewer becomes part of the diagnostic process [9]. Thus, to properly characterize the accuracy of the diagnostic test, we must include multiple reviewers in the study. This is the so-called MRMC, multiple-reader multiple-case, ROC study. Much has been written about the design and analysis of MRMC studies [10–20]. We mention here only the basic design of MRMC studies, and in a later subsection we describe their statistical analysis.

The usual design for the MRMC study is a factorial design, in which every reviewer interprets the image (or images if there is more than one test) of every patient. Thus, if there are *R* reviewers, *C* patients, and *I* diagnostic tests, then each reviewer interprets $C \times I$ images, and the study involves $R \times C \times I$ total interpretations. The accuracy of each reviewer with each diagnostic test is characterized by an ROC curve, so $R \times I$ ROC curves are constructed. Constructing pooled or consensus ROC curves is not the goal of these studies. Rather, the primary goals are to document the variability in diagnostic test accuracy between reviewers and report the average, or typical, accuracy of reviewers. In order for the results of the study to be generalizeable to the relevant patient and reviewer populations, representative samples from both populations are needed for the study. Often expert reviewers take part in studies of diagnostic test accuracy, but the accuracy for a nonexpert may be considerably less. An excellent illustration of the issues involved in sampling reviewers for an MRMC study can be found in the study by Beam et al. [21].

Fig. 4.—Three receiver operating characteristic (ROC) curves with same binormal parameter A (i.e., A = 1.0) but different values for parameter B of 3.0 ($3\sigma_1 = \sigma_0$), 1.0 ($\sigma_1 = \sigma_0$), and 0.33 ($\sigma_1 = 3\sigma_0$).When B = 3.0, ROC curve dips below chance diagonal; this is called an improper ROC curve [2].



Examples of ROC Studies in Radiology

The radiology literature, and the clinical laboratory and more general medical literature, contain many excellent examples of how ROC curves are used to characterize the accuracy of a diagnostic test and to compare accuracies of diagnostic tests. We briefly describe here three recent examples of ROC curves being used in the radiology literature.

Kim et al. [22] conducted a prospective study to determine if rectal distention using warm water improves the accuracy of MRI for preoperative staging of rectal cancer. After MRI, the patients underwent surgical resection, considered the gold standard regarding the invasion of adjacent structures and regional lymph node involvement. Four observers, unaware of the pathology results, independently scored the MR images using 4- and 5-point rating scales. Using statistical methods for MRMC studies [13], the authors determined that typical reviewers' accuracy for determining outer wall penetration is improved with rectum distention, but that reviewer accuracy for determining regional lymph node involvement is not affected.

Osada et al. [23] used ROC analysis to assess the ability of MRI to predict fetal pulmonary hypoplasia. They imaged 87 fetuses, measuring both lung volume and signal intensity. An ROC curve based on lung volume showed that lung volume has some ability to discriminate between fetuses who will have good versus those who will have poor respiratory outcome after birth. An ROC curve based on the combined information from lung volume and signal intensity, however, has superior accuracy. For more information on the optimal way to combine measures or test results, see the article by Pepe and Thompson [24].

In a third study, Zheng et al. [25] assessed how the accuracy of a mammographic computer-aided detection (CAD) scheme was affected by restricting the maximum number of regions that could be identified as positive. Using a sample of 300 cases with a malignant mass and 200 normals, the investigators applied their CAD system, each time reducing the maximum number of positive regions that the CAD system could identify from seven to one. A special ROC technique called "free-response receiver operating characteristic curves" (FROC) was used. The horizontal axis of the FROC curve differs from the traditional ROC curve in that it gives the average number of false-positives per image. Zheng et al. concluded that limiting the maximum number of positive regions that the CAD could identify improves the overall accuracy of CAD in mammography. For more information on FROC curves and related methods, I refer you to other articles [26–29].

Statistical Methods for ROC Analysis

Fitting Smooth ROC Curves

In Figure 1 we saw the empiric ROC curve for the test results in Table 1. The curve was constructed with line segments connecting the observed points on the ROC curve. Empiric ROC curves often have a jagged appearance, as seen in Figure 1, and often lie slightly below the "true," smooth, ROC curve—that is, the test's ROC curve if it were constructed with an infinite number of points (not just the four points in Fig. 1) and an infinitely large sample size. A smooth curve gives us a better idea of the relationship between the diagnostic test and the disease. In this subsection we describe some methods for constructing smooth ROC curves.

The most popular method of fitting a smooth ROC curve is to assume that the test results (e.g., the BI-RADS scores in Table 1) come from two unobserved distributions, one distribution for the patients with disease and one for the patients without the disease. Usually it is assumed that these two distributions can be transformed to normal distributions, referred to as the binormal assumption. It is the unobserved, underlying distributions that we assume can be transformed to follow a binormal distribution, and not the observed test results. Figure 3 illustrates the hypothesized unobserved binormal distribution estimated for the observed BI-RADS results in Table 1. Note how the distributions for the diseased and nondiseased patients overlap.

Let the unobserved binormal variables for the nondiseased and diseased patients have means μ_0 and μ_1, and variances σ_0 [2] and σ_1 [2], respectively. Then it can be shown [30] that the ROC curve is completed described by two parameters:

$$A = (\mu_1 - \mu_0) / \sigma_1 \qquad (2)$$

$$B = \sigma_0 / \sigma_1. \qquad (3)$$

(See Appendix 1 for a formula that links parameters A and B to the ROC curve.) Figure 4 illustrates three ROC curves. Parameter A was set to be constant at 1.0 and parameter B varies as follows: 0.33 (the underlying distribution of the diseased patients is three times more variable than that of the nondiseased patients), 1.0 (the two distributions have the same SD), and 3.0 (the underlying distribution of the nondiseased patients is three times more variable than that of the diseased patients). As one can see, the curves differ dramatically with changes in parameter B. Parameter A, on the other hand, determines how far the curve is above the chance diagonal (where A = 0); for a constant B parameter, the greater the value of A, the higher the ROC curve lies (i.e., greater accuracy).

Parameters A and B can be estimated from data such as in Table 1 using maximum likelihood methods [30, 31]. For the data in Table 1, the maximum likelihood estimates (MLEs) of parameters A and B are 2.27 and 1.70, respectively; the smooth ROC curve is given in Figure 1. Fortunately, some useful software [32] has been written to perform the necessary calculations of A and B, along with estimation of the area under the smooth curve (see next subsection), its SE and confidence interval (CI), and CIs for the ROC curve itself (see Appendix 1).

Dorfman and Alf [30] suggested a statistical test to evaluate whether the binormal assumption was reasonable for a given data set. Others [33, 34] have shown through empiric investigation and simulation studies that many different underlying distributions are well approximated by the binormal assumption.

When the diagnostic test results are themselves a continuous measurement (e.g., CT attenuation values, or measured lesion diameter), it may not be necessary to assume the existence of an unobserved, underlying distribution. Sometimes continuous-scale test results themselves follow a binormal distribution, but caution should be taken that the fit is good (see the article by Goddard and Hinberg [35] for a discussion of the resulting bias when the distribution is not truly binormal yet the binormal distribution is assumed). Zou et al. [36] suggest using a Box-Cox transformation to transform data to

TABLE 2 **Estimating Area Under Empirical Receiver Operating Characteristic Curve**

Test Results		Score	No. of Pairs	Score x No. of Pairs
Nondiseased	Diseased			
Normal	Normal	1/2	65 x 5 = 325	162.5
Normal	Benign	1	65 x 15 = 975	975
Normal	Probably benign	1	65 x 10 = 650	650
Normal	Suspicious	1	65 x 60 = 3,900	3,900
Normal	Malignant	1	65 x 10 = 650	650
Benign	Normal	0	10 x 5 = 50	0
Benign	Benign	1/2	10 x 15 = 150	75
Benign	Probably benign	1	10 x 10 = 100	100
Benign	Suspicious	1	10 x 60 = 600	600
Benign	Malignant	1	10 x 10 = 100	100
Probably benign	Normal	0	15 x 5 = 75	0
Probably benign	Benign	0	15 x 15 = 225	0
Probably benign	Probably benign	1/2	15 x 10 = 150	75
Probably benign	Suspicious	1	15 x 60 = 900	900
Probably benign	Malignant	1	15 x 10 = 150	150
Suspicious	Normal	0	7 x 5 = 35	0
Suspicious	Benign	0	7 x 15 = 105	0
Suspicious	Probably benign	0	7 x 10 = 70	0
Suspicious	Suspicious	1/2	7 x 60 = 420	210
Suspicious	Malignant	1	7 x 10 = 70	70
Malignant	Normal	0	3 x 5 = 15	0
Malignant	Benign	0	3 x 15 = 45	0
Malignant	Probably benign	0	3 x 10 = 30	0
Malignant	Suspicious	0	3 x 60 = 180	0
Malignant	Malignant	1/2	3 x 10 = 30	15
Total			10,000 pairs	8,632.5

TABLE 3 Fictitious Data Comparing the Accuracy of Two Diagnostic Tests

	ROC Curve	
	X	Y
Estimated AUC	0.841	0.841
Estimated SE of AUC	0.041	0.045
Estimated PAUC where FPR < 0.20	0.112	0.071
Estimated SE of PAUC	0.019	0.014
Estimated covariance	0.00001	
Z test comparing PAUCs	$Z = [0.112 - 0.071] / \sqrt{[0.019^2 + 0.014^2 - 0.00002]}$	
95% CI for difference in PAUCs	$[0.112 - 0.071] \pm 1.96 \times \sqrt{[0.019^2 + 0.014^2 - 0.00002]}$	

Note.—AUC = area under the curve, PAUC = partial area under the curve, CI = confidence interval.

binormality. Alternatively, one can use software like ROCKIT [32] that will bin the test results into an optimal number of categories and apply the same maximum likelihood methods as mentioned earlier for rating data like the BI-RADS scores.

More elaborate models for the ROC curve that can take into account covariates (e.g., the patient's age, symptoms) have also been developed in the statistics literature [37–39] and will become more accessible as new software is written.

Estimating the Area Under the ROC Curve

Estimation of the area under the smooth curve, assuming a binormal distribution, is described in Appendix 1. In this subsection, we describe and illustrate estimation of the area under the empiric ROC curve. The process of estimating the area under the empiric ROC curve is nonparametric, meaning that no assumptions are made about the distribution of the test results or about any hypothesized underlying distribution. The estimation works for tests scored with a rating scale, a 0–100% confidence scale, or a true continuous-scale variable.

The process of estimating the area under the empiric ROC curve involves four simple steps: First, the test result of a patient with disease is compared with the test result of a patient without disease. If the former test result indicates more suspicion of disease than the latter test result, then a score of 1 is assigned. If the test results are identical, then a score of 1/2 is assigned. If the diseased patient has a test result indicating less suspicion for disease than the test result of the nondiseased patient, then a score of 0 is assigned. It does not matter which diseased and nondiseased patient you begin with. Using the data in Table 1 as an illustration, suppose we start with a diseased patient assigned a test result of "normal" and a nondiseased patient assigned a test result of "normal." Because their test results are the same, this pair is assigned a score of 1/2.

Second, repeat the first step for every possible pair of diseased and nondiseased patients in your sample. In Table 1 there are 100 diseased patients and 100 nondiseased patients, thus 10,000 possible pairs. Because there are only five unique test results, the 10,000 possible pairs can be scored easily, as in Table 2.

Third, sum the scores of all possible pairs. From Table 2, the sum is 8,632.5.

Fourth, divide the sum from step 3 by the number of pairs in the study sample. In our example we have 10,000 pairs. Dividing the sum from step 3 by 10,000 gives us 0.86325, which is our estimate of the area under the empiric ROC curve. Note that this method of estimating the area under the empiric ROC curve gives the same result as one would obtain by fitting trapezoids under the curve and summing the areas of the trapezoids (so-called trapezoid method).

The variance of the estimated area under the empiric ROC curve is given by DeLong et al. [40] and can be used for constructing CIs; software programs are available for estimating the nonparametric AUC and its variance [41].

Comparing the AUCs or PAUCs of Two Diagnostic Tests

To test whether the AUC (or PAUC) of one diagnostic test (denoted by AUC_1) equals the AUC (or PAUC) of another diagnostic test (AUC_2), the following test statistic is calculated:

$$Z = [AUC_1 - AUC_2] / \sqrt{[var_1 + var_2 - 2 \times cov]}, \qquad (4)$$

where var_1 is the estimated variance of AUC_1, var_2 is the estimated variance of AUC_2, and cov is the estimated covariance between AUC_1 and AUC_2. When different samples of patients undergo the two diagnostic tests, the covariance equals zero. When the same sample of patients undergoes both diagnostic tests (i.e., a paired study design), then the covariance is not generally equal to zero and is often positive. The estimated variances and covariances are standard output for most ROC software [32, 41].

The test statistic Z follows a standard normal distribution. For a two-tailed test with significance level of 0.05, the critical values are –1.96 and +1.96. If Z is less than –1.96, then we conclude that the accuracy of diagnostic test 2 is superior to that of diagnostic test 1; if Z exceeds +1.96, then we conclude that the accuracy of diagnostic test 1 is superior to that of diagnostic test 2.

A two-sided CI for the difference in AUC (or PAUC) between two diagnostic tests can be calculated from

$$LL = [AUC_1 - AUC_2] - z_{\alpha/2} \times \sqrt{[var_1 + var_2 - 2 \times cov]} \qquad (5)$$

$$UL = [AUC_1 - AUC_2] + z_{\alpha/2} \times \sqrt{[var_1 + var_2 - 2 \times cov]}, \qquad (6)$$

where LL is the lower limit of the CI, UL is the upper limit, and $z_{\alpha/2}$ is a value from the standard normal distribution corresponding to a probability of $\alpha/2$. For example, to construct a 95% CI, $\alpha = 0.05$, thus $z_{\alpha/2} = 1.96$.

Consider the ROC curves in Figure 2A. The estimated areas under the smooth ROC curves of the two tests are the same, 0.841. The PAUCs where the FPR is greater than 0.20, however, differ. From the estimated variances and covariance in Table 3, the value of the Z statistic for comparing the PAUCs is 1.77, which is not statistically significant. The 95% CI for the difference in PAUCs is more informative: (–0.004 to 0.086); the CI for the partial area index is (–0.02 to 0.43). The CI contains large positive differences, suggesting that more research is needed to investigate the relative accuracies of these two diagnostic tests for FPRs less than 0.20.

Analysis of MRMC ROC Studies

Multiple published methods discuss performing the statistical analysis of MRMC studies [13–20]. The methods are used to construct CIs for diagnostic accuracy and statistical tests for assessing differences in accuracy between tests. A statistical overview of the methods is given elsewhere [10]. Here, we briefly mention some of the key issues of MRMC ROC analyses.

Fixed- or random-effects models.—The MRMC study has two samples, a sample of patients and a sample of reviewers. If the study results are to be generalized to patients similar to ones in the study sample and to reviewers similar to ones in the study sample, then a statistical analysis that treats both patients and reviewers as random effects should be used [13, 14, 17–20]. If the study results are to be generalized to just patients similar to ones in the study sample, then the patients are treated as random effects but the reviewers should be treated as fixed effects [13–20]. Some of the statistical methods can treat reviewers as either random or fixed, whereas other methods treat reviewers only as fixed effects.

Parametric or nonparametric.—Some of the methods rely on models that make strong assumptions about how the accuracies of the reviewers are correlated and distributed (parametric methods) [13, 14], other methods are more flexible [15, 20], and still others make no assumptions [16–19] (nonparametric methods). The parametric methods may be more powerful when their assumptions are met, but often it is difficult to determine if the assumptions are met.

Covariates.—Reviewers' accuracy may be affected by their training or experience or by characteristics of the patients (e.g., age, sex, stage of disease, comorbidities). These variables are called covariates. Some of the statistical methods [15, 20] have models that can include covariates. These models provide valuable insight into the variability between reviewers and between patients.

Software.—Software is available for public use for some of the methods [32, 42, 43]; the authors of the other methods may be able to provide software if contacted.

Determining Sample Size for ROC Studies

Many issues must be considered in determining the number of patients needed for an ROC study. We list several of the key issues and some useful references here, followed by a simple illustration. Software is also available for determining the required sample size for some ROC study designs [32, 41].

1. Is it a MRMC ROC study? Many radiology studies include more than one reviewer but are not considered MRMC studies. MRMC studies usually involve five or more reviewers and focus on estimating the average accuracy of the reviewers. In contrast, many radiology studies include two or three reviewers to get some idea of the interreviewer variability. Estimation of the required sample size for MRMC studies requires balancing the number of reviewers in the reviewer sample with the number of patients in the patient sample. See [14, 44] for formulae for determining sample sizes for MRMC studies and [45] for sample size tables for MRMC studies. Sample size determination for non-MRMC studies is based on the number of patients needed.

2. Will the study involve a single diagnostic test or compare two or more diagnostic tests? ROC studies comparing two or more diagnostic tests are common. These studies focus on the difference between AUCs or PAUCs of the two (or more) diagnostic tests. Sample size can be based on either planning for enough statistical power to detect a clinically important difference, or constructing a CI for the difference in accuracies that is narrow enough to make clinically relevant conclusions from the study. In studies of one diagnostic test, we often focus on the magnitude of the test's AUC or PAUC, basing sample size on the desired width of a CI.

3. If two or more diagnostic tests are being compared, will it be a paired or unpaired study design, and are the accuracies of the tests hypothesized to be different or equivalent? Paired designs almost always require fewer patients than an unpaired design, and so are used whenever they are logistically, ethically, and financially feasible. Studies that are performed to determine whether two or more tests have the same accuracy are called equivalency studies. Often in radiology a less invasive diagnostic test, or a quicker imaging sequence, is developed and compared with the standard test. The investigator wants to know if the test is similar in accuracy to the standard test. Equivalency studies often require a larger sample size than studies in which the goal is to show that one test has superior accuracy to another test. The reason is that to show equivalence the investigator must rule out all large differences between the tests—that is, the CI for the difference must be very narrow.

4. Will the patients be recruited in a prospective or retrospective fashion? In prospective designs, patients are recruited based on their signs or symptoms, so at the time of recruitment it is unknown whether the patient has the disease of interest. In contrast, in retrospective designs patients are recruited based on their known true disease status (as determined by the gold or reference standard) [2]. Both studies are used commonly in radiology. Retrospective studies often require fewer patients than prospective designs.

5. What will be the ratio of nondiseased to diseased patients in the study sample? Let k denote the ratio of the number of nondiseased to diseased patients in the study sample. For retrospective studies k is usually decided in the design phase of the study. For prospective designs k is unknown in the design phase but can be estimated by $(1 - \text{PREV}_p) / \text{PREV}_p$, where PREV_p is the prevalence of disease in the relevant population. A range of values for PREV_p should be considered when determining sample size.

6. What summary measure of accuracy will be used? In this article we have focused mainly on the AUC and PAUC, but others are possible (see [2]). The choice of summary measures determines which variance function formula will be used in calculating sample size. Note that the variance function is related to the variance by the following formula: variance = VF / N, where VF is the variance function and N is the number of study patients with disease.

7. What is the conjectured accuracy of the diagnostic test? The conjectured accuracy is needed to determine the expected difference in accuracy between two or more diagnostic tests. Also, the magnitude of the accuracy affects the variance function. In the following example, we present the variance function for the AUC; see Zhou et al. [2] for formulae for other variance functions.

Consider the following example. Suppose an investigator wants to conduct a study to determine if MRI can distinguish benign from malignant breast lesions. Patients with a suspicious lesion detected on mammography will be prospectively recruited to undergo MRI before biopsy. The pathology results will be the reference standard. The MR images will be interpreted independently by two reviewers; they will score the lesions using a 0–100% confidence scale. An ROC curve will be constructed for each reviewer; AUCs will be estimated, and 95% CIs for the AUCs will be constructed. If MRI shows some promise, the investigator will plan a larger MRMC study.

The investigator expects 20–40% of patients to have pathologically confirmed breast cancer (PREV_p = 0.2–0.4); thus, k = 1.5–4.0. The investigator expects the AUC of MRI to be approximately 0.80 or higher. The variance function of the AUC often used for sample size calculations is as follows:

$$VF = (0.0099 \times e^{-A \times A/2}) \times [(5 \times A^2 + 8) + (A^2 + 8) / k], \qquad (7)$$

where A is the parameter from the binormal distribution. Parameter A can be calculated from $A = \phi^{-1}(\text{AUC}) \times 1.414$, where ϕ^{-1} is the inverse of the cumulative normal distribution function [2]. For our example, AUC = 0.80; thus $\phi^{-1}(0.80) = 0.84$ and $A = 1.18776$. The variance function, VF, equals $(0.00489) \times [(15.05387) + (9.41077) / 4.0] = 0.08512$, where we have set $k = 4.0$. For $k = 1.5$, the $VF = 0.10429$.

Suppose the investigator wants a 95% CI no wider than 0.10. That is, if the estimated AUC from the study is 0.80, then the lower bound of the CI should not be less than 0.75 and the upper bound should not exceed 0.85. A formula for calculating the required sample size for a CI is

$$N = [z_{\alpha/2}{}^2 \times VF] / L^2 \qquad (8)$$

where $z_{\alpha/2} = 1.96$ for a 95% CI and L is the desired half-width of the CI. Here, $L = 0.05$. N is the number of patients with disease needed for the study; the total number of patients needed for the study is $N \times (1 + k)$. For our example, N equals $[1.96^2 \times 0.08512] / 0.05^2 = 130.8$ for $k = 4.0$, and 160.3 for $k = 1.5$. Thus, depending on the unknown prevalence of breast cancer in the study sample, the investigator needs to recruit perhaps as few as 401 total patients (if the sample prevalence is 40%) but perhaps as many as 654 (if the sample prevalence is only 20%).

Finding the Optimal Point on the Curve

Metz [46] derived a formula for determining the optimal decision threshold on the ROC curve, where "optimal" is in terms of minimizing the overall costs. "Costs" can be defined as monetary costs, patient morbidity and mortality, or both. The slope, m, of the ROC curve at the optimal decision threshold is

$$m = (1 - \text{PREV}_p) / \text{PREV}_p \times [C_{FP} - C_{TN}] / [C_{FN} - C_{TP}] \qquad (9)$$

where C_{FP}, C_{TN}, C_{FN}, and C_{TP} are the costs of false-positive, true-negative, false-negative, and true-positive results, respectively. Once m is estimated, the optimal decision threshold is the one for which sensitivity and specificity maximize the following expression: [sensitivity – m(1 – specificity)] [47].

Examining the ROC curve labeled X in Figure 2, we see that the slope is very steep in the lower left where both the sensitivity and FPR are low, and is close to zero at the upper right where the sensitivity and FPR are high. The slope takes on a high value when the patient is unlikely to have the disease or the cost of a false-positive is large; for these situations, a low FPR is optimal. The slope takes on a value near zero when the patient is likely to have the disease or treatment for the disease is beneficial and carries little risk to healthy patients; in these situations, a high sensitivity is optimal [3]. A nice example of a study using this equation is given in [48]. See also work by Greenhouse and Mantel [49] and Linnet [50] for determining the optimal decision threshold when a desired level for the sensitivity, specificity, or both is specified a priori.

Conclusion

Applications of ROC curves in the medical literature have increased greatly in the past few decades, and with this expansion many new statistical methods of ROC analysis have been developed. These include methods that correct for common biases like verification bias and imperfect gold standard bias, methods for combining the information from multiple diagnostic tests (i.e., optimal combinations of tests) and multiple studies (i.e., meta-analysis), and methods for analyzing clustered data (i.e., multiple observations from the same patient). Interested readers can search directly for these statistical methods or consult two recently published books on ROC curve analysis and related topics [2, 39]. Available software for ROC analysis allows investigators to easily fit, evaluate, and compare ROC curves [41, 51], although users should be cautious about the validity of the software and check the underlying methods and assumptions.

Acknowledgments

I thank the two series' coeditors and an outside statistician for their helpful comments on an earlier draft of this manuscript.

References

1. Weinstein S, Obuchowski NA, Lieber ML. Clinical evaluation of diagnostic tests. *AJR* 2005;184:14–19
2. Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine.* New York, NY: Wiley-Interscience, 2002
3. Metz CE. Some practical issues of experimental design and data analysis in radiologic ROC studies. *Invest Radiol* 1989;24:234–245
4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36
5. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989;9:190–195
6. Hilden J. The area under the ROC curve and its competitors. *Med Decis Making* 1991;11:95–101
7. Hilden J. Prevalence-free utility-respecting summary indices of diagnostic power do not exist. *Stat Med* 2000;19:431–440
8. Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. *Acad Radiol* 2001;8:328–334
9. 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
10. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader multicase receiver operating characteristic analysis: an empirical comparison of five methods. *Acad Radiol* 2004;11:980–995
11. Obuchowski NA. Multi-reader ROC studies: a comparison of study designs. *Acad Radiol* 1995;2:709–716
12. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. *Acad Radiol* 1997;4:587–600
13. 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
14. Obuchowski NA. Multi-reader multi-modality ROC studies: hypothesis testing and sample size estimation using an ANOVA approach with dependent observations. with rejoinder. *Acad Radiol* 1995;2:S22–S29
15. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Stat Med* 1996;15:1807–1826
16. Song HH. Analysis of correlated ROC areas in diagnostic testing. *Biometrics* 1997;53:370–382
17. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Acad Radiol* 2000;7:341–349
18. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structure across modalities. *Acad Radiol* 2001;8:605–615
19. Beiden SV, Wagner RF, Campbell G, Chan HP. Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis. *Acad Radiol* 2001;8:616–622
20. Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Can J Stat* 2000;28:731–750
21. 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
22. Kim MJ, Lim JS, Oh YT, et al. Preoperative MRI of rectal cancer with and without rectal water filling: an intraindividual comparison. *AJR* 2004;182:1469–1476
23. Osada H, Kaku K, Masuda K, Iitsuka Y, Seki K, Sekiya S. Quantitative and qualitative evaluations of fetal lung with MR imaging. *Radiology* 2004;231:887–892
24. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics* 2000;1:123–140

25. Zheng B, Leader JK, Abrams G, et al. Computer-aided detection schemes: the effect of limiting the number of cued regions in each case. *AJR* 2004;182:579–583
26. Chakraborty DP, Winter LHL. Free-response methodology: alternative analysis and a new observer-performance experiment. *Radiology* 1990;174:873–881
27. Chakraborty DP. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med Phys* 1989;16:561–568
28. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996;23:1709–1725
29. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. *Acad Radiol* 2000;7:516–525
30. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory: a direct solution. *Psychometrika* 1968;33:117–124
31. Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal detection theory and determination of confidence intervals: rating method data. *J Math Psychol* 1969;6:487–496
32. ROCKIT and LABMRMC. Available at: xray.bsd.uchicago.edu/krl/KRL_ROC software_index.htm. Accessed December 13, 2004
33. Swets JA. Empirical RO. Cs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychol Bull* 1986;99:181–198
34. Hanley JA. The robustness of the binormal assumption used in fitting ROC curves. *Med Decis Making* 1988;8:197–203
35. Goddard MJ, Hinberg I. Receiver operating characteristic (ROC) curves and non-normal data: an empirical study. *Stat Med* 1990;9:325–337
36. Zou KH, Tempany CM, Fielding JR, Silverman SG. Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. *Acad Radiol* 1998;5:680–687
37. Pepe MS. A regression modeling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika* 1997;84:595–608
38. Pepe MS. An interpretation for the ROC curve using GLM procedures. *Biometrics* 2000;56:352–359
39. Pepe MS. *The statistical evaluation of medical tests for classification and prediction***.** New York, NY: Oxford University Press, 2003
40. 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
41. ROC analysis. Available at: www.bio.ri.ccf.org/Research/ROC/index.html. Accessed December 13, 2004
42. OBUMRM. Available at: www.bio.ri.ccf.org/OBUMRM/OBUMRM.html. Accessed December 13, 2004
43. The University of Iowa Department of Radiology: The Medical Image Perception Laboratory. MRMC 2.0. Available at: perception.radiology.uiowa.edu. Accessed December 13, 2004
44. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. *Acad Radiol* (in press)
45. Obuchowski NA. Sample size tables for receiver operating characteristic studies. *AJR* 2000;175:603–608
46. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283–298
47. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561–577
48. Somoza E, Mossman D. “Biological markers” and psychiatric diagnosis: risk-benefit balancing using ROC analysis. *Biol Psychiatry* 1991;29:811–826
49. Greenhouse SW, Mantel N. The evaluation of diagnostic tests. *Biometrics* 1950;6:399–412
50. Linnet K. Comparison of quantitative diagnostic tests: type I error, power, and sample size. *Stat Med* 1987;6:147–158
51. Stephan C, Wesseling S, Schink T, Jung K. Comparison of eight computer programs for receiver-operating characteristic analysis. *Clin Chem* 2003;49:433–439
52. Ma G, Hall WJ. Confidence bands for receiver operating characteristic curves. *Med Decis Making* 1993;13:191–197

APPENDIX 1. Area Under the Curve and Confidence Intervals with Binormal Model

Under the binormal assumption, the receiver operating characteristic (ROC) curve is the collection of points given by

$$[1 - \phi(c), 1 - \phi(B \times c - A)]$$

where c ranges from $-\infty$ to $+\infty$ and represents all the possible values of the underlying binormal distribution, and ϕ is the cumulative normal distribution evaluated at c. For example, for a false-positive rate of 0.10, $\phi(c)$ is set equal to 0.90; from tables of the cumulative normal distribution, we have $\phi(1.28) = 0.90$. Suppose $A = 2.0$ and $B = 1.0$; then the sensitivity $= 1 - \phi(- 0.72) = 1 - 0.2358 = 0.7642$.

ROCKIT [32] gives a confidence interval (CI) for sensitivity at particular false-positive rates (i.e., pointwise CIs). A CI for the entire ROC curve (i.e., simultaneous CI) is described by Ma and Hall [52].

Under the binormal distribution assumption, the area under the smooth ROC curve (AUC) is given by

$$\text{AUC} = \phi[A / \sqrt{(1 + B^2)}].$$

For the example above, $\text{AUC} = \phi[2.0 / \sqrt{(2.0)}] = \phi[1.414] = 0.921$.

The variance of the full area under the ROC curve is given as standard output in programs like ROCKIT [32]. An estimator for the variance of the partial area under the curve (PAUC) was given by McClish [5]; a Fortran program is available for estimating the PAUC and its variance [41].

The reader’s attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005

Fundamentals of Clinical Research for Radiologists

Lawrence Joseph[1,2]
Caroline Reinhold[3,4]

Statistical Inference for Continuous Variables

Consider the following statements from an abstract reporting results from a study of CT in large cell neuroendocrine carcinoma of the lung [1]:

> In the 38 patients, six central tumors and 32 peripheral tumors, with diameters ranging from 12 to 92 mm (mean ± SD, 32 ± 19 mm), were identified. None of the tumors had air bronchograms or calcification in the mass or nodule… On contrast-enhanced CT scans, inhomogeneously enhanced tumors appeared to be larger (51 ± 18 mm) than homogeneously enhanced tumors (25 ± 10 mm; $p < 0.001$).

Proper interpretation of the above results, and of similar reports from much of the modern clinical literature, depends in large part on the understanding of statistical terms. In this case, terms such as "SD" were used for descriptive purposes, and p values were given to support evidence of between-group differences in tumor size. In other reports, one may see terms such as "confidence intervals," "t tests," "type 1 and type 2 errors," and so on. Clearly, radiologists who wish to keep pace with new technologies must at least have a basic understanding of statistical language. This is true not only if they desire to plan and perform their own research, but also if they simply want to read the medical literature with a keen critical eye or to make informed decisions about which new treatments or diagnostic techniques they may wish to use to treat their own patients.

Descriptive terms such as "means," "medians," and "SDs" have been covered in a previous article in this series [2]. Before reading this article, reviewing the previous modules on descriptive statistics [2] and probability and sampling [3] may be a good idea. In this module, we introduce the basic notions of inferential statistics—that is, we discuss how to draw inferences about one or more populations' characteristics using data from samples from these populations. We focus on continuous variables, including inferences for means and simple nonparametric methods. Rather than simply providing a catalogue of which formulas to use in which situation, we explain the logic behind each technique. In this way, informed choices and decisions can be made on the basis of a deeper understanding of exactly what information each type of statistical inference provides.

Recall from the discussion in a previous module [3] that there are two main schools of statistical inference: the frequentist school and the Bayesian school. These are each based on a different definition of probability, the frequentist school based on a long-run frequency definition and the Bayesian school based on a more subjective view of probability. We discuss these paradigms for statistical inference.

In the Statistical Inferences for Means section, the classical or frequentist school of statistical inferences for means is covered, and in the Nonparametric Inference section, we present a brief introduction to nonparametric inferences. In these sections, we explain exactly what is meant by ubiquitous statistical statements such as "$p < 0.05$"—which may not mean what many medical journal readers believe it to mean—and examine confidence intervals as an attractive alternative to p values. The problem of choosing an appropriate sample size for a given experiment is discussed in the Sample Size Calculations section. Increasingly important Bayesian alternatives to the

Received November 5, 2004; accepted after revision November 11, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 15th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology.* The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1] Department of Medicine, Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Ave., Montreal, QC H3G 1A4, Canada. Address correspondence to L. Joseph (Lawrence.Joseph@mcgill.ca).

[2] Department of Epidemiology and Biostatistics, 1020 Pine Ave. W, McGill University, Montreal, QC H3A 1A2, Canada.

[3] Department of Diagnostic Radiology, Montreal General Hospital, McGill University Health Centre, 1650 Cedar Ave., Montreal, QC H3G 1A4, Canada.

[4] Department of Oncology, Synarc, 575 Market St., San Francisco, CA 94105.

AJR 2005;184:1047–1056

0361–803X/05/1844–1047

classical statistical techniques are presented in the Bayesian Inference section.

Statistical Inferences for Means

In this section, we consider how to draw inferences about populations by statistically analyzing samples of data using standard frequentist methods. We first consider inferences for a single mean when the variance in the population is known. We also initially assume that the data follow a normal distribution, so we are estimating the mean of this normal distribution. Once the basic concepts are understood in this simple case, we indicate how to extend the same ideas to cases in which the variance is unknown or more than one mean is of interest and to cases in which the normal distribution is not assumed.

In addition to the two different schools of inference (i.e., frequentist or Bayesian), statistical inferences can be divided into procedures that test a hypothesis and those that estimate parameters. We begin with hypothesis testing procedures that lead to *p* values, and then compare the information they provide to that provided by parameter estimation via confidence intervals.

Standard Frequentist Hypothesis Testing

Suppose we wish to test the hypothesis that a new accelerated radiation schedule for patients with brain cancer leads to smaller mean tumor diameters compared with the standard schedule versus a null hypothesis that the tumor diameters are the same regardless of schedule. Suppose further that it is known that patients on the standard schedule have a tumor diameter of 3.5 cm, on average, after completing their radiation therapy. Although it is somewhat unrealistic to assume this perfect knowledge of past tumor diameters, this example approximates the situation in which a large case series (e.g., a historical control series) of tumor diameters is available, so that most uncertainty arises from the data from the new schedule. Formally, we can state the hypotheses as:

H_0 (null hypothesis): $\mu = 3.5$

H_A (alternative hypothesis): $\mu < 3.5$

where μ represents the unknown true average tumor diameter of the accelerated radiation schedule.

There are four possible results when considering hypothesis testing, depending on the true state of nature, which is typically unknown, and the statistical test result, which depends on the data collected. The four possibilities are shown in Table 1.

According to Table 1, if the accelerated schedule in fact leads to smaller tumor diameters than the standard and we reject the null hypothesis, then we have made a correct decision, as also happens if the null hypothesis is in fact correct and we do not reject it. On the other hand, if we reject the null hypothesis as false when it is in fact true, we make a so-called type 1 error, which occurs with probability α, and if we fail to reject the null hypothesis when it is in fact false, we make a type 2 error, which occurs with probability β. The power of a study is defined as the probability of rejecting the null hypothesis when the alternative hypothesis is in fact true, so that the power is equal to 1 – β. To summarize, we have equations 1–4:

$$\alpha = Pr\{ \text{ rejecting } H_0 | H_0 \text{ is true}\} = \text{Type I error} \tag{1}$$

$$1 - \alpha = Pr\{\text{not rejecting } H_0 | H_0 \text{ is true}\} \tag{2}$$

$$\beta = Pr\{\text{not rejecting } H_0 | H_A \text{ is true}\} = \text{Type II error, and} \tag{3}$$

$$1 - \beta = Pr\{ \text{ rejecting } H_0 | H_A \text{ is true}\} = \text{Power} \tag{4}$$

Recall from a previous module in this series [3] that probabilities written in the form of Pr{A | B} are called "conditional probabilities," and the notation is read as the probability that the event A occurs, given that the event B is known to have occurred. Thus, all of the quantities are conditional on knowing whether the null or alternative hypotheses are in fact true. Of course, we generally do not know whether the null hypothesis is true or not, so these conditional statements are at best of indirect interest. Once we obtain our data, we would ideally like to know the probability that the null hypothesis is true—not assume the null hypothesis is true. We will discuss this point further in the Bayesian Inference section.

Although it is important to understand the types of errors that can be made when hypothesis testing, the results of a hypothesis test are usually reported as a *p* value, which we now define: The *p* value is the probability of obtaining a result as extreme as or more extreme than that observed assuming that the null hypothesis is in fact true.

TABLE 1 Results of Hypothesis Testing

Test	True State of Nature	
	H_A	H_0
Reject H_0	$1 - \beta$	α
Do not reject H_0	β	$1 - \alpha$

It is important to note that the *p* value is not the probability that the null hypothesis is true after having seen the data, even though many clinicians often falsely interpret it this way. The *p* value does not directly or indirectly provide this probability and in fact can be orders of magnitude different from it. In other words, it is possible to have a *p* value equal to 0.05, when the probability of the null hypothesis is 0.5, different from the *p* value by a factor of 10 (see the Bayesian Inference section for how to calculate a more easily interpreted hypothesis test from a Bayesian viewpoint).

Given the definition of a *p* value, how would we calculate it? Suppose that we perform tumor measurements on 10 patients under the accelerated schedule and that these tumors have a mean diameter of $\bar{x}$ = 3.0 cm, with a known SD of σ = 1.5 cm. The definition implies that we need to calculate the probability of obtaining mean tumor diameters of 3.0 cm or less (i.e., as extreme as or more extreme than what was observed), given that the true mean tumor diameter under the standard treatment schedule is exactly 3.5 cm (i.e., given the null hypothesis is true). Now, recall from a previous article in this series [3] that the probability density of our mean, $\bar{x}$, is usually considered as normal. Because for purposes of calculating a *p* value the null hypothesis is considered as exactly correct, the mean of our normal distribution is assumed to be 3.5 cm. The SD of our mean (known as the SE) is given as the SD in the population (assumed here to be σ = 1.5 cm) divided by the square root of the sample size [3]. Thus here, our SE is given by 1.5 / √ 10 = 0.474.

Therefore, we calculate equations 5 and 6:

$$p = Pr\{\text{of obtaining data as or more extreme than observed} | H_0 : \mu = 3.5\} \tag{5}$$

$$= Pr\{\bar{x} < 3.0 | \bar{x} \sim N(3.5, 0.474)\}. \tag{6}$$

This probability can be calculated from tables of the normal distribution, as explained in Joseph and Reinhold [3]. Normalizing, we find *Z* = [(3.0 – 3.5) / 0.474] = –1.05, and looking up −1.05 on standard normal tables, we find *p* = 0.147. Thus, there is about a 14.7% chance of obtaining results as extreme as or more extreme than the 3.0 cm observed, if the true mean tumor diameter for the new schedule is exactly 3.5 cm. Therefore, the observed result is not unusual (i.e., it is compatible with the null hypothesis), so we cannot reject H_0.

Notice that if we had observed the same mean tumor diameter but with a larger sample size of 100, say, the p value would have been 0.0004. With a sample size of 100, the event of the observed data or data more extreme occurring would be a rare event if the null hypothesis were true, so the null hypothesis could be rejected. Therefore, p values depend not only on the observed mean tumor diameter, but also on the sample size.

The test described earlier was one-sided—that is, we a priori believed (perhaps from preliminary data or theoretic considerations) that the accelerated schedule would lead to equal or better results and not larger tumor sizes. To generalize, to perform a one-sided test of the null hypothesis that a single mean μ has value μ_0, calculate the statistic in equation 7:

$$z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{7}$$

and determine the p value from normal distribution tables as in equation 8:

$$p = Pr\{z > z^* | z \sim N(0,1)\} \tag{8}$$

On the other hand, often we may not want to specify the direction ahead of time. In this case, the alternative hypothesis is two-sided (i.e., the alternative hypothesis is H_A: $\mu \neq \mu_0$ rather than the one-sided H_A: $\mu < \mu_0$), and one performs the calculation in equation 9:

$$z^* = \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \tag{9}$$

where | a | indicates the absolute value of a, and one determines the p value from normal distribution tables as in equation 10:

$$p = 2 \times Pr\{z > z^* | z \sim N(0,1)\} \tag{10}$$

In the one-sided case, we reject the null hypothesis only if we observe an extreme result in the direction specified by the alternative hypothesis. In the two-sided case, we reject if we observe an extreme result in either direction (larger or smaller tumor sizes). This results in a doubling of the p value, so for a two-sided alternative hypothesis (H_A: $\mu \neq 3.5$ in this case), we find $p = 2 \times 0.147 = 0.294$. The doubling results from adding the areas under the normal curve both below −1.05 (as in the one-sided case) and above 1.05.

Similar methods are available for tests involving comparisons between two means. For example, to test the null hypothesis that means in two different groups are equal to each other versus a two-sided alternative hypothesis, calculate as in equation 11:

$$z^* = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| \tag{11}$$

For example, suppose we wish to again look at the difference in mean tumor diameter between two groups of patients with brain cancer, but this time in a clinical trial setting, with subjects randomized into accelerated and standard schedule groups (this would, of course, be a better design because concurrent groups are compared, minimizing potential confounding). Suppose we observe a mean tumor diameter of $\bar{x}_1 = 3.0$ cm ($\sigma_1 = 1.5$ cm) in 200 subjects under the new schedule, and a mean tumor diameter of $\bar{x}_2 = 3.7$ cm ($\sigma_2 = 1.4$ cm) in 200 subjects under the standard schedule. Plugging into the above formula, we get equation 12:

$$z^* = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| = \left| \frac{3.0 - 3.7}{\sqrt{\frac{1.5^2}{200} + \frac{1.4^2}{200}}} \right| = 4.82 \tag{12}$$

Looking up 4.82 on normal tables gives a p value of 2 × (0.0000007) = 0.0000014. Because this indicates a very rare event under H_0, we can reject the null hypothesis that the two means are equal.

These formulas can be extended in a variety of directions, which we describe in the subsequent sections.

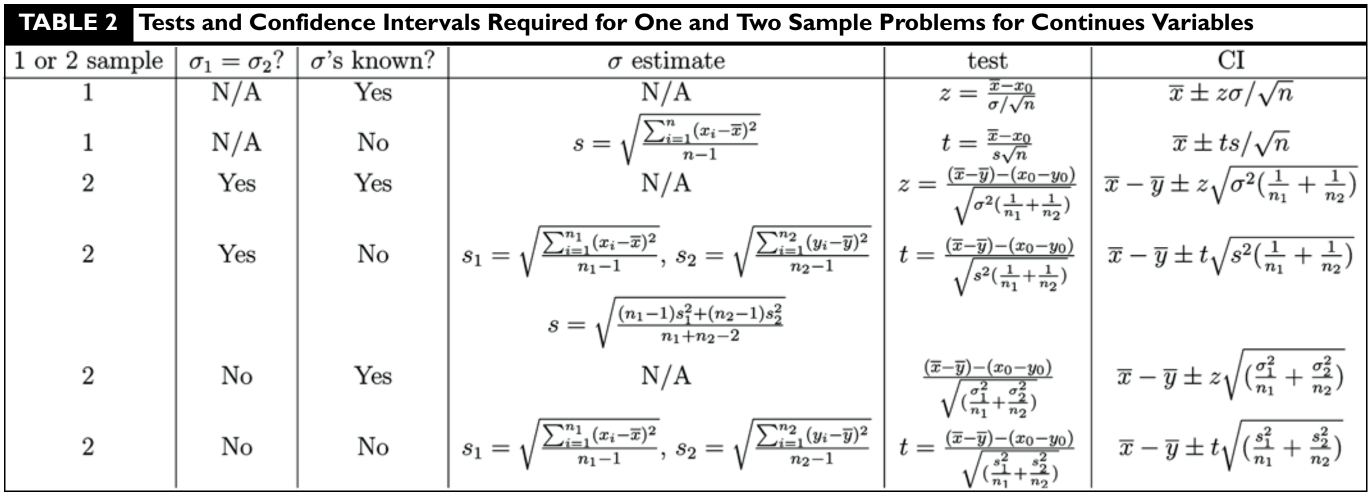
Paired versus unpaired tests.—In comparing the two mean tumor diameters, we have assumed that the design of this study was unpaired, meaning that the data were composed of two independent samples, one from each treatment group. In some experiments, for example, if one wishes to compare quality of life before and after any medical procedure is performed, a paired design is appropriate because the patient is being compared with him- or herself—that is, the patient serves as his or her own control. Here, one would subtract the value measured on an appropriate quality-of-life scale before the procedure from that measured on the same scale after the procedure to create a single set of before-to-after differences. Once this subtraction has been done for each patient, one in fact has reduced the two measures on each patient (i.e., before and after) to a single set of numbers representing the differences. Therefore, paired data can be analyzed using the same formulas as used for single-sample analyses. Paired designs are often more efficient than unpaired designs, as between-group variability is reduced by the pairing.

Assumptions behind the Z tests.—For ease of exposition, we have presented all of the test formulas using percentiles that came from the normal distribution, but in practice there are two assumptions behind this use of the normal distribution. The first assumption is that the data arise either from a normal distribution or the sample size is large enough for the central limit theorem [3] to apply. The second assumption is that the variance or variances involved in the calculations are known exactly.

The first of these assumptions is often satisfied at least approximately in practice, but the second assumption almost never holds in real applications. We usually have to use estimates s^2, s_1^2, and s_2^2 in the above formulas rather than the exact values σ^2, σ_1^2, and σ_2^2, respectively, because the variances would usually be estimated from the data rather than being known exactly. To account for the extra uncertainty due to the fact that the variance is estimated rather than known, the distribution of our test statistic changes. We thus use t distribution tables rather than normal distribution tables. In calculations, this means that the z values used in all of the formulas need to be switched to the corresponding values from t tables. This requires knowledge of the degrees of freedom (*df*), which for single-mean problems is simply the sample size minus 1. This of course applies to paired designs as well, because they reduce to single-sample problems. For two sample unpaired problems, a conservative number for the *df* is the minimum of the two sample sizes minus 1 ($n - 1$, where n is the sample size) [4]. These tests are called t tests.

Equal or unequal variances.—The tests described earlier assume that the variances in the two groups are unequal. Slightly more efficient formulas can be derived if the variances are the same, as a single pooled estimate of the variance can be derived from combining the information in both samples together. We do not discuss pooled variances further here, in part because in practice the difference in analyses done with pooled or unpooled variances is usually quite small and

TABLE 2 Tests and Confidence Intervals Required for One and Two Sample Problems for Continues Variables



1 or 2 sample	$\sigma_1 = \sigma_2$?	σ's known?	σ estimate	test	CI
1	N/A	Yes	N/A	$z = \frac{\overline{x}-x_0}{\sigma/\sqrt{n}}$	$\overline{x} \pm z\sigma/\sqrt{n}$
1	N/A	No	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$	$t = \frac{\overline{x}-x_0}{s\sqrt{n}}$	$\overline{x} \pm ts/\sqrt{n}$
2	Yes	Yes	N/A	$z = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}}$	$\overline{x} - \overline{y} \pm z\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}$
2	Yes	No	$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2}{n_1-1}}$, $s_2 = \sqrt{\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{n_2-1}}$	$t = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}}$	$\overline{x} - \overline{y} \pm t\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}$
			$s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$		
2	No	Yes	N/A	$\frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}}$	$\overline{x} - \overline{y} \pm z\sqrt{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}$
2	No	No	$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2}{n_1-1}}$, $s_2 = \sqrt{\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{n_2-1}}$	$t = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})}}$	$\overline{x} - \overline{y} \pm t\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}$

Note.—In all cases, the data are assumed to be normally distributed or the sample size large enough for the central limit theorem to apply. The data are assumed to be represented by x_i, $i = 1,\ldots, n$ for a single-sample problem or by x_i, $i = 1,\ldots, n_1$ and y_i, $i = 1,\ldots, n_2$ for a two-sample problem. Sample sizes are n for a single-sample problem and n_1 and n_2 for the two-sample problem. The z indicates a normal table is used, t indicates a t table is required. When a t table is required, the degrees of freedom are equal to $n - 1$ for a single-sample problem, while the degrees of freedom are $n_1 + n_2 - 2$ for a two-sample problem with equal variances, and $\min(n_1 - 1, n_2 - 1)$ for unequal variances (conservative value). The x_0 and y_0 indicate null values under the null hypothesis (usually but not always equal to zero). For paired two-sample problems, form the within-individual differences, and use the formulas for the one-sample case. N/A = not applicable.

in part because it is rarely appropriate to pool the variances, because the variability is usually not exactly the same in both groups.

Analysis of variance: more than two means.—We have seen tests for one or two means, but sometimes one wishes to test the equality of three or more means. Although this topic is not covered here, readers should be aware that analysis of variance is a technique that extends our two-sample procedure to three or more means. See, for example, Armitage and Berry [5] or Rosner [6] for details.

Table 2 provides the test statistics used for all possible cases with one or two means, as discussed earlier.

How Useful Are p *Values for Medical Decision Making?*

Although p values are still often found in the literature, there are several major problems associated with their use. First, as mentioned earlier, they are often misinterpreted as the probability of the null hypothesis given the data, when in fact they are calculated assuming the null hypothesis to be true. Second, clinicians often use them to dichotomize results into important or unimportant depending on whether $p < 0.05$ or $p > 0.05$, respectively. However, there is not much difference between p values of 0.049 and 0.051, so the cutoff of 0.05 is arbitrary. Third, p values concentrate attention away from the magnitude of treatment differences. For example, one could have a p value that is very small but is associated with a clinically unimportant difference. This is especially prone to occur in cases in which the sample size is large. Conversely, results of potentially great clinical interest are not necessarily ruled out if $p > 0.05$, especially in studies with small sample sizes. Therefore, one should not confuse statistical significance (i.e., $p < 0.05$) with practical or clinical importance. Fourth, the null hypothesis is almost never exactly true. In the example described, does one seriously think that the mean tumor diameter of the patients on the standard treatment schedule could be exactly 3.5 cm (rather than, say, 3.50001 cm)? Because one knows the null hypothesis is almost surely false to begin with, it makes little sense to test it. Instead, one should concern oneself with the question, By how much are the two treatments different?

There are so many problems associated with p values that most statisticians now recommend against their use, in favor of confidence intervals or Bayesian methods. In fact, some prominent journals have virtually banished p values from publication [7], others strongly discourage their use [8], and many others have published articles and editorials encouraging the use of Bayesian methodology [9, 10]. We cover these more informative techniques for drawing statistical inferences, starting with confidence intervals.

Frequentist Confidence Intervals

Although the p value provides some information concerning the rarity of events as extreme as or more extreme than that observed assuming the null hypothesis to be exactly true, it provides no information about what the true parameter values might be. In the two-mean example described earlier, we observed a tumor diameter difference of 0.7 cm, which was shown to be "statistically significant," with a p value of approximately 0.000001. Although we observed a difference of 0.7 cm, we know that our data are from a random sample of patients to whom this procedure could be applied, so the true mean difference could in fact be higher or lower than our observed difference. How likely is it that the true mean difference in tumor diameter is clinically important?

One way to answer this question is with a confidence interval. The formula in equation 13 provides 95% confidence interval limits for means (the value 1.96 could be changed to other values if intervals with coverage other than 95% are of interest) [3]:

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{13}$$

where $\overline{x}$ is the sample mean and σ is the known SD from a sample of size n. As before, if σ is not known, it is replaced by its estimate from the data, s, and 1.96 is increased somewhat, as a percentile from the t distribution replaces the normal percentile.

Applying this formula to the single-mean example we first discussed, where $\bar{x} = 3.0$, $n = 10$, and $\sigma = 1.5$, we obtain a 95% confidence interval of (2.1–3.9 cm). We cannot conclude very much from this interval because we have not ruled out mean tumor diameters as small as 2.1 cm, which is clinically superior to the 3.5 cm from the old schedule; however, on the other hand, diameters as large as 3.9 cm have also not been ruled out, which is even worse than the tumor diameter in the standard group. Thus, further data would need to be collected before any conclusions could be drawn about this new schedule.

Our two-group clinical trial example had larger sizes, so it will presumably provide a more accurate estimate. We can calculate a 95% confidence interval for the difference in means for the two groups using the formula in equation 14,

$$\left(\bar{x}_1 - \bar{x}_2 - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\; \bar{x}_1 - \bar{x}_2 + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \tag{14}$$

where the same comment regarding unknown variances again applies. Plugging in the values we obtained from our clinical trial example given earlier, we find a confidence interval of –0.46 to –0.94 cm. Thus, roughly speaking, it is likely that the true tumor diameter difference between our two schedules is between approximately 0.5 cm less under the new schedule (–0.46 cm) and up to almost a 1-cm reduction (–0.94 cm). Although our p value for this same data set was small, which enabled us to reject the null hypothesis, we can see that the confidence interval provides more clinically useful information about the magnitude of the difference. We can also see that, in contrast to what may be believed after seeing the p value, we are still uncertain about the clinical utility of the new schedule, because values near the lower limit of the confidence interval would not be interesting clinically—it would represent less than a 30% change from the mean baseline tumor size—while differences near 1 cm may be clinically interesting. Therefore, our conclusions from the confidence interval are more detailed than those from the p value. This is true in general, as we now discuss.

Interpreting Confidence Intervals

Confidence intervals are derived from procedures that are set up to "work" 95% of the time (if a 95% confidence interval is used). The two confidence interval equations discussed earlier provide procedures that, when used repeatedly across different problems, will capture the true value of the mean (or difference in means) 95% of the time and fail to capture the true value 5% of the time. In this sense, we have confidence that the procedure works well in the long run, although in any single application, of course, the interval either does or does not contain the true mean. Note that we are careful not to say that our confidence interval has a 95% probability of containing the true parameter value. For example, we did not say that the true difference in mean tumor diameter is in the interval –0.49 to –0.94 cm with 95% probability. This is because the confidence limits and the true mean tumor diameters are both fixed numbers, and it makes no more sense to say that the true mean is in this interval than it does to say that the number 2 is inside the interval (1, 6) with probability 95%. Of course, 2 is inside this interval, just like the number 8 is outside of the interval (1, 6). However, the procedure used to calculate confidence intervals provides random upper and lower limits that depend on the data collected; in repeated uses of this formula across a range of problems, we expect the random limits to capture the true value 95% of the time and exclude the true mean 5% of the time. Refer to Figure 1. If we look at the set of confidence intervals as a whole, we see that about 95% of them include the true parameter value. However, if we pick out a single trial, it either contains the true value (≈ 95% of the time) or excludes this value (≈ 5% of the time).

Despite their somewhat unnatural interpretation, confidence intervals are generally preferred to p values. This is because confidence intervals focus attention on the range of values compatible with the data on a scale of direct clinical interest. Given a confidence interval, one can assess the clinical meaningfulness of the result, as can be seen in Figure 2.

Depending on where the upper and lower confidence interval limits fall in relation to the upper and lower limits of the region of clinical equivalence, different conclusions should be drawn. The region of clinical equivalence, sometimes called the region of clinical indifference, is the region inside of which two treatments, say, would be considered to be the same for all practical purposes. The point 0, indicating no difference in results between the two treatments, is usually included in the region of clinical equivalence, but values above and below 0 are usually also included. How wide this region is depends on each individual clinical situation. For example, if one treatment schedule is much more expensive than another, one may want at least a 50% reduction in tumor diameter to consider it the preferred treatment.

There are five different conclusions that can be made after a confidence interval has been calculated, as illustrated by the five hypothetic intervals displayed in Figure 2. The first conclusion (interval 1) is that the confidence interval includes zero and that both upper and lower confidence interval limits, if they were the true values, would not be clinically interesting. Therefore, this variable has been shown to have no important effect.

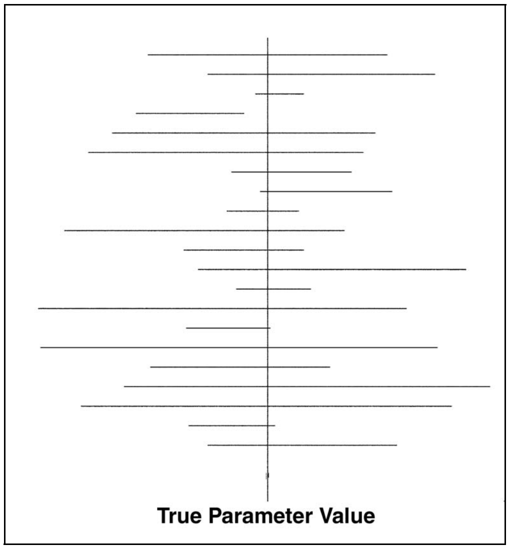


Fig. 1.—Drawing shows series of 95% confidence intervals for unknown parameter.

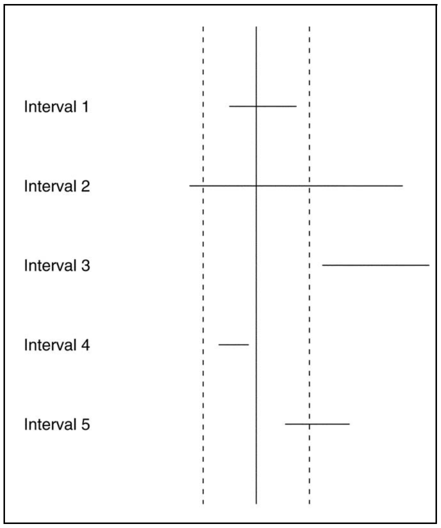


Fig. 2.—Drawing shows how to interpret confidence intervals. Depending on where confidence interval lies in relation to region of clinical equivalence, different conclusions can be drawn.

The second conclusion (interval 2) is that the confidence interval includes zero but that one or both of the upper or lower confidence interval limits, if they were the true values, would be interesting clinically. Therefore, the results of this variable in this study are inconclusive, and further evidence needs to be collected.

The third conclusion (interval 3) is that the confidence interval does not include zero and that all values inside the upper and lower confidence interval limits, if they were the true values, would be clinically interesting. Therefore, this study shows this variable to be important.

The fourth conclusion (interval 4) is that the confidence interval does not include zero but that all values inside the upper and lower confidence interval limits, if they were the true values, would not be clinically interesting. Therefore, this study shows this variable, although having some small effect, is not clinically important.

The fifth conclusion (interval 5) is that the confidence interval does not include zero but that only some of the values inside the upper and lower confidence interval limits, if they were the true values, would be clinically interesting. Therefore, this study shows this variable has at least a small effect and may be clinically important. Further study is required to better estimate the magnitude of this effect.

Revisiting the two confidence intervals discussed earlier in light of Figure 2, we see that the interval based on our single-sample experiment, which ranged from 2.1 to 3.9 cm, is clearly of type 2 and the interval based on the two-group clinical trial is of type 5. Once again, note that these confidence intervals provide much more detailed conclusions than the information contained in a *p* value.

The *p* values group together intervals 1 and 2 as “nonsignificant” and intervals 3, 4, and 5 as “significant.” This can lead to misleading conclusions from a clinical viewpoint. For example, similar clinical conclusions should be drawn from intervals 1 and 4, even though one is “significant” and the other is not. It should now be clear why many journals discourage reporting results in terms of *p* values and encourage confidence intervals.

Summary of Frequentist Statistical Inference

The main tools for statistical inference from the frequentist point of view are *p* values and confidence intervals. The *p* values have fallen out of favor among statisticians, and although they continue to appear in medical journal articles, their use is likely to greatly diminish in the coming years. Confidence intervals provide more clinically useful information than *p* values, so confidence intervals are to be preferred in practice. Confidence intervals still do not allow the formal incorporation of preexisting knowledge into any final conclusions. For example, in some cases there may be compelling medical reasons why a new technique may be better than a standard technique, so if faced with an inconclusive confidence interval, a radiologist may still wish to switch to the new technique, at least until more data become available. On what basis could this decision be justified? We return to this question in the Bayesian Inference section, which appears later in this article.

Nonparametric Inference

Thus far, statistical inferences on populations have been made by assuming a mathematic model for the population (e.g., a normal distribution) and estimating parameters from that distribution based on a sample. Once the parameters have been estimated—for example, the mean or variance for a normal distribution—the distribution is fully specified. This is known as parametric inference.

Sometimes we may be unwilling to specify the general shape of the distribution in advance and prefer to base the inference only on the data, without a parametric model. In this case, we have distribution-free or nonparametric methods.

For example, consider the following data, which represent the tumor diameters of the marker liver metastases for two different chemotherapy regimens in patients with colorectal carcinoma: conventional treatment, 21, 12, 11, 28, 3, 10, 9, 5, 7, 10, 6; new treatment, 4, 3, 4, 5, 20, 22, 5, 12, 15, 5, 1, 14, 13.

Because we are making nonparametric inferences, we no longer refer to tests of similarity of group means. Rather, the null and alternative hypotheses here are defined as follows: For the null hypothesis (H_0), there is no treatment effect—that is, conventional treatment tends to give rise to tumor sizes similar to those from the new treatment. For the alternative hypothesis (H_A), the new treatment tends to give rise to different values for tumor sizes compared with those from the conventional treatment group.

The first step to nonparametrically test these hypotheses is to order and rank the data from lowest to highest values, keeping track of which data points belong to each treatment group, as shown in Table 3.

Thus, in ranking the data, we simply sort the data from the smallest to the largest value regardless of group membership and assign a rank to each data point depending on where its value lies in relation to other values in the data set. Hence, the lowest value receives a rank of 1, the second lowest a rank of 2, and so on. Because there are many “ties” in this data set, we need to rank the data accounting for the ties, which we do by grouping all tied values together and distributing the sum of the available ranks evenly among the tied values. For example, the second and third lowest values in this data set are both 3, and there is a total of five ranks (2 + 3) to be divided among them. Hence, each of these values receives a rank of 2.5 (5 / 2). Similarly, the sixth through ninth values are all tied at 5. There are 30 total ranks (6 + 7 + 8 + 9) to divide up among four tied values, so each receives a value of 7.5 (30 / 4), and so on.

The next step is to sum the ranks for the values belonging to the conventional treatment group, which yields a total of 147.5 (2.5 + 7.5 + 10 + 11 + 12 + 13.5 + 13.5 + 15 + 16.5 + 22 + 24).

We now reason as follows: There is a total of 300 ranks (1 + 2 + 3 + …+ 23 + 24) that can be distributed among the conventional and new treatment groups. If the sample sizes were equal, therefore, and if the null hypothesis were exactly true, we would expect that these ranks should divide equally among the two groups, so each would have a sum of ranks of 150. Now, the sample sizes are not

TABLE 3 First Step to Nonparametrically Test Null and Alternative Hypotheses: Order and Rank the Data

Treatment Group	N	N	C	N	N	N	N	N	C	C	C	C	C	C	C	C	N	N	N	N	N	C	N	C
Data	1	3	3	4	4	5	5	5	5	6	7	9	10	10	11	12	12	13	14	15	20	21	22	28
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Ranks with ties	1	2.5	2.5	4.5	4.5	7.5	7.5	7.5	7.5	10	11	12	13.5	13.5	15	16.5	16.5	18	19	20	21	22	23	24

Note.—N = new treatment, C = conventional treatment.

quite equal, so here we expect 300 × (11 / 24) = 137.5 of the ranks to go to the conventional group, and 300 × (13 / 24) = 162.5 of the ranks to go to the new treatment group. Note that 137.5 + 162.5 = 300, which is the total sum of ranks available. We have in fact observed a sum of ranks of 147.5 in the conventional group, which is higher than expected. Is it high enough that we can reject the null hypothesis? To answer this question, we must refer to computer programs that will calculate the probability of obtaining a sum of ranks of 147.5 or greater given that the null hypothesis of no treatment difference is true (remember the definition of the p value discussed earlier). Most statistical computer packages will perform this calculation, which in this case gives $p = 0.58$. Hence, the null hypothesis cannot be rejected, because our result and those more extreme are not rare under the null hypothesis.

This nonparametric test is called the Wilcoxon's rank sum test. An exactly equivalent test can be based on counts rather than ranks, and it is called the Mann-Whitney test. The Mann-Whitney test always provides the same p value as the Wilcoxon's rank sum test, so either can be used. The analogous parametric test, the unpaired t test for the same data, also gives a p value of 0.58, so the same conclusion is reached.

Because the two tests do not always provide the same conclusions, which of these tests is to be preferred? The answer is situation-specific. Remember that the t test assumes either that the data are from a normal distribution—here, it would imply that the tumor diameters are approximately normally distributed—or that the sample size is large. A histogram would show that the data are skewed toward the right, so that normality is unlikely, and the sample sizes are 11 and 13, hardly large. Hence, in this example the nonparametric test is preferred because the assumptions behind the t test do not seem to hold. In general, if the assumptions required by a parametric test may not hold, a nonparametric test is to be preferred, whereas if the distributional assumptions do likely hold, a parametric test provides slightly increased power compared with a nonparametric test.

The Wilcoxon's rank sum test is appropriate for unpaired designs. A similar test exists for paired designs, called the Wilcoxon's signed rank test. Nonparametric confidence intervals are also available, as are tests for two or more groups, such as the Kruskal-Wallis test. See Sprent [11] for further details about these methods.

Sample Size Calculations

As previously discussed, there has been a strong trend away from hypothesis testing and p values toward the use of confidence intervals in the reporting of results from biomedical research. Because the design phase of a study should be in sync with the analysis that will eventually be performed, sample size calculations should be performed on the basis of ensuring adequate numbers for accurate estimation of important quantities that will be estimated in the study, rather than by power calculations. This distinction is important because it has been shown [12] that sample sizes calculated from a power viewpoint are often insufficient when viewed from a confidence interval viewpoint. In other words, although high power ensures rejection of the null hypothesis with high probability, it does not ensure than the confidence interval will be narrow enough to allow good clinical decision making. Therefore, in this section, we focus on sample size methods based on confidence interval width. For similar methods based on power, see the book by Lemeshow et al. [13].

The question of how accurate is "accurate enough" can be addressed by carefully considering the results you would expect to get (a bit of a catch-22 situation, because if you knew the results you will get, there would be no need to perform the experiment) and making sure your interval will be small enough to land in intervals numbered 1, 3, or 4 of Figure 2. The determination of an appropriate width is a nontrivial exercise, but a reasonable target confidence interval width can usually be found.

For estimating the sample size requirements in experiments involving population means, two different formulas are available, depending on whether there is a single sample or two samples. These are derived by solving for the sample size n in the formulas for the confidence intervals discussed.

Single Sample

Let μ be the mean that is to be estimated, and assume that we wish to estimate μ to an accuracy of a total confidence interval width of w (so that the confidence interval will be $\bar{x} \pm d$, where $2 \times d = w$). Let σ be the SD in the population.

Then the required sample size, n, is given by equation 15,

$$n = \frac{z^2\sigma^2}{d^2} = \frac{4 \times z^2\sigma^2}{w^2} \qquad (15)$$

where, as usual, z is replaced by the appropriate normal distribution quantile (z = 1.96, 1.64, or 2.58 for 95%, 90%, or 99% intervals, respectively).

For example, suppose that we would like to estimate average tumor size to an accuracy of d = 2 mm with a 95% confidence interval and that we expect the patient-to-patient variability will be σ = 10 mm. Then, from the previous formula, we need to perform the calculation in equation 16,

$$n \;=\; \frac{1.96^2 \times 10^2}{2^2} = 96 \qquad (16)$$

rounding up to the next highest integer. The most difficult problem in using this equation is to decide on a value for the SD σ, because it is usually unknown. A conservative approach would be to use the maximum value of σ that seems reasonably likely to occur in the experiment.

Two Samples

Let μ_1 and μ_2 be the means of two populations, and suppose that we would like an accurate estimate of $\mu_1 - \mu_2$. Again assume a total confidence interval width of w (so that again $2 \times d = w$). Let σ_1 and σ_2 be the SD in each population, respectively.

Then the sample size is given in equation 17,

$$n = \frac{z^2(\sigma_1^2 + \sigma_2^2)}{d^2} = \frac{4 \times z^2(\sigma_1^2 + \sigma_2^2)}{w^2} \quad (17)$$

where now n represents the required sample size for each group. As usual, z is chosen as we did earlier and is usually 1.96, corresponding to a 95% confidence interval.

Bayesian Inference

Consider again the single-sample tumor diameter problem introduced in the Statistical Inferences for Means section. Recall that in this example patients undergoing the standard radiation therapy schedule are assumed to have a mean of 3.5 cm, whereas the data collected so far for the new accelerated schedule indicate a mean of 3.0 cm, but are based on only 10 subjects. The frequentist confidence interval was wide, ranging from approximately 2.1 to 3.9 cm, so it has not been particularly helpful in making a decision about which technique to use for the next patient. At this point, with the data being relatively uninformative, the treating physician may decide to be conservative and remain with the standard schedule until more information becomes available about the new schedule or may go with their "gut feeling" as to the likelihood that the new schedule is truly better or not. If there have been data from animal experiments or strong theoretic reasons why the new schedule may be better, there may be temptation to try the new one. Can anything be done to aid in this decision-making process?

Bayesian analysis has several advantages over the standard or frequentist statistical analyses discussed in this article so far, including the ability to formally incorporate relevant information not directly contained in the current data set into any statistical analysis. We will see how this can help with the problem discussed earlier, but first we will cover some basics of Bayesian analysis.

Let us generically denote our parameter of interest by θ. Hence, θ can be a binomial parameter, the mean from a normal distribution, an odds ratio, a set of regression coefficients, and so on. Note in particular that θ can be two or more dimensional. The parameter of interest is sometimes usefully thought of as the "true state of nature."

The three basic elements of any Bayesian analysis are, first, the prior probability distribution, $f(\theta)$. This prior distribution summarizes what is known about θ before the experiment is performed. It is based on a "subjective" assessment of the available past information, so may vary from investigator to investigator.

The second basic element of Bayesian analysis is the likelihood function: $f(x \mid \theta)$. The likelihood function summarizes the information contained in the data, x. For instance, it may be created from a normal distribution for a mean. It is important to realize that Bayesians and frequentists can use the same likelihood function because both need to calculate the probability of data given various values for the parameter θ. The way the likelihood function is used, however, differs between the two paradigms.

The third basic element is the posterior distribution: $f(\theta \mid x)$. The posterior distribution summarizes the information in the data, x, together with the information in the prior distribution. Thus, it summarizes what is known about the parameter of interest θ after the data are collected.

Bayes' theorem, posthumously published by Thomas Bayes [14] in 1763, relates the three quantities: posterior distribution = [likelihood of the data × prior distribution] / a normalizing constant, or using our notation above in equation 18,

$$f(\theta|x) = \frac{f(x|\theta) \times f(\theta)}{\text{a normalizing constant}} \quad (18)$$

or, omitting the normalizing constant in equation 19,

$$f(\theta|x) \propto f(x|\theta) \times f(\theta) \quad (19)$$

where $\propto$ indicates "is proportional to."

Thus, we update the prior distribution to a posterior distribution after seeing the data via Bayes' theorem. The current posterior distribution can be used as a prior distribution for the next study, so Bayesian inference provides a natural way to represent the learning that occurs as science progresses.

Radiologists are already familiar with the Bayesian way of thinking, using it every day in the context of interpreting diagnostic tests. The prior probability used in Bayes' theorem is analogous to the background rate of a condition in the population, which is updated to a positive or negative predictive value (analogous to a posterior distribution) after seeing the results of a diagnostic test (analogous to seeing the data). It is thus just a short step from using predictive values in a clinical setting to using Bayes' theorem in a research setting.

The most contentious element in Bayesian analysis is the need to specify a prior distribution. Because there is no unique way to derive prior distributions, they are necessarily subjective, in the sense that one radiologist may derive a different prior distribution than another and, hence, arrive at a different posterior distribution. Several points can be made regarding this controversy.

First, Bayesians can use diffuse, flat, or reference prior distributions that, for all practical purposes, consider all values in the feasible range as equally likely. Hence, if little prior information exists or if a Bayesian wishes to see what information the data themselves provide, this choice of prior distribution can be used. In fact, in many situations, a Bayesian analysis using reference priors will result in similar interval estimates as those provided by frequentist confidence intervals, but with a more natural interpretation: Unlike confidence intervals, Bayesian intervals (often called credible intervals) can be directly interpreted as containing the true parameter value with the indicated probability. Thus, no references to long runs of other trials are necessary to properly interpret a credible interval.

Second, although many frequentists have been quick to criticize Bayesian analysis because of the difficulty in deriving prior distributions, frequentist analysis formally ignores this information, which can hardly be considered as a better solution.

Third, if different clinicians have a range of prior opinions and hence a range of prior distributions, there will also be a range of posterior distributions. Presenting several Bayesian analyses matching this range of prior opinions helps to raise the level of debate after the publication of results in medical journals, because it accurately reflects the range of clinical opinion that exists in the community. Furthermore, it can be shown that as more data accumulate, the posterior distributions from different priors tend to converge toward a single distribution, accurately mirroring the process of eventual consensus among clinicians as data accumulate. When viewed in this light, prior distributions can be seen as a great advantage. See Spiegelhalter et al. [15] or a more introductory level article [9] for more information on using a range of prior distributions when carrying out a Bayesian analysis.

Having discussed the basic elements, let us see how Bayesian analysis works in practice by again considering our example of tumor diameters after radiation for brain cancer. We will discuss the three elements that lead to the posterior distribution calculated from Bayes' theorem, which are listed in the previous section.

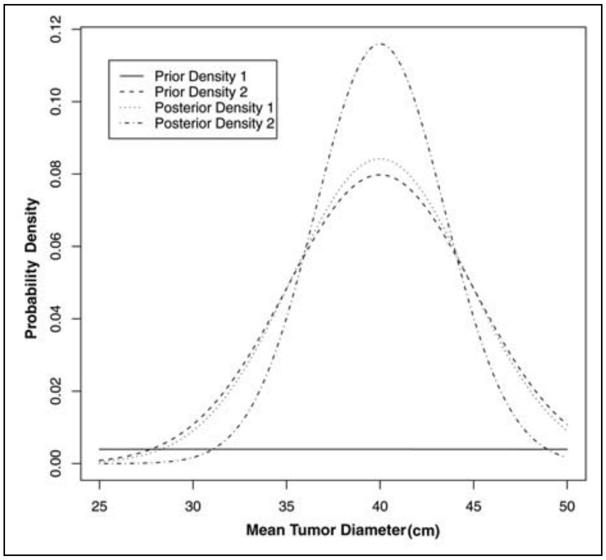
Recall that in our data set we had $\overline{x} = 3.0$, $\sigma = 1.5$, and $n = 10$, so that our likelihood function is a normal distribution with mean 3.5 and SE of 0.474, the same as was used in the frequentist inferences discussed previously. In general, the choice of prior distribution is based on any information that is available at the time of the experiment. We will consider two different prior distributions. The first (prior distribution 1 in Fig. 3) will be a normal distribution with a mean of 3.5 cm and a very large variance, say, 10,000. This is a noninformative prior, because all values in the likely range have an approximately equal chance of being the true value, the curve being quite flat over a wide range. Note that an equal 50% chance is given to both the null and alternative hypotheses that the new schedule is superior to that of the old, because the distribution is centered at 3.5 cm. The second prior distribution (prior distribution 2 in Fig. 3) will be centered at 3.0, with an SD of 0.5 (variance of 0.25). This would represent the opinion of a radiologist who is enthusiastic about the new schedule, with a prior opinion that the new mean tumor diameter will be between about 2.0 and 4.0 cm, with 95% probability (as calculated from the range of the normal [μ = 3.0, τ^2 = 0.25] distribution, where τ^2 is our prior variance). Do not be confused by the two distinct SDs that are used here: σ represents the variability of the tumor diameters among the patients, whereas τ represents how certain we are of our prior mean value.

We now wish to combine this prior density with the information in the data as represented by the likelihood function to derive the posterior distribution, using Bayes' theorem. After some algebra, the posterior distribution can be shown to be given by a normal distribution shown in equation 20,

$$N\left(A \times \mu + B \times \overline{x}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right) \qquad (20)$$

where $A = [(\sigma^2 / n) / (\tau^2 + \sigma^2 / n)]$ and $B = [(\tau^2) / (\tau^2 + \sigma^2 / n)]$. Note that the posterior mean value depends on both the prior mean, μ, and the observed mean in the data set, $\overline{x}$. Plugging these values into the previous equation and using the first (very flat) prior distribution, we find that the posterior distribution for our mean tumor diameter is $N (A \times \mu + B \times \overline{x} = 3.0, [(\tau^2\sigma^2) / (n\tau^2 + \sigma^2)] = 0.225)$. For the second more informative prior, the corresponding posterior distribution is N (3.0, 0.118).

Fig. 3.—Graph shows two prior and corresponding posterior densities for tumor diameter example.



The two prior and two posterior densities are displayed in Figure 3. Note that the second posterior distribution is narrower, because a stronger prior distribution was used. These posterior distributions can be used to derive 95% credible intervals and to test hypothesis from a Bayesian viewpoint. These calculations can be done using normal tables. Because these posterior distributions directly represent the probability distribution for our unknown parameter, interpretation of these quantities is straightforward.

For example, a 95% credible interval from posterior distribution 1 is given by (2.1–3.9). In comparing this interval to the prior 95% confidence interval calculated in the Statistical Inferences for Means section, we see that they are numerically identical (at least to one decimal place). However, the interpretations of these two intervals are different because the Bayesian credible interval is directly interpreted as the probability that the true mean tumor diameter lies in the given interval, given the data and the prior information used. This is in contrast to the less direct interpretation of a confidence interval, discussed earlier. Many people misinterpret confidence intervals as if they were Bayesian intervals. This error is often not too serious, because if little prior information is available, the two intervals are numerically similar. Therefore, even though it is technically incorrect, one does not go too far wrong thinking of confidence intervals as approximate Bayesian intervals, when there is little prior information. A 95% credible interval from our second posterior distribution is given by (2.3–3.7), which is somewhat narrower than the first interval.

We can also perform Bayesian hypothesis tests, again just using the posterior distributions. For example, suppose we wish to test H_0 ($\mu \geq 3.5$) versus H_A: ($\mu < 3.5$). We can calculate $\Pr\{H_0 \mid \text{data}\} = \Pr\{\mu \geq 3.5 \mid \text{data}\}$, which is equal to 14.5% for posterior 1 and 7.3% for posterior 2. Thus, we are approximately 85.5% or 92.7% sure that the tumor diameter under the accelerated schedule is better than the standard schedule, depending on which prior we use. Based on this, each clinician can make a decision about which schedule to apply to the next patient. Note again the very direct statements available for Bayesian hypothesis tests, compared with the nonintuitive interpretation of a p value. This clarity, however, comes at the expense of having to specify a prior distribution.

Carrying out Bayesian analyses is made easier via the use of freely available customized software. The posterior distributions shown earlier were performed using the First Bayes package [16], and more complex Bayesian analyses can be done via specialized Monte Carlo numeric routines implemented in WinBUGS software [17] made freely available by the Medical Research Council of Great Britain [18]. An excellent introductory text on Bayesian analysis is one written by Gelman et al. [19].

Conclusions

This module has introduced some of the major ideas behind statistical inference, with emphasis on the simple methods for continuous variables. Rather than a simple catalogue listing of which tests to use for which types of data, we have tried to explain the logic behind the common statistical procedures seen in the medical literature, the correct way to interpret the results, and what their advantages and drawbacks may be. We have also introduced Bayesian inference as a strong alternative to standard frequentist statistical methods, both for its ability to incorporate the available prior information into the analysis and for its ability to address questions of direct clinical interest.

The next few modules in this series will cover techniques suitable for other types of data, including proportions and regression methods.

References

1. Oshiro Y, Kusumoto M, Matsuno Y, et al. CT findings of surgically resected large cell neuroendocrine carcinoma of the lung in 38 patients. *AJR* 2004;182:87–91
2. Karlik SJ. Visualizing radiologic data. *AJR* 2003;180:607–619
3. Joseph L, Reinhold C. Fundamentals of clinical research for radiologists: introduction to probability theory and sampling distributions. *AJR* 2003; 180:917–923
4. Moore D, McCabe G. *Introduction to the practice of statistics,* 3rd ed. New York, NY: Freeman and Company, 1988
5. Armitage P, Berry G. *Statistical methods in medical research,* 3rd ed. Oxford, England: Blackwell Scientific Publications, 1994
6. Rosner B. *Fundamentals of biostatistics.* Belmont, MA: Duxbury, 1995
7. Rothman K. Writing for epidemiology. *Epidemiology* 1998;9:333–337
8. Evans S, Mills P, Dawson J. The end of the p-value. *Br Heart J* 1988;60:177–180
9. Brophy J, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *JAMA* 1995;273:871–875
10. Lilford R, Braunholz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;313:603–607
11. Sprent P. *Applied nonparametric statistical methods.* New York, NY: Chapman and Hall, 1989
12. Bristol D. Sample sizes for constructing confidence intervals and testing hypotheses. *Stat Med* 1989;8:803–811
13. Lemeshow S, Hosmer D, Klar J, Lwanga S. *Adequacy of sample size in health studies.* Chichester, England: Wiley, 1990
14. Bayes T. An essay towards solving a problem in the doctrine of chances: 1763. *Philos Trans R Soc* 1763;53:370–418
15. Spiegelhalter D, Freedman L, Parmar M. Bayesian approaches to randomized trials. *J R Stat Soc [Ser A]* 1994;157:387–416
16. O'Hagan A. First Bayes software. Available at: www.shef.ac.uk/st1ao/1b.html. Accessed December 25, 2003
17. Spiegelhalter D, Thomas A, Best N. *WinBUGS version 1.4 user manual.* Cambridge, England: MRC Biostatistics Unit, 2003
18. WinBUGS, version 1.4. Available at: www.mrc-bsu.cam.ac.uk/bugs/. Accessed December 25, 2003
19. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian data analysis*, 2nd ed. London, England: Chapman and Hall, 2003

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005

Fundamentals of Clinical Research for Radiologists

Lawrence Joseph[1,2]
Caroline Reinhold[3,4]

Statistical Inference for Proportions

This module will discuss the most commonly used statistical procedures when the parameters of interest arrive in the form of proportions. Understanding these methods is especially important to radiologists because so much radiologic research and clinical work involves dichotomous (e.g., yes or no, present or absent) outcomes summarized as proportions. For example, a given disease or condition may be present or absent in any given subject, and any time a diagnostic tool is used, test characteristics such as sensitivity, specificity, and positive and negative predictive values are all summarized as proportions.

We will continue to use the three basic methods for statistical inferences, including p values and confidence intervals (CIs) from a frequentist viewpoint, and posterior distributions leading to credible intervals from a Bayesian viewpoint. We will only briefly review the basic principles behind these generic inferential principles, so readers may wish to ensure they have a good understanding of the previous module [1] in this series before tackling this one. It may also be useful to recall the basic properties of the binomial distribution [2] because it is the central distribution used for inferences involving proportions.

We begin with inferences for single proportions, which are covered in the next section. Then we discuss inferences for two or more proportions from independent groups, inferences for dependent proportions, sample size determination for studies involving one or two proportions, and Bayesian methods for proportions. Finally, we will summarize what we have learned in this module.

Inferences for Single Proportions

Standard Frequentist Hypothesis Testing

Suppose a new computer-aided automated system for the detection of lung nodules on chest radiographs has been developed [3]. Suppose further that one wishes to investigate whether this new system provides improved sensitivity compared with standard detection via non-computer-aided methods of analyzing chest radiographs. In other words, suppose that chest radiographs are taken from a series of subjects who all truly have lung nodules, and we know that using standard (non-computer-aided) methods 90% of them will be found to have lung nodules and 10% of these cases will be missed. Is there evidence that the new computer-aided automated system provides increased sensitivity compared with the standard method of detection?

To look for evidence of improved sensitivity in the new automated system, we might wish to test the null hypothesis (H_0) that the automated system is in fact not better than standard detection, versus an alternative hypothesis (H_A) that it is better. Formally, we can state these hypotheses as:

$$H_0: p \leq 0.9$$
$$H_A: p > 0.9$$

where p represents the unknown true probability of success of the new automated system in detecting lung nodules.

Suppose that we observe the results from 10 subjects with lung nodules, and all 10 test positively with the new automated system. Recalling the correct definition of a p value [1] (it is the probability of obtaining a result as extreme as or more extreme than the result observed, given that the null hypothesis is exactly correct), how would we calculate the p value in this case? For our example of the new automated technique, the definition implies that we need to calculate the probability of obtaining 10 (or more, but in this case more than 10 is impossible) successful

Received November 5, 2004; accepted after revision November 10, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 16th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, 1650 Cedar Ave., Montreal, QC H3G 1A4, Canada.

[2]Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave. W, Montreal, QC H3A 1A2, Canada. Address correspondence to L. Joseph (Lawrence.Joseph@mcgill.ca).

[3]Department of Diagnostic Radiology, Montreal General Hospital, McGill University Health Centre, 1650 Cedar Ave., Montreal, QC H3G 1A4, Canada.

[4]Synarc Inc., 575 Market St., San Francisco, CA 94105.

AJR 2005;184:1057–1064

0361–803X/05/1844–1057

lung nodule detections in the 10 patients to whom the technique was applied, given that the true rate of success is exactly 90%. Recall [2] that if x follows a binomial distribution with probability of success p, then Pr (x successes in n trials) = $[n!/(x!(n–x)!)]p^x(1 – p)^{n–x}$, where $x!$ is read as "x factorial" and is equal to $x(x – 1)$ $(x – 2)$… (2) (1). For example, 5! = (5) (4) (3) (2) (1) = 120, and by convention 0! = 1. Using this binomial probability function, we can calculate the probability of 10 successes in a row with p = probability of success = 0.9 as shown in equation 1:

$$\frac{10!}{10!0!}0.9^{10}(1-0.9)^0 = 0.9^{10} = 0.3487 \qquad (1)$$

So there is about a 34.9% chance of obtaining results as extreme as or more extreme than the 10 of 10 results observed, if the true rate for the new technique is exactly 90%. Therefore, the observed result is not unusual, and hence compatible with the null hypothesis, so we cannot reject H_0.

This calculation could be done exactly, because the sample size was quite small. For larger sample sizes, the normal approximation to the binomial distribution [2] could be used. Also, this test was one-sided, but two-sided hypotheses are also of interest. For example, suppose we wish to test a similar null hypothesis as above (H_0: p = 0.9) but against a two-sided alternative (H_A: $p \neq 0.9$). Suppose we observed 98 successes in 100 trials. Because our test is two-sided, according to the definition of a p value we need to calculate the probability of obtaining data as extreme as or more extreme than the observed 98 of 100. Now, 98 is 8 higher than the 90 expected under the null hypothesis, so that to be as extreme as or more extreme than the 98 observed, we need to be 8 or more above or below the expected 90. That is, we need to calculate the probability of 98, 99, or 100 successes on one side, and 82, 81, 80, …, 2, 1, 0 on the other side. This lengthy calculation, involving the sum of 85 binomial calculations, can be well approximated by using the normal approximation to the binomial distribution [2]. Let our estimate of the unknown proportion be $\hat{p}$ = 98 / 100 = 0.98. We can calculate equations 2–4:

$$\begin{aligned} z &= \left|\frac{\hat{p}-0.9}{\sqrt{\frac{p(1-p)}{n}}}\right| && (2)\\ &= \frac{.98-0.9}{\sqrt{.9(1-.9)/100}} && (3)\\ &= 2.67 && (4) \end{aligned}$$

Looking up 2.67 on normal tables, we find 0.004, and doubling this value gives us our two-sided p value, which is 0.008. It is unlikely that rates of 98% or more extreme will be observed in 100 trials if the true rate is in fact only 90%. Therefore, in this case, sufficient evidence exists to reject the null hypothesis in favor of the alternative.

Although p values are still often found in the literature, several major problems are associated with their use, as we have previously discussed [1]. Briefly, the null hypothesis is virtually never exactly true (is it possible that the true underlying sensitivity is exactly 90%, as opposed to, say, 89.9999% or 90.0001%?), so we know it should be rejected regardless of the data we observe. Furthermore, the p value says nothing about the effect size, which is crucial to clinical decision making, with large sizes usually implying a more clinically important effect than small sizes. A much more interesting question is to estimate the rate or proportion of interest, together with a measure of the accuracy of the estimate. CIs are one answer to this question, and we discuss them next. The Bayesian solution—credible intervals—is discussed later.

Confidence Intervals for Single Proportions

Continuing the previous example, we have observed rates of 100% (10/10 in our smaller sample) or 98% (98/100 in our larger sample), but we know that these are estimates only, not guaranteed (in fact, unlikely) to exactly equal the true rates. On the basis of these data, however, what can we say about what we would expect the true rate to be?

One way to answer this question is with a CI. CIs usually have the form

$$\text{estimate} \pm k \times \text{standard error}$$

where the estimate and SE are calculated from the data, and where k is a constant dependent on the width of the CI desired. The value of k is usually near 2 (e.g., k is 1.96 for a 95% CI).

If one observes x = 98 positive tests in n = 100 subjects known to have lung nodules, a point estimate of the success rate is $\hat{p}$ = x / n = 0.98 or 98%. We use the notation $\hat{p}$ rather than p to indicate that this is an estimated rate, not necessarily equal to the true rate, which we denote by p. Following this generic formula, a CI for a binomial probability of success parameter is given by the formula in equation 5,

$$\left(\hat{p} - z \times \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}},\ \hat{p} + z \times \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}}\right) \qquad (5)$$

where z is derived from normal tables, and is given by z = 1.96 for the usual 95% CI (z = 1.64 for a 90% CI and z = 2.56 for a 99% CI). Therefore, the 95% CI in our example is calculated as shown in equation 6,

$$\left(0.98 - 1.96 \times \sqrt{\frac{0.98\times 0.02}{100}}, 0.98 + 1.96 \times \sqrt{\frac{0.98\times 0.02}{100}}\right) (6)$$

which here gives (0.930–0.994).

Technical note: This formula uses the normal approximation to the binomial distribution [2]. Exact formulae are also available [4], which are especially useful for small sample sizes or for estimates $\hat{p}$ near 0 or 1. For example, using an exact approach to this CI yields (0.930–0.998), which is very close to but not identical to that given by the indicated normal approximated interval. In addition, when $\hat{p}$ equals 0 or 1 exactly, the normal approximation breaks down, because the variance is estimated to be 0. Here one has no choice but to use a different procedure. The exact method yields a wider 95% CI of (0.741–1.000) in the case of our smaller data set, where 10 positive values were found in 10 subjects. There is also an easy-to-use and reasonably accurate rule of thumb when calculating a binomial CI and one observes 0 events. The rule is this: If you observe n patients, and none of these patients have an event, then a 95% CI for the probability of the event goes from 0 to 3 / n. For example, if you observe 0 events in 10 binomial trials, then an approximate 95% CI would go from 0 to 3 / 10 = 0.3. By symmetry, the rule would say that if you observe only events in n trials, then the 95% CI would go from (1 – 3 / n) to 1. For example, if you observe 10 events in 10 trials, then the 95% CI would go from 0.7 to 1, which is reasonably close to the exact solution of (0.741–1.000) given here.

How does one interpret this CI? Recall from the previous module [1] that the 95% confidence value (often called the confidence coefficient) is a long-run probability over repeated uses of the CI procedure. In practice, there are five different interpretations associated with CIs, depending on where the upper and lower CI limits fall with respect to clinical cut points of interest (see Fig. 2 of Joseph and Reinhold [1]). The formula displayed in equation 5 of this article provides a procedure that, when used repeatedly across different problems, will capture the true value of p 95% of the time and fail to capture the true value 5% of the time. In this sense, we have confidence that the procedure works well in the long run, although in any single application, of course, the interval either does or does not contain the true proportion p.

For our smaller data set, with 10 subjects found to be positive in 10 trials, the 95% CI ranges from 74.1% to 100%, providing a large and inconclusive interval, because it may well be better or worse than the standard diagnosis, which is assumed to be successful 90% of the time. In our larger data set, the 95% CI ranged from 93.0% to 99.4%, so we can be quite certain that it is better than standard diagnoses. However, it can be as little as 3% better (90% compared with the lower CI

TABLE 1 Data from a Two-Group Study

Diagostic Method	Test Positive	Test Negative	Total
Automated system	285	15	300
Standard diagnosis	265	45	310
Total	550	60	610

TABLE 2 Expected Data for the Example in Table 1 Under the Null Hypothesis

Diagnostic Method	Test Positive	Test Negative	Total
Automated system	270.49	29.51	300
Standard diagnosis	279.51	30.49	310
Total	550	60	610

limit of 93%). Whether this is enough evidence to switch to the new automated system or not depends on clinical judgment. This in turn depends on many factors, including the cost and availability of the new automated system and the average clinical benefits that will accrue to those diagnosed earlier by the more sensitive diagnostic method.

Inferences for Two or More Independent Proportions

Let us continue with our example comparing the diagnostic properties of a new automated system for the detection of lung nodules on chest radiographs compared with standard detection via non-computer-aided methods. Earlier we assumed that the rate in the standard diagnosis group was exactly known before the study, but this is somewhat unrealistic. We will now relax this assumption, and consider the data from the two-group study shown in Table 1 (presented in the form of a 2 × 2 table of data because we have two possible outcomes in each of the two groups being compared).

Again, we assume that all 610 subjects studied are truly positive, so that one would like to draw inferences about whether the automated system has increased sensitivity compared with the usual diagnosis group. Although one observes $\hat{p}_1 = [285 / 300] = 0.95$ sensitivity for the automated system compared with $\hat{p}_2 = [265 / 310] = 0.855$ sensitivity using standard diagnosis, for a 9.5% observed difference, a CI will provide us with a range of values compatible with the data that will help draw a better conclusion than simply looking at the observed point estimates. To calculate a CI for this difference in proportions, we can use the formula in equation 7,

$$\left(\hat{p}_1 - \hat{p}_2 - z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\right.$$
$$\left.\hat{p}_1 - \hat{p}_2 + z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right) \tag{7}$$

which extends equation 5 to the case of two proportions. In this formula, $\hat{p}_1$ and $\hat{p}_2$ are the observed proportions in the two groups out of sample sizes n_1 and n_2, respectively, and z is the relevant percentile from normal tables, chosen according to the desired level of the CI. For example, for a 95% CI $z = 1.96$, for a 90% interval $z = 1.64$, and so on. Using this formula for the diagnosis data given, one finds that a 95% CI for the difference in sensitivity is (0.049–0.141). This interval suggests that the automated system is indeed better, likely by at least as much as 0.049. Unless cost is a prohibitive factor, from these data it looks like the automated system is worthwhile (at least in these hypothetical data).

Although CIs are preferred for reasons we have briefly discussed here and which were more extensively discussed in a previous module in this series [1], we will also discuss hypothesis testing for proportions, because one often sees such tests in the literature. Suppose we wish to test the null hypothesis that $p_1 = p_2$—that is, the null hypothesis states that the success rates are identical in the two units. Because we hypothesize $p_1 = p_2$, we expect to observe, on average, the data in Table 2.

Why do we expect to observe this table of data if the null hypothesis is true? We have observed a total of 550 "successes" divided among the two groups. If $p_1 = p_2$ and if the sample sizes were equal in the two groups, we would have expected (550 / 2) = 275 successes in each group. However, because the sample sizes are not equal, we expect 550 × (300 / 610) = 270.49 to go to the automated system group, and 550 × (310 / 610) = 279.51 to go the standard diagnosis group. Similarly, expected values for the 60 negatively testing patients can be calculated. Observed discrepancies from these expected values are evidence against the null hypothesis. To perform a chi-square test, we now calculate as shown in equations 8–10:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{8}$$

$$= \frac{(285 - 270.49)^2}{270.49} + \frac{(15 - 29.51)^2}{29.51} + \frac{(265 - 279.51)^2}{279.51} + \frac{(45 - 30.49)^2}{30.49} \tag{9}$$

$$= 15.57. \tag{10}$$

Comparing the $\chi^2 = 15.57$ value on chi-square tables with 1 degree of freedom (*df*) (see Armitage and Berry [4] or almost any basic textbook on statistics to find such tables), we find that $p \approx 0.0001$ so that we have strong evidence to reject the null hypothesis. This coincides with our conclusion from the CI, but note that the CI is more informative than simply looking at the *p* value from the chi-square test, because a range for the difference in sensitivities is provided by the CI. Thus, the clinical importance of any differences can be more easily evaluated.

The chi-square test can be extended to include tables larger than the so-called 2 × 2 table of this example. For instance, a 3 × 2 table could arise if, rather than classifying patients as positive or negative, we included a third outcome category, such as "chest radiograph is inconclusive." A 3 × 2 table could also arise if we considered comparing a third method of diagnosis rather than the two considered here. In these cases we would sum over 3 × 2 = 6 terms rather than the four terms of a 2 × 2 table. Although for 2 × 2 tables the *df* is always equal to 1, in general the *df* for chi-square tests is given by $(r - 1) \times (c - 1)$, where the number of rows in the table is r and the number of columns is c. Thus, in the case of a 3 × 2 table, we would have (3 – 1) × (2 –1) = 2 *df*. In general, cases with arbitrary numbers of rows and columns can be constructed and analyzed using the chi-square test.

In order for the chi-square test to be valid, one needs to ensure that the expected value for each cell in the table is at least 5. This was satisfied in the previous example, in which our smallest expected table value was 29.51, much larger than 5. Fisher's exact test [4] is often used if this criterion is not satisfied for a particular table. The Fisher's exact test is valid for tables of any size, in particular for small sample sizes.

TABLE 3 Generic Setup of a 2 × 2 Table

Second Test	First Test		Total
	Positive	Negative	
Positive	*a*	*b*	*a* + *b*
Negative	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*N* = *a* + *b* + *c* + *d*

Inferences for Dependent Proportions

A two-group clinical trial, where n_1 subjects receive treatment A and n_2 different subjects receive treatment B, usually results in independent samples. That is, the results under treatment regimen A (number of successful outcomes among the n_1 subjects given treatment A) do not depend on the outcomes in group B (number of successful outcomes among the n_2 subjects given treatment B).

Sometimes, however, subjects or data points may come in pairs, so that dependencies among the groups are naturally induced. Consider, for example, the frequently occurring situation in which two diagnostic tests are given to each of a series of subjects. Each subject may test positively or negatively on each of the two tests, so that the data arising from such a study may be summarized in a 2 × 2 table, as seen in Table 3.

Thus, we observe *a* number of subjects who are positive on both tests, *b* subjects who are negative on the first test but positive on the second test, *c* subjects who are positive on the first test but negative on the second, and *d* subjects who test negatively on both tests. The cells with *a* and *d* contain concordant pairs, because the two test results agree with each other, whereas the cells with *b* and *c* contain discordant pairs.

Similar data can arise from a matched case–control study. In this type of study design, cases (e.g., those with a particular disease) are first found and then matched to a particular control case with similar characteristics but without the condition of interest.

As a concrete example, suppose we wish to investigate whether impaired renal function is related to diminished renal size. Because we would otherwise require large numbers of subjects to be followed up over a long period of time, a case–control design may be considered. Thus, one finds patients with impaired renal function and control subjects without impaired renal function, and discovers whether there is a tendency of those with impaired renal function to show diminished renal size on sonography compared with those without impaired renal function. Of course, patients with impaired renal function may tend to be different from subjects without (control subjects) in many ways, so to minimize possible confounding one may want to control for age, sex, height, hypertension, diabetes, and so on. For each patient, one may want to find a control subject with similar age, sex, height, and other characteristics, thus forming a series of matched pairs. Within each of these pairs, one then classifies each patient and control subject into whether they have diminished renal size at sonography or not.

Within each matched pair are four possibilities: Both the patients and control subjects show diminished renal size, or both may not show diminished renal size. These two possibilities form concordant pairs (introduced in previous text) because similar renal size is shown for each subject forming the pair. Of course, the other two possibilities are that the patient shows diminished renal size and the control does not, and vice versa, forming the nonconcordant pairs. As was the case with diagnostic test studies, the data may be formed into a 2 × 2 table, as shown in Table 4.

Note that there are a total of *N* pairs of subjects in this study, meaning that we in fact have 2*N* individuals (similarly, in the diagnostic test case, we have 2*N* tests, but only *N* subjects). We have *a* subjects in whom both the patient and the matched control subject showed diminished renal size, *b* subjects in whom the control but not the patient showed diminished renal size, and so on.

Suppose we would like to test the null hypothesis that diminished renal size is unrelated to impaired renal function versus the alternative hypothesis that a relation exists between diminished renal size and impaired renal function. The McNemar test focuses on the discordant pairs, represented in Table 4 by *b* and *c*. We can formulate the statistic shown in equation 11,

$$X^2 = \frac{(|b-c|-1)^2}{b+c} \qquad (11)$$

which approximately follows a chi-square distribution with 1 *df.* Thus, a *p* value can be calculated for this test.

For example, suppose we observe the following data: $a = 200$, $b = 100$, $c = 75$, and $d = 300$. According to the McNemar test, we calculate as shown in equation 12:

$$X^2 = \frac{(|100-75|-1)^2}{175} = 3.29 \qquad (12)$$

Looking up 3.29 on chi-square tables yields a *p* value of 0.069, so that it is close to but does not cross the (admittedly arbitrary) threshold of 0.05. Thus, at least at the type 1 error level of 0.05, we do not have evidence to reject the null hypothesis.

Of course, the McNemar test can also be used for testing hypotheses relating to diagnostic test data of the type described at the beginning of this section.

The general criticisms relating to hypothesis testing and *p* values carry over the particular case of testing dependent proportions through the McNemar test. Odds ratios and associated CIs can be calculated from matched pair studies, and these will be covered in a future module in this series.

Sample Size Determination for One and Two Proportions

As previously discussed [1], there has been a strong trend away from hypothesis testing and *p* values toward the use of CIs in the reporting of results from biomedical research. Because the design phase of a study should synchronize with the analysis that will be eventually performed, sample size calculations should be performed on the basis of ensuring adequate numbers for

TABLE 4 Data in a Case Control Study

Diminished Renal Size	Diminished Renal Size		
	Patient Has	Patient Does Not Have	Total
Control has	*a*	*b*	*a* + *b*
Control does not have	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*N* = *a* + *b* + *c* + *d*

accurate estimation of important quantities that will be estimated in the study, rather than by power calculations. For one- and two-sample problems, the formulae are as given in the following paragraphs.

Single Sample

Let p be the proportion that is to be estimated, and assume that we wish to estimate p to an accuracy of a total CI width of $w = 2 \times h$, where h is half the total CI width.

Then we can perform the calculation shown in equation 13,

$$n = \frac{z^2}{h^2}p(1-p) = \frac{4z^2}{w^2}p(1-p), \qquad (13)$$

where, again, z is the appropriate normal quantile (e.g., $z = 1.96$ for a 95% CI).

Two Sample

Let p_1 and p_2 be the two proportions whose difference we would like to estimate to a total CI width of $w = 2 \times h$.

Then we can perform the calculation shown in equation 14,

$$n = \frac{4 \times (p_1 \times (1-p_1) + p_2 \times (1-p_2)) \times z^2}{w^2} = \frac{(p_1 \times (1-p_1) + p_2 \times (1-p_2)) \times z^2}{h^2} \qquad (14)$$

where n represents the required sample size for each group.

As an example, suppose we want to design a study to measure the difference in diagnostic accuracy for two types of imaging techniques, say MRI versus CT for staging cervical carcinoma. Suppose that CT is thought to be successful in staging patients with cervical carcinoma, with probability $p_1 = 0.70$, and MRI may improve this to $p_2 = 0.80$. We would like to estimate the true difference to within $h = 0.05$, so that not only will we be able to detect any differences of 10%, but the 95% CI will be far enough away from 0 (if our predicted rates are correct) so that we can make a more definitive conclusion as to the clinical usefulness of MRI. We calculate as shown in equations 15 and 16

$$n = \frac{(p_1 \times (1-p_1) + p_2 \times (1-p_2)) \times z^2}{h^2} \qquad (15)$$

$$= \frac{(0.7 \times (1-0.7) + 0.8 \times (1-0.8)) \times 1.96^2}{0.05^2} = 569 \qquad (16)$$

so that 569 patients are required in each group.

The main practical difficulty with equations 13 and 14 is assigning appropriate values for p, p_1, and p_2. It is therefore useful to note that equation 13 is maximized when $p = 0.5$, so using this value is conservative in the sense that the desired CI width will be respected regardless of the estimated value of p that will be observed in the study. This conservative value, however, may provide too large a sample size and therefore be wasteful of resources if the true proportion is far from 0.5. A conservative rule of thumb is to use the value of p that is closest to 0.5, selected from the set of all plausible values. Similarly, equation 14 is maximized for $p_1 = p_2 = 0.5$, so a similar rule of thumb applies for each of p_1 and p_2.

Bayesian Inference for Proportions

Consider again the problem introduced in the section called Inferences for Single Proportions. Recall that in that example the sensitivity of standard interpretation of radiographs is assumed to be 90%, whereas the small data set collected so far for the new automated radiograph interpretation system indicates a 100% success rate but is based on only 10 subjects. The frequentist CI was very wide, ranging from 74.1% to 100%. Therefore, the data themselves have not been particularly helpful in making a decision as to which technique to use for the next patient, because values indicating a new test that is both more and less sensitive than the standard diagnostic method have not been ruled out by the CI. At this point, with the data being relatively uninformative, the radiologist may decide to be conservative and remain with the standard method until more information becomes available about the new automated technique, or may go with his or her "gut feeling" as to the likelihood that the new therapy is truly better or not better. If there have been data from animal experiments or strong theoretic reasons why the new technique may be better, the radiologist may be tempted to try the new one. Can anything be done to aid in this decision-making process?

Bayesian analysis has several advantages over standard or frequentist statistical analyses. These advantages include the following:

First is the ability to address questions of direct clinical interest, such as direct probability statements about hypotheses of interest and credible intervals with similarly easy interpretations [1]. Hence, results of Bayesian analyses are straightforward to interpret, in contrast to the obscure and difficult-to-understand (and frequently misinterpreted) inferences provided by p values and CIs [1].

Second is the ability to incorporate relevant information not directly contained in the data into any statistical analysis. This enters in the form of prior information about parameters of interest.

The third advantage is that Bayesian analysis is a natural way to update statistical analyses as new information becomes available.

A main theoretic difference between frequentist and Bayesian statistical analyses is that Bayesian analysis permits parameters of interest (binomial probabilities, population means, and so on) to be considered as random quantities, so that probabilities can be attached to the possible values that they may attain. On the other hand, frequentists consider these parameters to be fixed (albeit possibly unknown) constants, so they have no choice but to attach their probabilities to the data that could arise from the experiment, rather than to the parameters. This distinction is the main reason Bayesian analysis can answer direct questions of interest, whereas frequentist analyses must settle for answering more obscure questions in the form of p values and CIs.

The ability to address questions of direct interest, however, comes at the cost of having to do a bit more work. Not only do Bayesians have to collect data from their experiments, but they also have to quantify the state of knowledge of all parameters before their collecting this data. This nontrivial step is summarized in a prior distribution. The information in the prior distribution is updated by the information in the data to arrive at a posterior distribution, which summarizes all available information, past and current. We will apply a Bayesian analysis to our radiologist's decision later in this section, but first we need to recall the basic elements of all Bayesian analyses and see how they are applied to drawing inferences about our parameter of interest here, the binomial success rate of the new automated radiographic technique.

Let us generically denote our parameter of interest as θ. Hence, θ can be a binomial parameter, a set of two independent or dependent binomial parameters, or the mean and variance from a normal distribution, or an odds ratio, or a set of regression coefficients, and so on. Note in particular that θ can be two- or more dimensional. The parameter of interest is sometimes usefully thought of as the "true state of nature." As discussed in more detail in the previous module in this series [1], the basic elements of a Bayesian analysis then are as follows:

First is the prior probability distribution, $f(\theta)$. This subjective prior distribution summarizes what is known about θ before the experiment is performed.

Second is the likelihood function, $f(x \mid \theta)$. The likelihood function provides the distribution of the data, x, given the parameter value θ.

For instance, for proportions it may be a binomial likelihood, as in equation 17:

$$l(x|p) = Pr\{x \text{ successes in } N \text{ trials}\} = \frac{N!}{(N-x)!\,x!}p^x(1-p)^{(N-x)} \tag{17}$$

Third is the posterior distribution, $f(\theta \mid x)$. The posterior distribution summarizes the information in the data, x, together with the information in the prior distribution, $f(\theta)$. Thus, it summarizes what is known about the parameter of interest θ after the data are collected.

Bayes' theorem relates the above three quantities:

posterior distribution =
[likelihood of the data × prior distribution] /
a normalizing constant,

or using our notation and omitting the normalizing constant, as shown in equation 18,

$$f(\theta|x) \propto f(x|\theta) \times f(\theta) \tag{18}$$

where ∝ indicates "is proportional to."

Thus, we update the prior distribution to a posterior distribution after seeing the data via Bayes' theorem. The current posterior distribution can be used as a prior distribution for the next study; hence, Bayesian inference provides a natural way to represent the learning that occurs as science progresses.

The prior distribution is subjective and chosen by each investigator according to his or her appreciation of the past literature regarding the unknown parameters of interest. Hence, the prior distribution is not unique to each experiment but can vary from investigator to investigator. This can be seen as accurately reflecting clinical reality. Different clinicians can have different initial opinions about a parameter value, although these opinions tend to concentrate about a constantly narrowing range of values as more data accumulate. This is how Bayes' theorem operates, because the prior becomes a less important contributor to the posterior distribution as more data become available. See the previous module for more discussion about prior distributions [1].

We now will apply the general Bayesian technique we have described to the specific problem of inferences for binomial proportions.

Suppose that in a given experiment x "successes" are observed in N binomial trials. Let $\theta = p$ denote the parameter of interest—the true but unknown probability of success—and suppose that the problem is to find an interval that covers the most likely locations for p given the data.

The Bayesian solution to this problem follows the usual pattern, as outlined previously. Hence, the main steps can be summarized as first, write down the likelihood function for the data. Second, write down the prior distribution

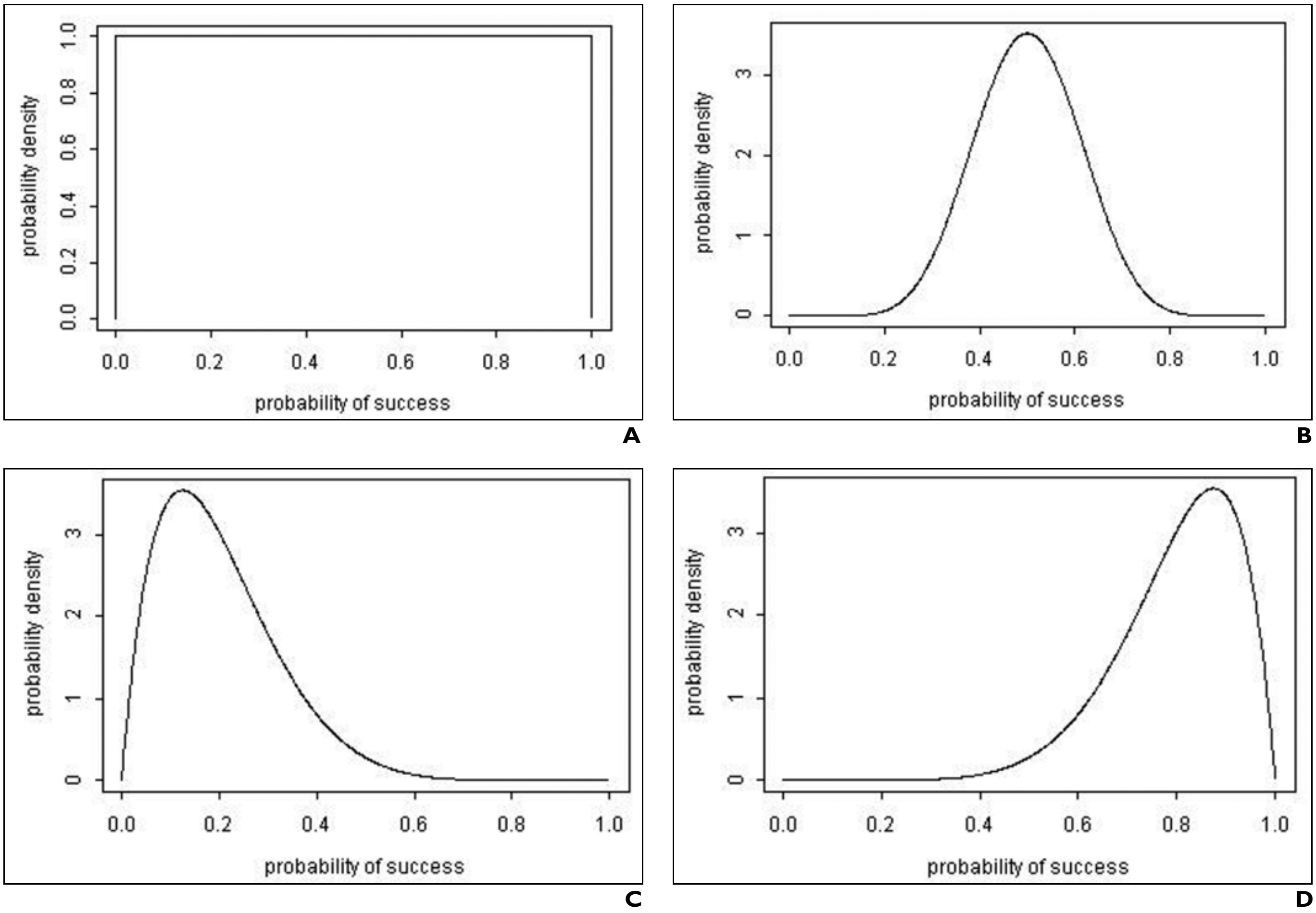


Fig. 1.—Series of four beta densities.
A–D, Graphs show beta(1,1) (**A**), beta(10,10) (**B**), beta(2,8) (**C**), and beta(8,2) (**D**) densities. Beta(1,1) distribution (**A**) is also known as the uniform density.

for the unknown parameter p. Third, use Bayes' theorem (i.e., multiply the equation for the likelihood function of the data by the prior distribution) to derive the posterior distribution. Use this posterior distribution, or summaries of it like 95% credible intervals, for statistical inferences. Credible intervals are the Bayesian analogues to frequentist CIs.

For the case of a single binomial parameter, these steps are realized in this manner:

Step 1

The likelihood function is the usual binomial probability formula shown in equation 17, where $l(x \mid p)$ represents the likelihood function for the success rate p given data x.

Step 2

Although any prior distribution can be used, two distributions are of particular interest. The first prior distribution we will discuss is the uniform prior distribution, which specifies that all possible values (for proportions, this implies all values in the range of 0–0) are equally probable, a priori. See Figure 1A. The uniform distribution is suitable for use as a "diffuse" or a "noninformative" distribution, when little or no prior information is available or when one wishes to see the information contained in the data by itself.

A second particularly convenient prior distribution, for reasons to be explained, is the beta distribution. A random variable, θ, has a distribution that belongs to the beta family if it has a probability density given by equation 19

$$f(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (19)$$

for $0 \le \theta \le 1$, and $\alpha, \beta > 0$. $B(\alpha,\beta)$ represents the beta function evaluated at (α,β). It is simply the normalizing constant that is necessary to make the total area under the curve equal to 1, but otherwise plays no role.

Some beta distributions are illustrated in Figure 1. For example, using a beta($\alpha = 1, \beta = 1$) distribution reproduces the perfectly flat or uniform distribution discussed previously. Thus, the uniform distribution is really just a special case of the beta distribution. On the other hand, a beta($\alpha = 10, \beta = 10$) density produces a curve similar in shape to a normal density centered at $\theta = 0.5$. If $\alpha > \beta$ the curve is skewed toward values near 1, whereas if $\alpha < \beta$ the curve is skewed toward values near 0.

The mean of the beta distribution is given by equation 20,

$$\mu = \frac{\alpha}{\alpha+\beta}, \quad (20)$$

and the SD is given by equation 21.

$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \quad (21)$$

To choose a prior distribution, one needs only to specify values for α and β. This can be done by finding the α and β values that give the correct prior mean and SD values. Solving these two equations in two unknowns, the formulae are shown in equations 22 and 23.

$$\alpha = -\frac{\mu\,(\sigma^2+\mu^2-\mu)}{\sigma^2} \quad (22)$$

$$\beta = \frac{(\mu-1)\,(\sigma^2+\mu^2-\mu)}{\sigma^2} \quad (23)$$

For example, if we wish to find a member of the beta family centered near $\mu = 0.9$ and with $\sigma = 0.05$, then plugging these values for μ and σ into these two equations gives $\alpha = 31.5$ and $\beta = 3.5$, so that a beta(31.5, 3.5) will have the desired properties. This curve, pictured in Figure 2, may be an appropriate prior distribution for the problem introduced at the beginning of this section if the radiologist believes, a priori, that the new technique is likely to be successful between 80% and 100% of the time, and whose best guess of the rate is 90%. Note that this clinician has centered the prior around the rate thought to be equal to the standard treatment. Thus, this prior distribution would give equal a priori weight to both the null and alternative hypotheses given at the start of the section on Inferences for Single Proportions. We will return to this example again shortly.

Step 3

As always, Bayes' theorem says

posterior distribution ∝ prior distribution ×
likelihood function.

In this case, it can be shown (by relatively simple algebra) that if the prior distribution is beta(α,β) and the data are x successes in N trials, then the posterior distribution is again a beta distribution, beta($\alpha + x$, $\beta + N -$ x). This simplicity arises from noticing that both the beta prior distribution as represented in equation 19 and the binomial likelihood as given in equation 17 have the general form $p^a \times (1-p)^b$, so that when multiplying them as required by Bayes' theorem, the exponents simply add, and the form is once again recognized to be from the beta family of distributions.

Hence, if we observe the new automated computer-aided radiologic method to correctly identify 10 patients in a row with lung nodules, and if we use the prior distribution discussed previously, then the posterior distribution is a beta(31.5 + 10, 3.5 + 0) = beta(41.5, 3.5) distribution, which is illustrated in Figure 2. The mean of this distribution is [41.5 / (41.5 + 3.5)] = 0.922, and the 95% posterior credible interval is (0.844–0.988). The

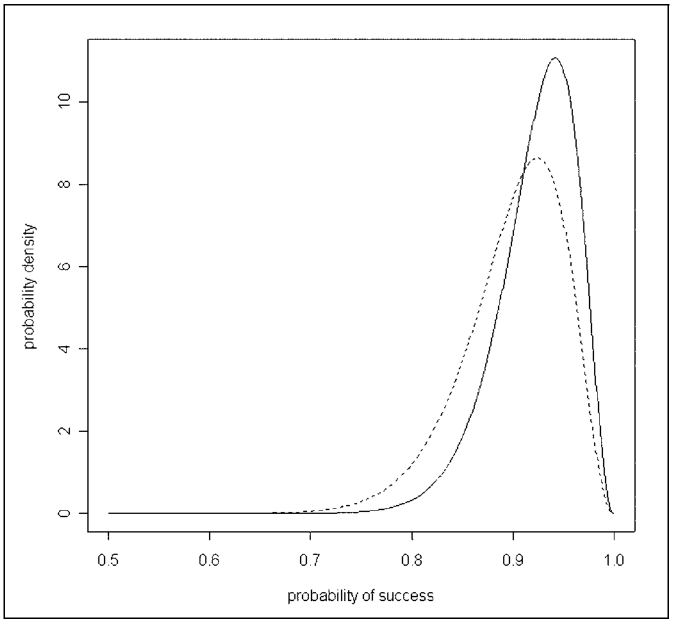


Fig. 2.—Prior (*dotted line*) and posterior (*solid line*) beta densities for automated radiology example.

probability of being greater than 90% is 0.748 (area under the curve to the right of 0.9 in Fig. 2). Therefore, the radiologist may or may not be tempted to try the automated technique on the next patient but should realize that this decision is mostly based on the prior information, to which the data contributed only a small amount of new information. Looking at Figure 2, we see that the prior density was shifted only a small amount by the data. If instead the radiologist "lets the data speak for themselves" by using a beta(1,1) or uniform prior distribution (Fig. 1), then the 95% interval is (0.773–0.971), very similar numerically to the frequentist CI of the section Inferences for Single Proportions, although their interpretations are quite different. Bayesian intervals (deliberately called credible intervals to distinguish them from frequentist confidence intervals) are interpreted directly as the posterior probability that p is in the interval, given the data and the prior distribution. No references to long-run frequencies or other experiments are required, as is the case for CIs.

In general, one should usually perform a Bayesian analysis using a diffuse prior distribution like a beta(1,1) distribution, to examine what information the current data set provides. Then one or more Bayesian analyses with more informative prior distributions could be performed, depending on the available prior information. If opinions in the medical community are widely divergent concerning the parameters of interest, then several prior distributions should be used. If the data set is large, then similar conclusions will be reached no matter which prior distribution one starts with. On the other hand, with smaller data sets, diversity of opinions will still exist, even after the new data are analyzed. Bayesian analysis allows this situation to be accurately represented and assessed.

Although we discuss only the simple case of Bayesian inference for a single binomial proportion, these methods are easily extended to the case of two or more proportions. For a clinical example using Bayesian analysis to compare two proportions, see Brophy and Joseph [5]. This example also illustrates the use of a range of prior distributions and shows that Bayesian analysis can often come up with answers that are quite different from those obtained using a frequentist approach.

Discussion

This module has introduced some of the major ideas behind statistical inference for proportions, with emphasis on the simple methods for one and two samples. Rather than a simple catalogue listing of which methods to use for which types of dichotomous data, we have tried to explain the logic behind the common statistical procedures seen for binary data in the medical literature, the correct way to interpret the results, and what their advantages and drawbacks may be. We have also introduced Bayesian inference as a strong alternative to standard frequentist statistical methods, for both its ability to incorporate the available prior information into the analysis and its ability to address questions of direct clinical interest.

For more information about inferences on proportions, see the books by Fleiss [6] for the frequentist perspective and by Gelman et al. [7] for the Bayesian view. General books on statistical inferences in medicine [8–10] all contain many techniques on inferences for proportions that are beyond the scope of this module.

Software is available that makes carrying out all the analyses discussed in this module relatively easy. From the frequentist viewpoint, there are literally dozens of statistical packages available for purchase, but much excellent free software is also available. For example, the *R* package [11] is freely available for most computer platforms, including Windows (Microsoft) and Linux PCs and MacOS (Apple). It is a comprehensive package that is constantly being updated. Free Bayesian software includes First Bayes [12] for simple problems and WinBUGS [13, 14] for more complicated problems.

The previous module covered similar techniques to those covered here for continuous data, and future modules in this series will cover techniques suitable for other types of study designs and questions that arise in radiology, including linear and logistic regression methods. The latter is especially relevant because logistic regression allows one to analyze dichotomous outcomes from one or more groups while adjusting the analysis for potential confounding factors.

References

1. Joseph L, Reinhold C. Fundamentals of clinical research for radiologists. Statistical inferences for continuous variables. *AJR* 2005;184:1047–1056
2. Joseph L, Reinhold C. Fundamentals of clinical research for radiologists. Introduction to probability theory and sampling distributions. *AJR* 2003;180:917–923
3. Kakeda S, Moriya J, Sato H, et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. *AJR* 2004;182:505–510
4. Armitage P, Berry G. *Statistical methods in medical research*, 3rd ed. Oxford, England: Blackwell Scientific Publications, 1994
5. Brophy J, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *JAMA* 1995;273:871–875
6. Fleiss J. *Statistical methods for rates and proportions.* New York, NY: Wiley, 1981
7. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian data analysis,* 2nd ed. London, England: Chapman and Hall, 2003
8. Rosner B. *Fundamentals of biostatistics.* Belmont, MA: Duxbury, 1995
9. Bland M. *An introduction to medical statistics,* 3rd ed. Oxford, England: Oxford University Press, 2000
10. Le C. *Introductory biostatistics.* New York, NY: Wiley, 2003
11. R, version 1.8.0. Available at: cran.r-project.org/. Accessed February 2, 2004
12. O'Hagan A. First Bayes software. Available at: www.shef.ac.uk/~st1ao/1b.html. Accessed December 25, 2003
13. Spiegelhalter D, Thomas A, Best N. *WinBUGS version 1.4 user manual*. Cambridge, UK: MRC Biostatistics Unit, 2003
14. WinBUGS, version 1.4. Available at: www.mrc-bsu.cam.ac.uk/bugs/. Accessed February 2, 2004

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005

Fundamentals of Clinical Research for Radiologists

Philip E. Crewson[1]

Reader Agreement Studies

This article presents several approaches for evaluating reader agreement. The dominant technique in the radiology literature is weighted and unweighted Cohen's kappa and the associated measure, percent agreement. Percent agreement is an intuitive approach to measuring agreement but does not adjust for chance. Kappa provides a measure of agreement beyond that which would be expected by chance, as estimated by the observed data. Both the bi-rater and multirater kappa statistics have several limitations that are difficult to resolve. Although there are alternative approaches to measuring agreement, kappa remains the most commonly used measure.

Reader agreement studies have an important role in advancing radiology practice, technique, training, and quality control. Extremely common in the radiology literature, reader agreement studies determine the magnitude of agreement between or among readers. Potential applications include developing reliable diagnostic rules [1], understanding variability in treatment recommendations [2], evaluating the effects of training on interpretation consistency [3], determining the reliability of classification systems (lexicon development) [4], and comparing the consistency of different sources of medical information [5]. Agreement studies should not be confused with studies of accuracy in which measures of sensitivity and specificity and ROC curves are commonplace for comparisons when a reference standard (known truth) exists. Although these studies evaluate the validity of a measure and require a reference standard, agreement studies most commonly focus on the reliability of evaluations between different readers or in the same reader on different occasions; agreement studies do not require a reference standard.

Several methods are available for evaluating reader agreement, but the dominant technique in the radiology literature is weighted and unweighted Cohen's kappa and the associated measure, percent agreement. Because of the popularity of kappa in radiology research, this paper will focus on bi-rater and multirater kappa. Included in this presentation will be a discussion of the basic data requirements, calculation formulas, interpretation of the kappa coefficient as a measure of strength of agreement, and statistical significance testing. This discussion will be followed by an exploration of several limitations of kappa, especially those that pertain to comparability across studies. Formulas are provided in sufficient detail for those who wish to replicate the calculations, but an in-depth understanding of the mathematics is not necessary to appreciate the application and limitations of kappa.

Bi-Rater Kappa

Cohen's kappa is a common technique for estimating paired interrater agreement for nominal and ordinal-level data [6]. Kappa is a coefficient that represents agreement obtained between two readers beyond that which would be expected by chance alone [7]. A value of 1.0 represents perfect agreement. A value of 0.0 represents no agreement. Although such instances are rare, kappa can also exhibit negative values when observed agreement is less (worse) than chance. Key assumptions for using kappa include the following: elements being rated (images, diagnoses, clinical indications, and so forth) are independent of each other, one rater's classifications are made

Received November 17, 2004; accepted after revision November 23, 2004.

This is the 17th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1] Health Services Research and Development Service (124), Department of Veterans Affairs, 810 Vermont Ave., NW, Washington, DC 20420. Address correspondence to P. E. Crewson (philip.crewson@va.gov).

AJR 2005;184:1391–1397

0361–803X/05/1845–1391

independently of the other rater's classifications, the same two raters provide the classifications used to determine kappa, and the rating categories are independent of one another [8]. The last assumption may be difficult to satisfy in some imaging studies in which there are subtle differences in lesion characteristics and decision criteria. When differences between rating categories are not clear, careful study design is essential to maximize the independence among rating categories. Alternatives include dropping confusing categories or merging related categories. Although not always possible, adjustments in the classification scheme should be consistent with clinical practice.

Bi-rater kappa is used to test the hypothesis that agreement exists between two raters beyond that which would be expected by chance. Bi-rater kappa provides a measure of the relative intensity of agreement or disagreement between two readers rating the same elements using an identical classification system. A two-by-two contingency table illustrates hypothetic data in which two readers independently viewed the same set of 100 images from diagnostic mammograms with a simple classification criterion, malignant or benign (Table 1). To estimate kappa, both raters must use the same number of rating criteria so that the number of columns representing the rating categories used by rater 1 equals the number of rows representing the rating categories used by rater 2. Kappa is calculated using the formula:

$$\hat{k} = \frac{p_0 - p_e}{1 - p_e}$$

where p_o is the proportion of cases in which agreement exists between two raters, and p_e is the proportion of cases in which raters would agree by chance.

If we divide each cell count by the total sample size ($n = 100$), a matrix of probabilities is created (Table 2). Each cell contains the proportion of the total number of images ($n = 100$), not the count. As an example, the proportion of images in which reader 1 and reader 2 agree that an image is benign is 0.20 (20/100) or 20% (0.20 × 100).

TABLE 1 Two Readers Evaluating 100 Images (Counts)

Reader 2	Reader 1		
	Benign	Malignant	Total
Benign	20	5	25
Malignant	15	60	75
Total	35	65	100

The overall proportion of readings in which reader 1 and reader 2 agree is calculated by summing the diagonal probabilities in Table 2:

$$p_o = 0.20 + 0.60...p_0 = 0.80.$$

This "proportion agreement" is converted to a percentage and reported as "percent agreement." The interpretation of percent agreement is straightforward: Reader 1 and reader 2 agreed with each other on 80% of the classifications. The approach and calculations are the same for larger tables in which readers must consider more than two options in their decision making. Using as an example the American College of Radiology's BI-RADS lexicon [9] for final assessment, agreement could be based on each reader assigning each case to one of four categories: benign, probably benign, suspicious, or highly suggestive of malignancy. The resulting data would be reported in a four-by-four table in which the sum of the probabilities in the four diagonal cells represents the proportion agreement (p_o).

The advantage of the kappa statistic over percent agreement is its adjustment for the proportion of cases in which the raters would agree by chance alone. Because we are unlikely to know the true value of chance, the marginal probabilities from the observed data are used to estimate a surrogate for chance. The proportions in the total column and in the total row represent the marginal probabilities. Chance agreement is derived from the observed data, so it will likely change if different readers evaluate the same images. Using Table 2, the proportion of chance agreement (p_e) is computed as follows:

$$p_e = (0.35 \times 0.25) + (0.65 \times 0.75)$$
$$p_e = 0.09 + 0.49$$
$$p_e = 0.58$$

Once the proportion of observed agreement (p_o) and the proportion of chance agreement (p_e) are established, kappa is calculated using the formula:

$$\hat{k} = \frac{p_0 - p_e}{1 - p_e}$$

$$\hat{k} = \frac{.80 - .58}{1 - .58}$$

$$\hat{k} = .52$$

Using a common interpretation guideline offered by Landis and Koch [7], a kappa of 0.52 reflects a moderate level of agreement (Table 3).

Statistical Significance

To test the null hypothesis that the kappa coefficient is not different from zero (i.e., no better than chance), an estimate of the standard error (SE) for a one-sample test is calculated from the formula [10]:

$$SE_{k_0} = \sqrt{\frac{p_e}{n(1 - p_e)}}$$

$$SE_{k_0} = \sqrt{\frac{.58}{100(.42)}}$$

$$SE_{k_0} = 0.12$$

A kappa test statistic is compared with the standard normal distribution. The equation for obtaining the test statistic is as follows:

$$z = \frac{\hat{k}}{SE_{k_0}}$$

$$z = \frac{.52}{.12}$$

$$z = 4.33$$

Using a one-tailed test, the test statistic is statistically significant because it exceeds the critical value of 1.645 (alpha, 0.05) [6]. This result supports the alternative hypothesis that the kappa coefficient is different from zero (i.e., better than chance).

Although some effort has been directed toward estimating sample size requirements for comparisons among two or more kappa coefficients [11, 12], methods for calculating power for one kappa coefficient have not received much attention [11]. As a general rule of thumb, 30 cases with two readers is a reasonable minimum sample size as long as a moderate-level or better kappa coefficient ($\kappa > 0.40$) is expected and you want to show that kappa is different from a value of zero.

Confidence Intervals

For estimating confidence intervals, a different formula is used for SE [10]. There are other more accurate and complicated formulas for SE [6, 13, 14]:

$$SE_k \cong \sqrt{\frac{p_0(1-p_0)}{n(1-p_e)^2}}$$

$$SE_k \cong \sqrt{\frac{.80(.20)}{100(.42)^2}}$$

$$SE_k \cong .095$$

Given an estimate of kappa of 0.52, the 95% confidence interval would be 0.33–0.71:

$CI_{95\%} = \kappa \pm 1.96\ (SE_\kappa)$
$CI_{95\%} = 0.52 \pm 1.96(0.095)$
$CI_{95\%} = 0.52 \pm 0.19$

Weighted Kappa

Kappa treats disagreements the same regardless of whether a close decision on a rank-ordered classification system has clinical relevance. As an example, a rank-ordered rating scale from benign, probably benign, suspicious, and highly suspicious of malignancy, from which one rater concludes a lesion is suspicious and the other rater concludes that the lesion is highly suspicious, may result in the same clinical decision, immediate follow-up with biopsy. In this event, a disagreement between these two categories is much less important than a disagreement in which one rater rates a lesion as highly suspicious and the other rater rates the same lesion as probably benign.

Weighted kappa was developed to provide partial credit. The observed and expected proportions of each cell are multiplied by a weight before using them to calculate kappa. Weights can be established a priori (before data collection) using clinical experience [10], or they can be calculated after data collection using a simple algorithm for assigning weights that uses the same weighting strategy regardless of the data characteristics or rating criteria. Weighted kappa and unweighted kappa will be the same when there are only two decision categories. An example based on the BI-RADS classification system is provided in Appendix 1. For another example of calculating kappa weights, see Kundel and Polansky [15].

Special Considerations When Using Bi-Rater Kappa

For small sample sizes, kappa may be underestimated. In this case, a resampling technique (jackknifing) can be used to calculate an unbiased estimate of kappa [8]. Kappa may also be lower if the number of decision categories is excessive. Possible responses to compensate for this effect are to use weighted kappa if the categories are rank-ordered or to combine similar categories, or both. In any good study design, the choice of a weighting or classification scheme should be addressed and resolved before data collection. Overall, the precision (SE) of kappa is expected to improve as the number of patients and raters increases [16]. Although the preceding discussion was limited to two raters, the next section presents a technique for improving precision by comparing more than two raters.

TABLE 2 Two Readers Evaluating 100 Images (Proportions)

Reader 2	Reader 1		
	Benign	Malignant	Total
Benign	0.20	0.05	0.25
Malignant	0.15	0.60	0.75
Total	0.35	0.65	1.00

TABLE 3 Interpretation Guidance for Strength of Agreement

Kappa Coefficient	Strength of Agreement
< 0.00	Poor
0.00–0.20	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 taken from Landis and Koch [17].

Multirater Generalized Kappa

When there are more than two raters, generalized kappa is the recommended approach for evaluating interrater agreement [6, 13, 17]. This statistic measures the degree to which interpretation variability arises from differences among cases relative to differences among readers interpreting the same case. It is analogous to analysis of variance and the intraclass correlation used in the assessment of agreement when measured on a continuous scale.

The discussion that follows focuses on estimating agreement among more than two raters, when the number of raters is kept constant and the number of rating categories is greater than two. Slight modifications in the calculations are required when generalized kappa is estimated for only two rating categories or when the number of raters does not remain constant from one classification to another (see Fleiss [6] for alternative calculations). The approach presented here satisfies the likely characteristics of a prospective imaging study design [3].

Table 4 presents hypothetic data for five raters evaluating imaging from 10 patients using three decision categories—benign, suspicious, and malignant. The formulas that follow are from Woolson [13]. Assume the following notation: N = total number of patients, K = total number of raters, R = number of decision categories, and n_{ij} = number of raters who classified patient i (rows in Table 4) in category j (columns in Table 4).

The proportion ($\hat{p}_j$) of all classifications that fall within each decision category is presented at the bottom of each column. In this example, 0.40 (40%) of the classifications are in the benign category, 0.24 (24%) are suspicious, and 0.36 (36%) are classified as malignant.

TABLE 4 Ratings by Five Radiologists for 10 Patients

Patient (n_i)	Classification (n_j)			$\sum_{j=1}^{R}\frac{n_{ij}(n_{ij}-1)}{K(K-1)}$
	Benign	Suspicious	Malignant	
1	1	4	0	0.60
2	2	0	3	0.40
3	0	0	5	1.00
4	4	0	1	0.60
5	3	0	2	0.40
6	1	4	0	0.60
7	5	0	0	1.00
8	0	4	1	0.60
9	1	0	4	0.60
10	3	0	2	0.40
Total	20	12	18	
$\hat{p}_j$	0.40	0.24	0.36	$\bar{p}$ = 0.62

TABLE 5 **Implications of Case Distribution**

Benign and Malignant Cases Evenly Distributed				Malignant Cases Dominate Distribution (90%)			
Reader 2	Reader 1			Reader 2	Reader 1		
	Benign	Malignant	Total		Benign	Malignant	Total
Benign	0.45	0.05	0.50	Benign	0.05	0.05	0.10
Malignant	0.05	0.45	0.50	Malignant	0.05	0.85	0.90
Total	0.50	0.50		Total	0.10	0.90	
p_0	0.90			p_0	0.90		
p_e	0.50			p_e	0.82		
Kappa	0.80			Kappa	0.44		

Note.—p_0 = proportion of cases in which agreement exists between two raters (proportion observed), p_e = proportion of cases in which raters would agree by chance (proportion expected).

For each patient, the proportion of all possible pairings on which radiologists agree is calculated using the formula:

$$\sum_{j=1}^{R}\frac{n_{ij}\left(n_{ij}-1\right)}{K\left(K-1\right)}$$

For patient 1, this would be calculated as shown in:

$$\sum_{j=1}^{R}\frac{n_{ij}\left(n_{ij}-1\right)}{K\left(K-1\right)} = \frac{1(1-1)}{5(5-1)} + \frac{4(4-1)}{5(5-1)} + \frac{0(0-1)}{5(5-1)} = \frac{12}{20} = .60$$

The proportion of pairs agreeing for each patient is provided in Table 4 in the right column. The overall proportion of agreement ($\bar{p}$) is the mean agreement of all patients, or 0.62. In other words, we estimate that, on average, any two of the five radiologists will agree on a classification about 62% of the time.

As in bi-rater kappa, a correction for chance agreement is necessary to calculate the kappa coefficient. To estimate chance agreement for generalized kappa, the proportion ($\hat{p}_j$) of classifications in each decision category is squared and summed. For Table 4, the expected chance agreement is:

$$\bar{p}_e = \sum_{j=1}^{R}\hat{p}_j^2$$

$$\bar{p}_e = .40^2 + .24^2 + .36^2$$

$$\bar{p}_e = .35$$

Using the proportion of observed agreement and chance agreement, the generalized kappa statistic is:

$$\hat{K}_G = \frac{\bar{p}-\bar{p}_e}{1-\bar{p}_e}$$

$$\hat{K}_G = \frac{.62-.35}{1-.35}$$

$$\hat{K}_G = .42$$

Statistical Significance

To test the null hypothesis that the kappa coefficient is not different from zero (i.e., no better than chance), the generalized kappa statistic is compared with the standard normal distribution. The equation for obtaining the test statistic is as follows (see Appendix 2 for SE calculations):

$$z = \frac{\hat{K}_G}{SE_{K_G}} \qquad z = \frac{.42}{.075}$$

$$z = 5.6$$

For a one-tailed test (alpha = 0.05), the kappa coefficient is statistically significantly different from zero. Because of rounding, the SE and *z*-test statistic will be slightly different when calculated by computer algorithm, and there are other calculation methods for SE not presented here [6]. A confidence interval is created using the same procedure as that presented for bi-rater kappa using the generalized kappa coefficient and its SE.

Limitations of Kappa

Considerable debate surrounds the use of bi-kappa and generalized kappa as a measure of agreement [18]. As a result, several alternative approaches to measuring agreement have been proposed but have yet to gain wide acceptance in the peer-reviewed literature. A convenient listing of several alternative approaches and references is available on the Internet [19]. Given the dominance of kappa as a measure of agreement in imaging studies, it is important for both investigators and consumers of the literature to understand the limitations of kappa. Following is a brief discussion of the negative effects resulting from variations in case distribution, improper use of weights, and restrictions on the overall generalizability (external validity) of studies using kappa. This is not a complete listing of all the limitations, but rather basic considerations in interpreting any agreement study that uses kappa.

Effects of Case Distribution

A fundamental aspect of agreement studies is the distribution of cases. Because it is unlikely that a study reflects the population prevalence, marginals (row and column totals) based on reader agreement patterns are routinely used as surrogates for prevalence [18]. This surrogate measure of chance agreement is based on the distribution of the cases classified by readers (both bi-rater kappa and generalized kappa). It is possible to find a consistently high level of percent agreement while reporting widely differing kappa values from one study or one comparison to another because of the case distributions. Table 5 provides an example in which two readers with the same percent agreement are presented with differing distributions of cases. The examples provided in Table 5 assume a high level of accuracy by both readers, so that the marginal probabilities match the study case distribution. In both examples, the readers agree in 90% of the classifications; however, kappa is significantly reduced if one classification category dominates. As shown, an increase in the dominance of malignant cases from 50% to 90% resulted in kappa dropping from 0.80 to 0.44.

Limitation 1.—Because of variations in case mix, reported kappa values may vary dramatically from one study to another even when the overall percent agreement is similar.

Limitation 2.—Because varying rater pairs will likely change the category distributions, bi-kappa values on the same set of elements may vary dramatically from one reader pair to another, even when percent agreement is relatively stable.

Weighted Kappa

Adding to the limited comparability of the kappa statistic from one study to another is the use of weighted kappa. There are multiple methods to weight kappa, so the comparability between studies is often limited. This concern, however, is minor when compared with the problem of weight justification [13]. The assignment of weights is an arbitrary exercise, even when an established algorithm is used [6, 7]. The subjectivity of assigning weights should be balanced with a clear explanation of why and how the weights are used [10]. Unfortunately, it is not rare for agreement studies to report weighted kappa with little if any discussion regarding the justification for the weighting scheme used in the study.

Limitation 3.—Weighting schemes are often subjective.

Generalizability

Several factors affect the generalizability (external validity) of an agreement study. These include rater background, clarity of the decision categories, and clinical relevance.

Rater Background.—When using kappa, we assume that the raters have similar levels of experience, training, and specialization (e.g., general radiology residents are not paired with seasoned subspecialists). If this is not the case, kappa may not be an appropriate technique [6].

Limitation 4.—Agreement is likely to be underestimated when raters have dissimilar experience and training.

Characteristic Clarity.—Clear classification definitions and independence are essential in an agreement study. As a result, if a general understanding regarding the basic concepts being rated has not been reached, conducting an agreement study is premature and inappropriate. Similarly, if the difference between classification categories is not clear, agreement will suffer and may not reflect the actual domain of interest. As an example, is there an actual difference between "probably benign" and "suspicious," or do radiologists treat them clinically the same? In this case, reasons for possible differences among radiologists may include variation in attitudes toward the risk associated with false-negatives and unfamiliarity with subtle differences among the rating categories [2, 20]. It is unwise to give much credence to an agreement study that was based on a questionable classification scheme. An exception would be pilot studies such as lexicon development efforts, but they should be treated as experimental (efficacy) studies.

TABLE 6 Calculations for Weighted Kappa: Cell Counts

Reader B (*i*)	Reader A (*j*)					
	Benign	Probably Benign	Suspicious	Malignant	Row Total	Row Proportion Observed (p_{oj})
Benign	4	1	0	0	5	0.17
Probably benign	1	3	1	1	6	0.20
Suspicious	1	4	5	0	10	0.33
Malignant	0	1	2	6	9	0.30
Column total	6	9	8	7	30	
Column proportion observed (p_{oi})	0.20	0.30	0.27	0.23		

TABLE 7 Calculations for Weighted Kappa: Proportions Observed and Proportions Expected

Proportions Observed (p_o)

Reader B	Reader A				Formula
	Benign	Probably Benign	Suspicious	Malignant	
Benign	0.13	0.03	0.00	0.00	$p_o = n_{ij} / n$
Probably benign	0.03	0.10	0.03	0.03	
Suspicious	0.03	0.13	0.17	0.00	
Malignant	0.00	0.03	0.07	0.20	

Proportions Expected (p_e)

Reader B	Reader A				Formula
	Benign	Probably Benign	Suspicious	Malignant	
Benign	0.03	0.05	0.04	0.04	$p_e = p_o \times p_{oj}$
Probably benign	0.04	0.06	0.05	0.05	
Suspicious	0.07	0.10	0.09	0.08	
Malignant	0.06	0.09	0.08	0.07	

Limitation 5.—Agreement is likely to be underestimated and not generalizable when rating categories have questionable face validity.

Clinical Relevance.—A general question for any agreement study is whether the observed agreement is representative of clinical practice. Factors to consider include the type of imaging technology used, amount of background information provided, type of imaging (diagnostic or screening), prior imaging results, time allowed for interpretation, prior risk of disease, and comorbidity.

Limitation 6.—Agreement studies often do not reflect actual clinical practice (less information) or imaging prevalence (case mix), so the generalizability of the findings may be overstated.

Conclusion

Reader agreement studies have an important role in advancing radiology practice, technique, training, and quality control. Although the limitations of kappa are known, it remains a common statistical technique for estimating agreement for nominal and ordinal scale variables. The purpose of this article has been to build a better understanding of both the bi-rater and multirater kappa statistic. As has been shown, several weaknesses are intrinsic to kappa that are difficult to resolve.

TABLE 8 Calculations for Weighted Kappa: Quadratic Cell Weights (*w*)

Reader B Weights (i_w)	Reader A Weights (j_w)				Formula
	1	2	3	4	
1	1.00	0.89	0.56	0.00	$w_{ij} = 1 - \frac{(i_w - j_w)^2}{(k-1)^2}$
2	0.89	1.00	0.89	0.56	
3	0.56	0.89	1.00	0.89	
4	0.00	0.56	0.89	1.00	

TABLE 9 Calculations for Weighted Kappa: Weighted Proportions Observed and Weighted Proportions Expected

Weighted Proportions Observed: $p_o(w)$					
Reader B	Reader A				Formula
	Benign	Probably Benign	Suspicious	Malignant	
Benign	0.13	0.03	0.00	0.00	$Po(w) = \sum(Po_{ij} * w_{ij})$ $Po(w) = 0.93$
Probably benign	0.03	0.10	0.03	0.02	
Suspicious	0.02	0.12	0.17	0.00	
Malignant	0.00	0.02	0.06	0.20	
Weighted Proportions Expected: $p_e(w)$					
Reader B	Reader A				Formula
	Benign	Probably Benign	Suspicious	Malignant	
Benign	0.03	0.04	0.02	0.00	$Pe(w) = \sum(Pe_{ij} * w_{ij})$ $Pe(w) = 0.75$
Probably benign	0.04	0.06	0.05	0.03	
Suspicious	0.04	0.09	0.09	0.07	
Malignant	0.00	0.05	0.07	0.07	

Although there are alternative approaches to measuring agreement, kappa will likely remain the most commonly used measure. Issues hindering the use of alternatives include mathematic complexity, reduced understanding and interpretability, and lack of consistency with prior research.

At present, agreement studies will continue to use bi-rater kappa, multirater kappa, and weighted kappa as a measure of agreement. However, it is essential that researchers respond to the limitations of kappa not only by improving study design but also by reporting and interpreting the findings appropriately. Recommended steps to improve the quality and usefulness of published reader agreement studies include reporting the characteristics of the raters and their similarities and differences; reporting the source and characteristics of the elements (images) presented to raters; including percent agreement with any kappa coefficient, and including both percent agreement and unweighted kappa if weighted kappa is used; and tempering overgeneralization by reflecting on how the raters, the elements they rated, and the study design differ from general clinical practice. Although the limitations of the kappa statistic may seem insurmountable, the key to proper use and interpretation of kappa, and any other statistic, is understanding its limitations and reporting sufficient data so that others may judge the results.

Acknowledgments

I thank Caryn Cohen, the series editors, and anonymous reviewers for comments on earlier drafts of the manuscript.

References

1. Kinkel K, Helbich TH, Esserman LJ, et al. Dynamic high-spatial-resolution MR imaging of suspicious breast lesions: diagnostic criteria and interobserver variability. *AJR* 2000;175:35–43
2. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists' interpretations of mammograms. *N Engl J Med* 1994;331:1493–1499
3. Berg WA, D'Orsi CJ, Jackson VP, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? *Radiology* 2002;224:871–880
4. Ikeda DM, Hylton NM, Kinkel K, et al. Development, standardization, and testing of a lexicon for reporting contrast-enhanced breast magnetic resonance imaging studies. *J Magn Reson Imaging* 2001;13:889–895
5. Kashner TM. Agreement between administrative files and written medical records: a case of the Department of Veterans Affairs. *Med Care* 1998;36:1324–1336
6. Fleiss JL. *Statistical methods for rates and proportions,* 2nd ed. New York, NY: Wiley, 1981
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174
8. Cyr L, Francis K. Measures of clinical agreement for nominal and categorical data: the kappa coefficient. *Comput Biol Med* 1992;22:239–246
9. American College of Radiology. *Illustrated Breast Imaging Reporting and Data System (BI-RADS),* 3rd ed. Reston, VA: American College of Radiology, 1998
10. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–220
11. Lin H-M, Williamson JM, Lipsitz SR. Calculating power for the comparison of dependent K-coefficients. *Appl Stat* 2003;52:391–404
12. Donner A. Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Stat Med* 1998;17:1157–1168
13. Woolson RF. *Statistical methods for the analysis of biomedical data.* New York. NY: Wiley, 1987
14. Lee JJ, Tu ZN. A better confidence interval for kappa on measuring agreement between two raters with binary outcomes. *J Computat Graph Stat* 1994;3:301–321
15. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003;228:303–308
16. Kraemer HC. *Evaluating medical tests: objective and quantitative guidelines.* Newbury Park, CA: Sage Publications, 1992
17. Landis JR, Koch GG. A one-way components of variance model for categorical data. *Biometrics* 1977;33:671–679
18. Feinstein AR, Cicchetti DV. High agreement but low kappa. I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543–549
19. Uebersax J. ourworld.compuserve.com/homepages/jsuebersax/agree.htm Accessed November 16, 2003
20. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. *Acad Radiol* 1996;3:891–897

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005
16. Statistical Inference for Proportions, April 2005

APPENDIX 1. Weighted Kappa

The data and formulas used to calculate weighted kappa are shown in Tables 6–9. This example of weighted kappa is based on a four-category BI-RADS scale. Using a weighting scheme from Fleiss [6], a weight factor of 1 is used for benign, 2 for probably benign, 3 for suspicious, and 4 for malignant. The difference between the weight factors is used to estimate a weight for each cell. For example, if both readers classify the same set of lesions as malignant (an exact match), each decision has a weight factor of 4. Using the Fleiss formula results in a cell weight of 1. A weight of 1 allows the entire proportion of lesion classifications in this cell (observed and expected proportions) to contribute to the kappa estimate (see diagonal data in Table 8).

In contrast, all other lesion classification alternatives (mismatches) are adjusted according to the difference between their weight factors. As an example, if one reader classifies a set of lesions as malignant (weight factor of 4) and the other reader classifies the same set of lesions as suspicious (weight factor of 3), the proportion of lesion classifications in this cell that contribute to kappa are reduced (i.e., given less importance for estimating kappa than an exact match)—in this case, 89% (.89) as much weight as an exact match. As the difference between weight factors increases, the contribution to kappa from that cell decreases to the point at which none of the observations in a cell contribute to kappa (i.e., instances in which the same set of lesions is classified malignant by one rater and benign by the other).

Observed proportions of agreement and expected proportions of agreement are calculated for each cell and then weighted (multiplied) by the quadratic cell weight and summed. The resulting weighted kappa is 0.72, which is greater than the unweighted kappa (0.47).

$$k(w) = \frac{Po(w) - Pe(w)}{1 - Pe(w)}$$

$$k(w) = \frac{0.93 - 0.75}{1 - 0.75}$$

$$k(w) = 0.72$$

APPENDIX 2. SE for Generalized Kappa

To test the null hypothesis that the kappa coefficient obtained in Table 4 is not different from zero (i.e., no better than chance), an estimate of the SE is calculated using the formula shown.

$$SE_{K_G} = \sqrt{\left(\frac{2}{NK(K-1)}\right) * \frac{\sum_{j=1}^{R} \hat{p}_j^2 - (2K-3)\left(\sum_{j=1}^{R} \hat{p}_j^2\right)^2 + 2(K-2)\sum_{j=1}^{R} \hat{p}_j^3}{\left(1 - \sum_{j=1}^{R} \hat{p}_j^2\right)^2}}$$

$$SE_{K_G} = \sqrt{\left(\frac{2}{50(4)}\right) * \frac{.35 - (7)(.35)^2 + 2(3)(.40^3 + .24^3 + .36^3)}{(1 - .35)^2}}$$

$$SE_{K_G} = .075$$

Correlation and Regression

Nandini Dendukuri[1,2]
Caroline Reinhold[3,4]

Received November 17, 2004; accepted after revision November 23, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 18th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology.* The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Technology Assessment Unit, Royal Victoria Hospital, Montreal, QC H3A 1A1, Canada.

[2]Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave. W, Montreal QC H3A 1A2, Canada. Address correspondence to N. Dendukuri (nandini.dendukuri@mcgill.ca).

[3]Department of Diagnostic Radiology, Montreal General Hospital, McGill University Health Centre, 1650 Cedar Ave., Montreal QC H3G 1A4, Canada.

[4]Department of Oncology, Synarc, 575 Market St., San Francisco CA, 94105.

AJR 2005;185:3–18

0361–803X/05/1851–3

This module covers common statistical methods used in radiologic applications for measuring relations between variables. Under the topic of correlation we describe Pearson's and Spearman's correlation coefficients and partial correlation, all of which are suitable for evaluating the association between two continuous variables. In the section on regression we cover linear and logistic regression models. Regression models are used to study the association between an outcome variable and one or more predictor variables that may be continuous or dichotomous. For linear regression models the outcome variable is continuous, whereas for logistic regression models it is dichotomous. We also briefly describe methods for model selection and sample size determination.

In a hypothetical study evaluating the use of MRI for the assessment of myocardial viability, researchers were interested in characterizing the nature of the relation between myocardial infarct volume and ejection fraction. Their objective was to answer questions such as: Is there any relation between infarct volume and ejection fraction? What is the strength of this relation? Does ejection fraction increase or decrease with increasing myocardial infarct volume? By how much would we expect the ejection fraction to change when the myocardial infarct volume increases by 1 mL? Can we predict a patient's ejection fraction when given his or her myocardial infarct volume? How accurate is this prediction?

Questions such as these arise in situations in which more than one variable has been measured on each patient (or observational unit) in a sample, and the relationship between the different variables is of interest. This module covers some of the most commonly used statistical tools to answer such questions: correlation coefficients and regression models. We will cover methods for studying the relation between two variables that may be both continuous, both dichotomous (i.e., having only two values), or a mix (one dichotomous and the other continuous). We will also cover situations in which we wish to study the relation between more than two variables.

To illustrate the methods in this tutorial we have used hypothetical examples that are all inspired from studies appearing in radiology research journals. Some of the concepts covered in this tutorial assume knowledge of earlier articles in this series, to which the reader is encouraged to refer [1–4].

Correlation

In Figure 1 we have two scatterplots between ejection fraction and myocardial infarct volume. At first glance, it appears that the relation between the two variables is stronger in Figure 1A than in Figure 1B. In fact, the two figures are based on the same data from a hypothetical study of 30 patients. Altering the scale of the ejection fraction axis makes the relation observed in Figure 1B appear less strong than in Figure 1A. The purpose of this figure is to illustrate that a scatterplot alone is not sufficient to make conclusions about the strength of the relationship between two variables. The plot needs to be accompanied by an objective measure.

Pearson's Correlation Coefficient

Pearson's correlation coefficient is one such objective measure of the linear relation between two variables. Pearson's correlation coefficient (which we denote by r_P) between two variables X (e.g., infarct volume) and Y (e.g., ejection fraction) is given by:

$$r_P = \text{Correlation (X, Y)} = \frac{\text{Covariance(X, Y)}}{\sqrt{\text{Variance(X) Variance(Y)}}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}},$$

where x_i and y_i are the values of variables X and Y observed on each individual in the sample, $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, and N is the number of individuals in the sample. The denominator of this expression is the square root of a positive quantity and is always taken to be positive. The numerator, on the other hand, can be positive or negative depending on the nature of the relation between X and Y. If X tends to increase when Y increases, then it is likely that when an individual x_i exceeds the sample mean $\bar{x}$, the corresponding y_i also exceeds its mean $\bar{y}$. This would cause the numerator, and thus r_P itself, to be positive. If, on the other hand, X decreases as Y increases, it is likely that x_i is less than $\bar{x}$ when y_i is greater than $\bar{y}$. This would result in a negative value of the numerator and of r_P. In the example in Figure 1, we find that ejection fraction tends to decrease with increasing myocardial infarction volume. Thus, patients whose ejection fraction exceeds the mean ejection fraction of the sample are more likely to have myocardial infarct volumes that are smaller than the mean myocardial infarct volume of the sample, resulting in a negative value of r_P.

Pearson's correlation coefficient can range from a minimum value of −1 to a maximum value of 1. Figure 2 illustrates the value of r_P in various prototypical situations. A value of $r_P = 1$ is obtained when an increase in X is always associated with an increase in Y and the points in the scatterplot between X and Y can be joined to form a perfect straight line (Fig. 2A). A value of $r_P = -1$ is indicative of a perfect negative linear relation between X and Y (Fig. 2B). As the strength of the linear relation between X and Y diminishes, the value of r_P approaches 0 (Figs. 2C and 2D). A correlation coefficient of 0 indicates that there is no relation between the two variables. For the hypothetical data in Figure 1 we find that r_P is −0.91, suggesting a fairly strong negative relation between myocardial infarct volume and ejection fraction. The interested reader is referred to the table at the end of the appendix for a more detailed explanation of how to calculate the correlation coefficient.

Figures 2E and 2F illustrate two situations in which there is a perfect, though nonlinear, relation between X and Y. In Figure 2E, an increase in X is always accompanied by an increase in Y. Here, r_P is quite high (0.92), although not equal to 1. In Figure 2F we have a U-shaped relation between the variables, with both low and high values of X being associated with high values of Y. Here r_P is close to 0, suggesting only a weak relation between X and Y. These plots serve to illustrate that a value of r_P close to 0 does not rule out the possibility of a strong nonlinear relationship between the variables.

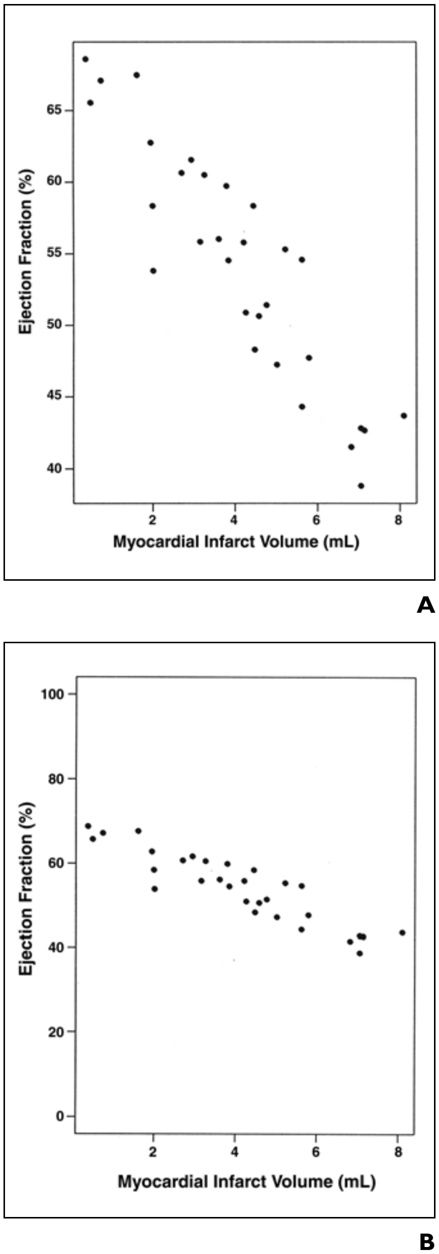


Fig. 1—Scatterplots show relation between myocardial infarct volume and ejection fraction and illustrate effect of changing scale of ejection fraction axis.
A and **B**, Relation between the two variables may appear stronger in **A** than in **B**, but both figures are based on same data. Altering scale of ejection fraction axis makes relation in **B** appear less strong than in **A**.

Interpreting Pearson's correlation coefficient—A few things need to be kept in mind when interpreting a correlation coefficient:

(1) Correlation is independent of the units in which the two variables are measured. If our interest is in measuring the strength of the relation between ejection fraction and myocardial infarct volume, it does not matter whether the latter was measured in milliliters (mL) or liters (L).

(2) High correlation may indicate a strong association but not causation. Note that in the expression for r_P, X and Y may be interchanged with no difference to the result. This means that the variables X and Y are not distinguished as "predictor" and "outcome" and it does not matter whether X causes Y or vice versa. It would be incorrect to assume that a high correlation between myocardial infarct volume and ejection fraction means that one of them is the cause of the other. Rather, we can only say that there is a strong association between them.

(3) The observed correlation (or lack of it) may be due to a confounding variable. In some situations the observed association (or lack of it) may be spurious and, in fact, reflect the effect of a third variable, referred to by epidemiologists as a "confounding variable" [5]. Such a variable is associated with both X and Y. Figure 3A is a scatterplot of the relation between endometrial thickness (measured at transvaginal sonography) and peak systolic velocity (measured at Doppler imaging) in postmenopausal women presenting with abnormal vaginal bleeding. The value of r_P for the entire sample is only moderate (r_P = 0.36). The sample was then divided into women with endometrial atrophy, those with endometrial hyperplasia, and those with endometrial carcinoma, and r_P was calculated separately within each group. We find that the true strong relation between endometrial thickness and peak systolic velocity is obscured because both variables have an association with the histologic subgroups.

(4) Correlation between aggregate values is stronger than at the individual level. In Figure 3B, the blank circles form a scatterplot of endometrial thickness versus peak systolic velocity in postmenopausal women presenting at three health centers (university-based, community hospital, and walk-in clinic). The dark circles plot the relation between the average endometrial thickness and average peak systolic velocity for each of the three health centers. The correlation between the average values is almost 1, despite a weaker correlation at the patient level.

(5) Correlation is influenced by the range of the X and Y variables. The greater the range

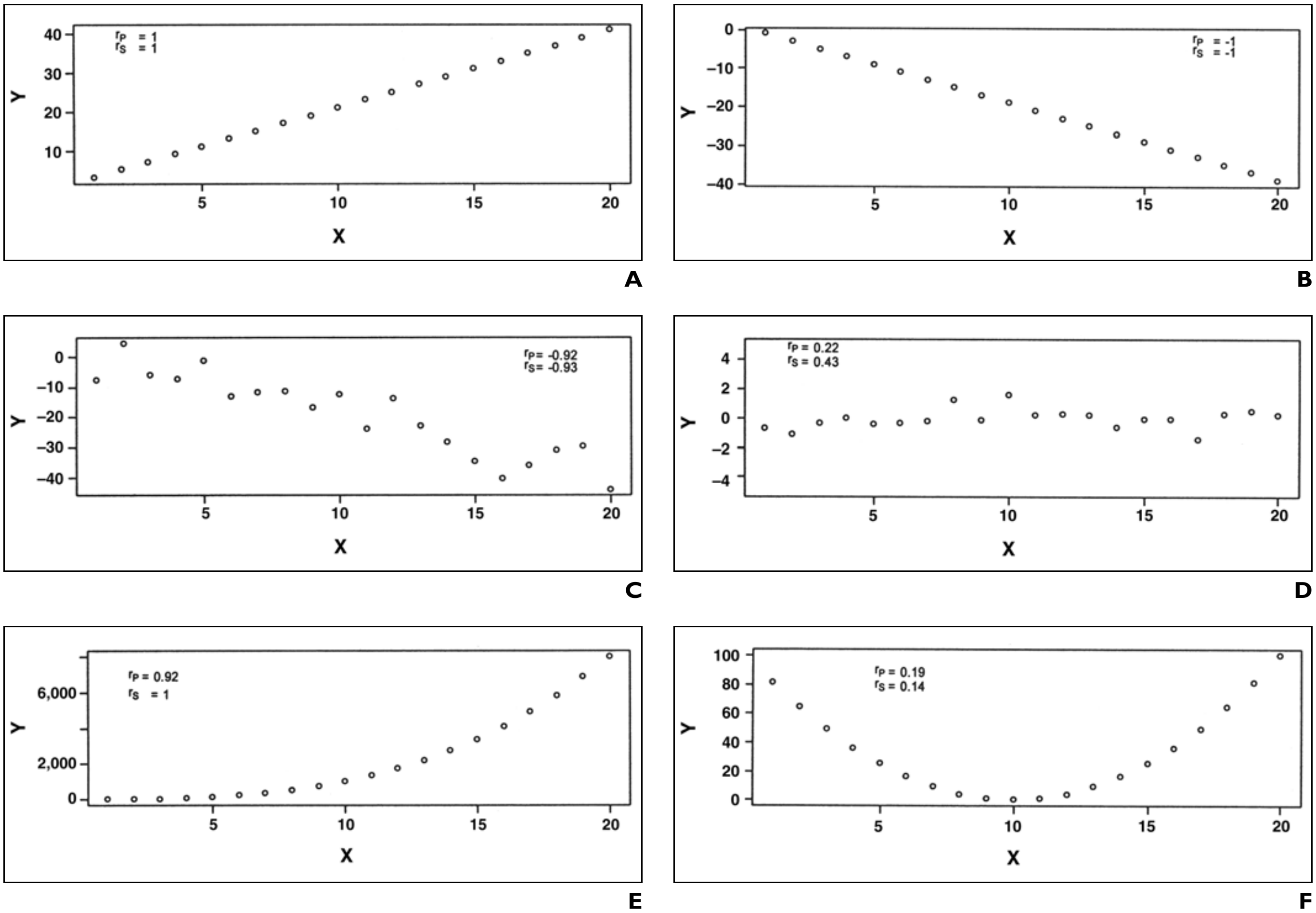


Fig. 2—Examples of different values of Pearson's (r_P) and Spearman's (r_S) correlation coefficients.
A, Value of $r_P = 1$ is obtained when increase in *X* is always associated with increase in *Y* and points in scatterplot form a straight line.
B, Value of $r_P = -1$ is indicative of negative linear relation between *X* and *Y*.
C and **D,** As strength of linear relation between *X* and *Y* diminishes, value of r_P approaches 0.
E and **F,** Plots show nonlinear relation between *X* and *Y*.

of the *X* and *Y* variables in the sample, the greater the correlation between them. Thus, a single outlying observation might give us a falsely elevated correlation coefficient.

(6) High correlation does not mean measurement equivalence. When comparing two imperfect measurements of the same underlying quantity, a high correlation is often used as a proof of strong agreement, but that is not correct. For example, we might be interested to determine whether measurements of the length of liver lesions using MRI and sonography are equivalent. A high positive correlation suggests only that increasing values of one measure are associated with increasing values of the second; it does not necessarily mean that they are measuring the same thing. A better approach to evaluating equivalence would be to examine the difference in magnitude of the observations on each patient. A large mean difference would suggest that the two measures are in fact not equivalent [6].

Assumptions used in calculating Pearson's correlation coefficient—Some important things need to be kept in mind before calculating r_P. First, it is based on the assumption that both *X* and *Y* are measured on an interval scale. When we say myocardial infarct volume has been measured on an interval scale, we mean that a myocardial infarct volume of 4 mL is twice as large as a myocardial infarct volume of 2 mL. This would not have been true if it were measured by a nominal variable having values 1 (small), 2 (medium), and 3 (large) because we cannot say that a patient rated as "medium" has twice the myocardial infarct volume of a patient rated as "small." Second, both *X* and *Y* are assumed to follow a normal probability distribution [2]. This assumption allows us to perform hypothesis tests and construct confidence intervals for r_P, as we will see.

Inference for Pearson's correlation coefficient—The sample correlation coefficient, r_P, is a statistic the value of which changes depending on the sample collected. It is only an estimate of the population correlation coefficient, ρ_P, that we would have obtained if it were possible to observe the entire population of patients (or study units) from which the sample was collected. When reporting the sample correlation coefficient, we also need to report some measure of our uncertainty in the knowledge of the population correlation coefficient. This uncertainty may be expressed in terms of a *p* value or a confidence interval [3]. Confidence intervals are preferred to *p* values because they provide more information regarding the parameter estimated. An earlier article in this series ex-

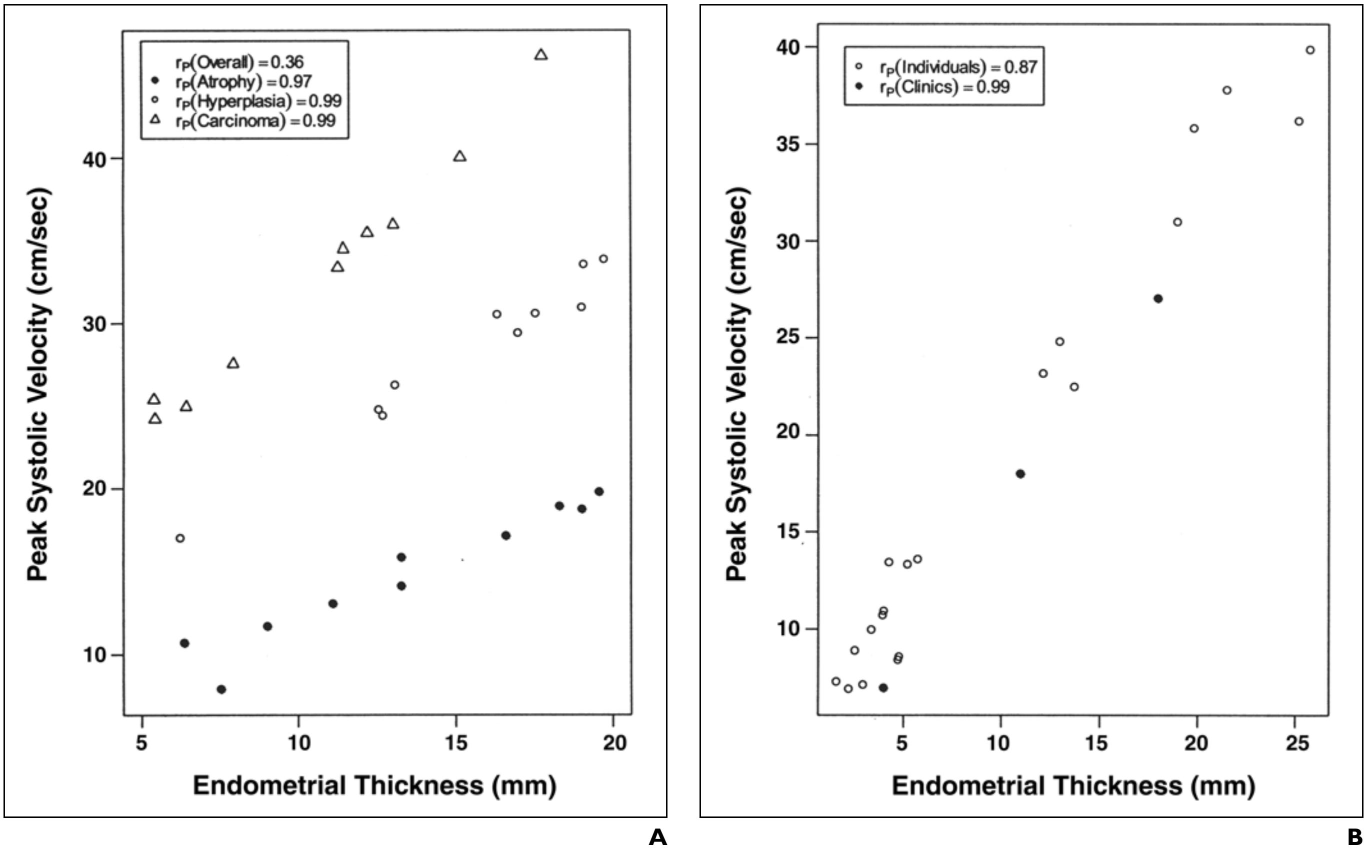


Fig. 3—Pearson's correlation coefficients.
A and **B**, Graphs show correlation coefficients in the presence of confounding (**A**) and from aggregate data (**B**).

plains in detail the distinction between confidence intervals and *p* values [3]. However, *p* values are still frequently reported in the medical literature, so we cover methods for their calculation and interpretation here.

p value: A *p* value measures the strength of the evidence in favor of a null hypothesis of the form H_0: $\rho_P = \rho_0$, where ρ_0 is a predetermined value of the correlation coefficient of interest. In our example on myocardial infarct volume and ejection fraction, we can set $\rho_0 = 0$ to measure the evidence in favor of "no association between the two variables." When the *p* value is very low (typically < 0.05 or 0.01) we reject the null hypothesis. Details on how to calculate the *p* value are provided for the interested reader in Appendix 1. We find that the *p* value for our example is very, very small (<< 0.001). In other words, the probability that we would have observed a correlation as strong as $r_P = -0.91$, when in fact the true correlation between myocardial infarct volume and ejection fraction was $\rho_P = 0$, is very, very small—much less than 0.0001. Therefore, we reject the null hypothesis of H_0: $\rho_P = 0$ and conclude that there is an association between myocardial infarct volume and ejection fraction.

Confidence interval: The hypothesis testing approach limits us to a single hypothesis, which is often artificially set up. Rather than simply concluding that the population correlation coefficient is not 0, we might want to say a little more about the strength of the correlation. A confidence interval is more informative in that it gives us the range of possible values of ρ_P that are compatible with the observed value of the correlation coefficient. Details of the calculation of the confidence interval are given in Appendix 1. The 95% confidence interval for the correlation coefficient between myocardial infarct volume and ejection fraction is (–0.96 to –0.81). If our hypothetical study were repeated several times and a confidence interval calculated each time, then 95% of the confidence intervals would capture the true value of ρ_P. However, we cannot say if the interval obtained from our sample is one of the 95% that capture the true value of ρ_P (see [3] for more details on how to interpret a confidence interval). The 95% confidence interval may also be interpreted as the range of values of the null hypothesis (ρ_0) that cannot be rejected at the $1 - 0.95 = 0.05$ level of significance.

The fact that our 95% confidence interval does not include 0 means that the null hypothesis of $\rho_0 = 0$ would be rejected, which is the same conclusion we reached earlier using the *p* value. A better approach would be to compare the confidence interval with a predetermined range of values indicative of no relation between the variables. For example, let us say that a correlation coefficient in the range from –0.1 to 0.1 is in practice indicative of no relation between myocardial infarct volume and ejection fraction. Then the fact that our confidence interval clearly lies outside this region leads us to conclude there is a strong, negative relation between myocardial infarct volume and ejection fraction.

Partial Correlation

It is possible that the observed correlation between two variables (*X* and *Y*) may be in part because of a third variable (*Z*) that is related to both of these variables. When this third confounding variable is also observed, we may be interested in estimating the correlation between *X* and *Y* after eliminating the effect of their correlation with *Z*. For example, in a study of liver lesion characterization using three diagnostic tests—sonography, CT, and MRI—the Pearson's correlation coefficient between the accuracy of the different diagnostic tests was as shown in the following equations:

$$r_P\ (\text{sonography, MRI}) = 0.7$$

$$r_P\ (\text{CT, sonography}) = 0.8$$

$$r_P\ (\text{CT, MRI}) = 0.9$$

Clearly, all three methods are correlated with each other. What is the correlation between the diagnostic performance of sonography and MRI alone, after eliminating the effect of the correlation that both have with CT? To estimate this, we can calculate a partial correlation coefficient. The partial correlation between *X* and *Y* after having eliminated the effect of a third variable *Z* is given by:

$$r_{XY.Z} = \frac{r_P(X,Y) - r_P(X,Z)r_P(Y,Z)}{\sqrt{1-r_P(X,Z)^2}\sqrt{1-r_P(Y,Z)^2}}$$

If *Z* is not a confounding variable, one or both of r_P (*X,Z*) and r_P (*Y,Z*) would be 0 or very small. In such a situation, the partial correlation between *X* and *Y* ($r_{XY.Z}$) would be similar to the Pearson's correlation coefficient between them (r_P [*X,Y*]).

The partial correlation coefficient between performance in sonography and MRI in our example is shown in these equations (where *US* = sonography):

$$r_{US\ MRI.CT} = \frac{r_P(US,MRI) - r_P(US,CT)r_P(MRI,CT)}{\sqrt{1-r_P(US,CT)^2}\sqrt{1-r_P(MRI,CT)^2}}$$

$$= \frac{0.7 - 0.8 \times 0.9}{\sqrt{1-0.8^2}\sqrt{1-0.9^2}} = -0.08$$

Thus, after eliminating the contribution of CT, we find that the strong relation between sonography and MRI vanishes. Moreover, it appears that the direction of the relation changes as well, suggesting that after removing the contribution of CT, lesions that are accurately diagnosed with sonography in fact are poorly diagnosed with MRI and vice versa.

This concept can be extended to calculate the partial correlation between two variables after adjusting for the effect of two or more variables. Multiple regression, which is discussed later in this article, can be used for the same purpose and is more straightforward to perform using commonly available statistical software packages.

Spearman's Rank Correlation

Spearman's rank correlation, which we denote by r_S, is another statistic used for measuring the correlation between a pair of variables. It is called a nonparametric measure and is preferred when assumptions required for calculating Pearson's correlation coefficient are violated—that is, when *X* and/or *Y* are not measured on an interval scale, or when *X* and/or *Y* do not follow a normal probability distribution. To calculate Spearman's correlation coefficient, we need to assign a rank to the individual values of *X* and *Y*—that is, sort each of *X* and *Y* in increasing order and assign them ranks so that the smallest observation has a rank of 1 and the highest observation has a rank of *N*. The expression for Spearman's correlation coefficient is similar to Pearson's correlation coefficient, except that x_i and y_i are replaced by the rank(x_i) and rank(y_i) as follows:

$$r_S = \frac{\sum_{i=1}^{N}\left(rank(x_i) - \frac{N+1}{2}\right)\left(rank(y_i) - \frac{N+1}{2}\right)}{\sqrt{\sum_{i=1}^{N}\left(rank(x_i) - \frac{N+1}{2}\right)^2 \sum_{i=1}^{N}\left(rank(y_i) - \frac{N+1}{2}\right)^2}}$$

Spearman's correlation coefficient ranges between −1 and 1, with these extreme values indicating a perfect negative or positive relationship, respectively, between *X* and *Y*. It takes the value 0 when there is no relation between the variables (Figs. 2A–2D). An advantage of Spearman's correlation coefficient over Pearson's correlation coefficient is that it can be used to evaluate a nonlinear relation between variables when the direction of the relationship does not change. In Figure 2E, where *Y* continuously increases with *X*, we see that the perfect nonlinear relationship between the variables is captured by Spearman's correlation coefficient, although not by Pearson's correlation coefficient. However, like $r_{P,}$ r_S is inappropriate for measuring the strength of a nonlinear relationship that both increases and decreases, such as the U-shaped relation in Figure 2F.

Regression

The correlation coefficients described thus far can be used to measure the strength and the direction of an association. Regression models go a step further and can be used to predict the value of one variable given the other. This quality makes them suitable for the study of relationships when the two variables can be distinguished as "predictor" and "outcome." Note, however, that fitting a regression equation between two variables does not imply a causal relation between them. Regression models also provide a more straightforward approach to adjusting for the effect of confounding variables. They can be used to deal with a variety of types of outcome variables (continuous, dichotomous, ordinal, count data, and so forth). Here, we focus on two of the most commonly used models for radiologic applications—linear regression models, in which the outcomes are continuous, and logistic regression models, in which the outcomes are dichotomous.

Regression is a broad area to which this article provides but a brief introduction. Greater detail on estimation and inference for linear and logistic regression is covered in introductory biostatistics textbooks [7–9]. More complex topics, such as regression model diagnostics, variable selection, and logistic regression for ordinal variables, are covered in greater depth in advanced textbooks [10–13].

Simple Linear Regression

Like Pearson's correlation coefficient, simple linear regression is also used to characterize linear relationships between variables. It is distinguished from multiple variable linear regression (discussed later) in that it involves only two variables, the outcome or dependent variable and the predictor or independent variable. The standard form of the simple linear regression equation is as follows:

$$Y = \alpha + \beta X + \varepsilon,$$

where *X* and *Y* are the observed values of the predictor and the outcome variables, respectively. The parameters α and β are called the intercept and the slope, respectively. For a given value of *X*, the predicted value of *Y* is $\alpha + \beta X$. The term ε, the residual (or error), is the difference between the observed value of *Y* and the predicted value of *Y*. The intercept and slope parameters are estimated with the aim of reducing this difference. The estimated values of the intercept and slope are denoted by *a* and *b*, respectively. An important assumption of the linear regression model is that the residuals are assumed to follow a normal distribution with mean 0 and a variance σ^2, which remains constant for all values of *X*.

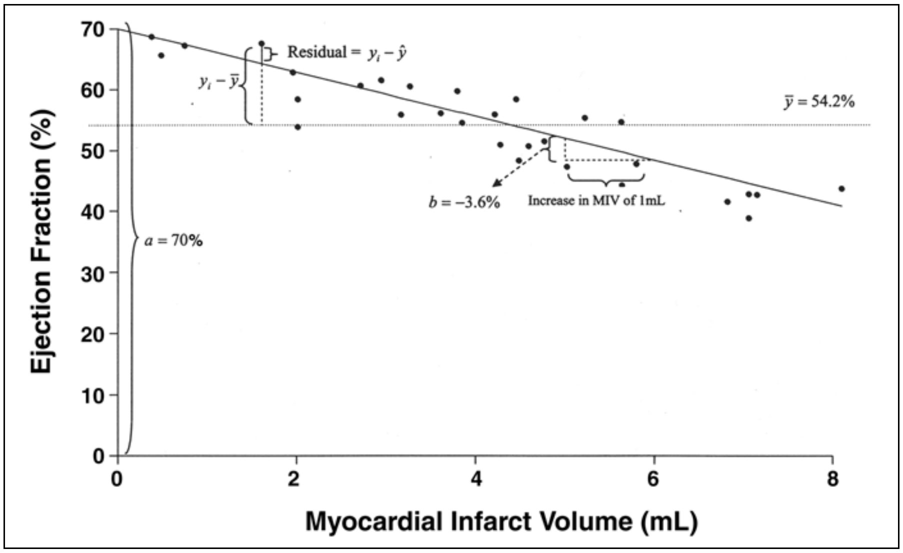


Fig. 4—Graph shows simple linear regression line between ejection fraction and myocardial infarct volume (MIV).

These assumptions imply that for a given value of *X,* the error in predicting the outcome is 0 on the average. Moreover, the magnitude of the error is not associated with *X*.

For our hypothetical example of the relation between myocardial infarct volume and ejection fraction, the estimated simple linear regression equation is as follows:

$$\text{ejection fraction} = 70 - 3.6\ (\text{myocardial infarct volume}) + \varepsilon$$

(see the solid line in Fig. 4).

The intercept of the regression model is equal to the predicted value of the outcome when the predictor variable is 0. This parameter is of interest only in those situations in which 0 lies within the plausible range of *X* values. Figure 4 shows that when the myocardial infarct volume is 0 mL, the ejection fraction is predicted to be equal to the intercept, or 70%. The slope of the regression model is the change in the outcome corresponding to a unit change in the predictor variable. A slope of 0 indicates that no relation exists between the predictor and outcome variables. From Figure 4, we see that when the myocardial infarct volume increases by 1 mL, the predicted value of the ejection fraction decreases by an amount equal to the slope, or –3.6%.

Selecting the "best-fitting" line—We need an objective criterion to help us estimate α and β so that we have a best-fitting straight line. As explained earlier, we would like to use the regression equation to predict the outcome variable using the predictor variable. Clearly, we would like to do so in a way that minimizes the error in prediction (i.e., results in the lowest possible residual), ε_i, for each patient. We use a criterion that minimizes the sum of the squared residual terms:

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

This is known as the method of least squares. The expressions for the estimated values of the intercept and the slope obtained using the method of least squares are given in

$$a = \bar{y} - b\bar{x}$$

where

$$b = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

(See the table in Appendix 2 for an illustrative example of how to calculate *a* and *b* for a smaller sample of five patients. Notice that much of the calculation involves the terms already used in the calculation of Pearson's correlation coefficient.) In addition to *a* and *b,* we also obtain an estimate for the SE (i.e., square root of the variance) of the residuals, which we denote by *s*:

$$s = \sqrt{\frac{\sum_{i=1}^{N} (y_i - a - bx_i)^2}{N - 2}}$$

For our example, the SE of the residuals is given by *s* = 3.53. This tells us that the average error in predicting the ejection fraction by the myocardial infarct volume is about 3.53%. This error is quite small when compared with the range of ejection fraction values—roughly 40–70%—suggesting that our regression equation has a good predictive ability on average.

The residual SE, *s*, can be used to obtain estimates of the SEs of *a* and *b* and of the predicted value of the outcome variable using the formulae given in Appendix 2. These SEs can be used to perform inferences for these parameters via hypothesis tests or confidence intervals. In our example we find that the confidence interval for the slope of the regression line is (−4.3% to −2.9%). Because this interval does not include 0, we can conclude that there is an association between myocardial infarct volume and ejection fraction.

Model diagnostics—After having obtained the intercept and slope of a regression model, we need to verify whether the basic assumptions on which the model was built were satisfied. We need to evaluate whether the residuals follow a normal probability distribution, whether the variance of the residuals is constant for all values of *X*, and whether the relation between *Y* and *X* is linear. All of these assumptions can be verified using the following simple plots of the residuals.

Normal probability plot—A normal probability plot is used to verify whether the residuals follow a normal probability distribution. Most standard statistical software packages can be used to produce this plot. Figure 5A illustrates the ideal situation, in which the residuals do indeed follow a normal distribution

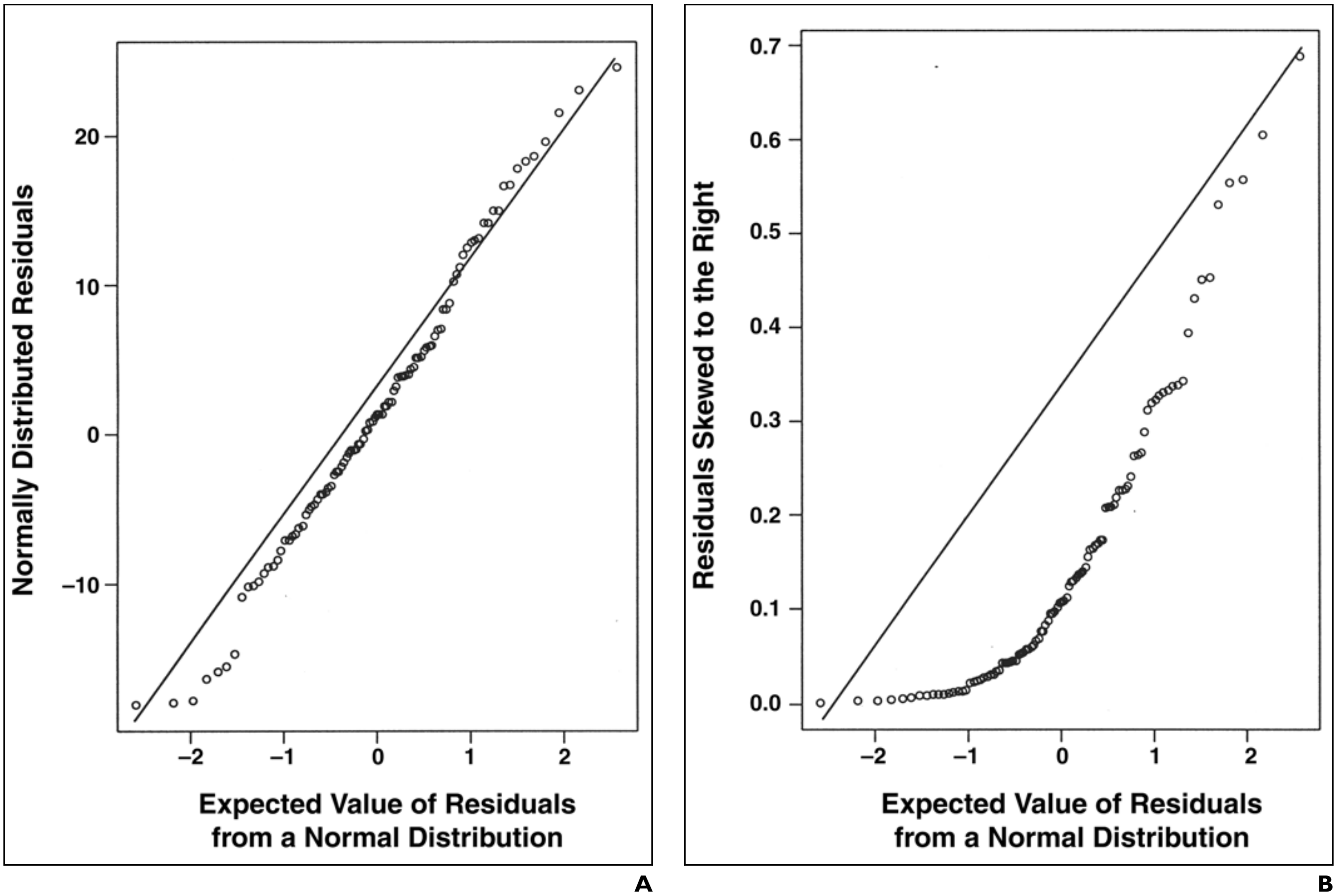


Fig. 5—Prototype normal probability plots.
A and **B,** Graphs show plots with normally distributed residuals (**A**) and with residuals skewed to the right (**B**).

and we observe a straight line along the diagonal of the plot. Any departure of the residuals from a normal distribution will show up as a deviation from this straight line. Figure 5B illustrates a case in which the residuals are skewed to the right and we observe a curved line below the diagonal. A possible corrective measure for this problem is to model the natural logarithm of the outcome instead of the outcome itself.

Scatterplot of residuals versus X—Figures 6A–6C are prototype scatterplots of the residuals versus the predictor variable, *X*. In Figure 6A, we have the ideal situation, in which the model is appropriate. The residuals are randomly scattered about the value of 0 for the entire range of *X*. Furthermore, the residuals fall in a horizontal band of equal width for the entire range of *X*, meaning that they have a constant variance. In Figure 6B, we have a situation in which the residuals indicate that the relation between outcome and predictor is nonlinear. We find that values of *X* that are close to its minimum or maximum are associated with positive residuals, whereas values of *X* in the middle of its range are associated with negative residuals. The parabolic relation between the residuals and *X* in this plot suggests that *Y* is in fact a quadratic function of *X*—that is, *Y* is a function of both *X* and X^2. In Figure 6C, we see an increase in the magnitude of the residuals with increasing *X*. This tells us that our assumption of a constant variance has been violated. As a result, the prediction of the outcome is better for lower values of *X* than for higher values.

Model fit—The usefulness of the regression model is determined by how well it predicts the outcome—that is, how well it fits the data. In the absence of information on myocardial infarct volume, our best guess at predicting the ejection fraction for patients in our sample would have been the sample mean ejection fraction $\bar{y}$—that is, the predicted value of the ejection fraction would be identical for all patients and equal to $\bar{y} = 54.2\%$. This would be equivalent to assuming $a = \bar{y}$ and $b = 0$ (the horizontal dotted line in Fig. 4) and would result in the maximum possible value for the sum of the squared residuals. A commonly used method to estimate the usefulness of a linear regression line is to compare the decrease in the sum of the squared residuals with this maximum value. This is done using the R^2 statistic, which is an estimate of the proportion of the total variation in *Y* that is explained by *X*. The R^2 statistic ranges from a minimum of 0% when *X* is not related to *Y* to 100% when there is a perfect relation between the two variables. In our example, we found that $R^2 = 82.5\%$, meaning that myocardial infarct volume explains 82.5% of the observed variation in the ejection fraction.

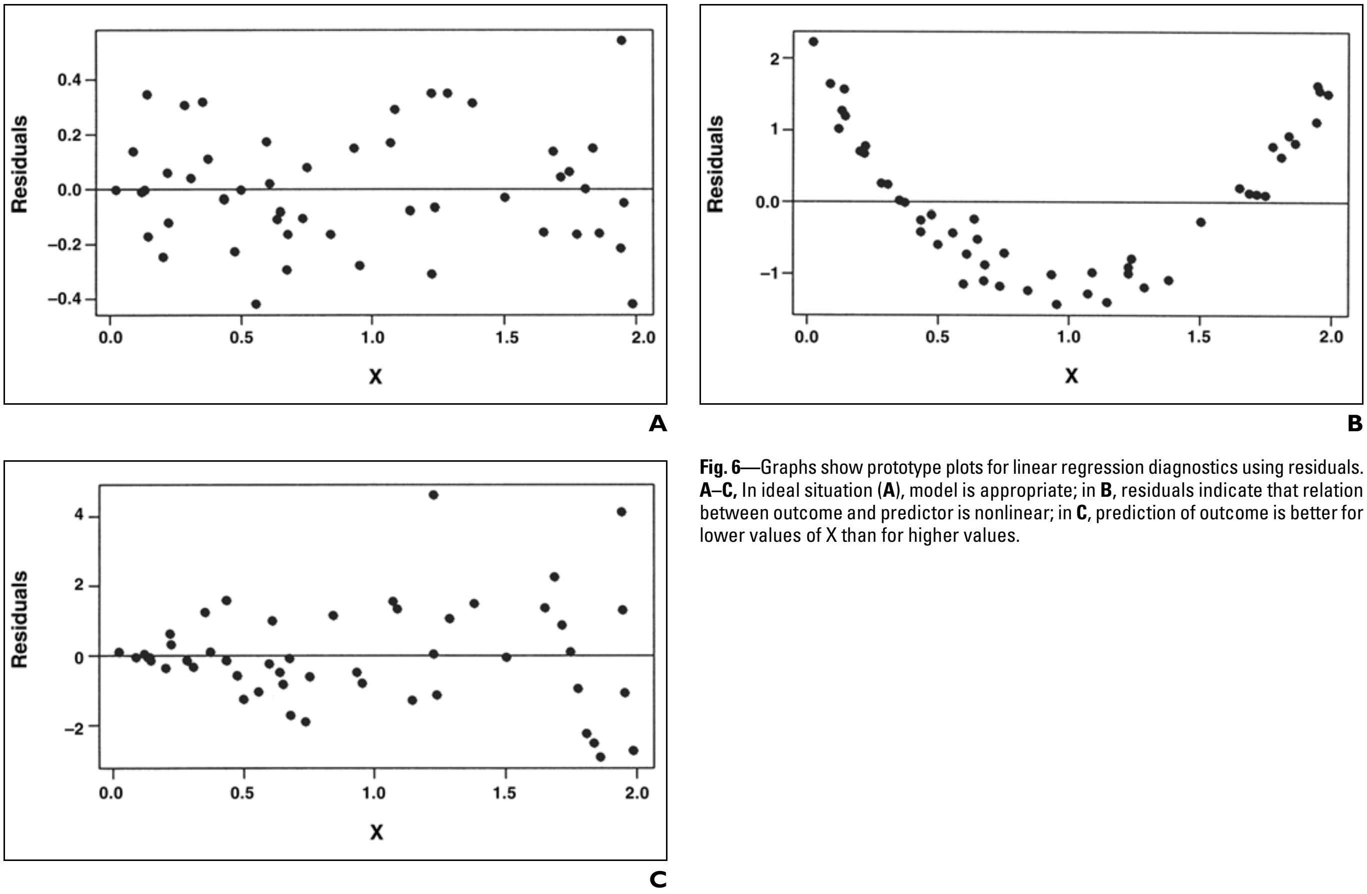


Fig. 6—Graphs show prototype plots for linear regression diagnostics using residuals. **A–C,** In ideal situation (**A**), model is appropriate; in **B**, residuals indicate that relation between outcome and predictor is nonlinear; in **C**, prediction of outcome is better for lower values of X than for higher values.

Multiple Variable Linear Regression

Simple linear regression can be extended to accommodate more than one predictor variable. For example, a patient's glomerular filtration rate (GFR) can be predicted by a linear combination of the patient's age, weight, sex, and the inverse of his or her serum creatinine value by using an equivalent of the form:

$$\text{GFR} = \alpha + \beta_1(\text{age}) + \beta_2(\text{weight}) + \beta_3(\text{sex}) + \beta_4\left(\frac{1}{\text{serum creatinine}}\right) + \varepsilon .$$

As in the case of the simple linear regression model, the unknown parameters α, β_1, β_2, β_3, and β_4 are estimated with the objective of minimizing the sum of the squared residuals (i.e., the sum of the squared differences between the observed GFR values for each patient and the predicted values according to the regression model). We do not present the expressions for calculating the different coefficients and their confidence intervals because these are cumbersome, requiring knowledge of matrix theory. Moreover, most widely available statistical software programs can calculate these quantities. We focus instead on the interpretation of the model.

Table 1 presents the results from a hypothetical study relating the GFR to the predictor variables mentioned here among 100 patients with ages ranging from 40 to 60 years, weight ranging from 40 to 100 kg, and serum creatinine levels between 180 and 200 mmol/L. The intercept is the predicted value of the outcome in the event that all predictor variables are equal to 0. This quantity is of interest only when it is possible for all predictor variables in the model to be simultaneously equal to 0. In the example in Table 1, the intercept is not of interest because the values age = 0, weight = 0, and 1 / serum creatinine = 0 are not possible. The regression coefficients (estimates of the β_1 parameters) corresponding to continuous predictors are interpreted as the change in the outcome variable for a unit change in the predictor variable, while the remaining predictor variables are constant. This means that among a group of patients with a common weight, sex, and serum creatinine, an increase of 1 year in a patient's age is associated with a decrease in the GFR of 0.06 mL/min.

Ordinal and nominal predictor variables—When including nominal predictors (e.g., variables such as sex or country of origin that have no natural ordering) or ordinal predictors (e.g., age measured in 5-year categories) in a regression model, we need to create what are called "dummy variables" or "indicator variables." To do this, we identify one of the categories of the predictor as a reference category. In the case of ordinal variables, the reference category is typically the lowest category. For example, if age is a three-category ordinal variable having values 61–65 years, 66–70 years, and 71–75 years, the 61–65 year category could be selected as the reference. In the case of nominal variables, where there is no clear ordering of the categories, any category may be arbitrarily selected as the reference. Once the reference category has been determined, we create indicator variables corresponding to each of the remaining categories of the predictor. The indicator variables

TABLE 1: Multiple Variable Linear Regression Model for Predicting Glomerular Filtration Rate

Predictor	Estimated Regression Coefficient	SE of Regression Coefficient	*t* Statistic	p^a	95% CI for Regression Coefficient
Intercept	–9.30	17.90	–0.52	0.60	(–44.38 to 25.78)
1 / SCR	9859.28	3194.50	3.09	0.003	(3598.05–16120.51)
Age	–0.06	0.08	–0.75	0.46	(–0.21 to 0.09)
Sex	–2.60	1.06	–2.46	0.02	(–4.67 to –0.53)
Weight	0.07	0.05	1.42	0.16	(–0.03 to 0.17)

Note—CI = confidence interval, SCR = serum creatinine.

[a]Obtained from the tables of the *t*-distribution with $N - k = 100 - 4 = 96$ degrees of freedom.

TABLE 2: Comparing Different Candidate Models for Predicting Glomerular Filtration Rate

Independent Variables in Model	R^2 (%)	Bayesian Information Criterion
1 / SCR	10	–6.22
Age	1	4.02
Sex	10	–5.65
Weight	7	–2.76
1 / SCR + sex + weight	20	–9.44
1 / SCR + sex + weight + age	21	–5.42

Note—SCR = serum creatinine.

take the value of 1 if a patient is in the category to which it corresponds or 0 otherwise. Because three categories were defined for the variable age, this means we need to create two indicator variables—one would take the value 1 for patients in the 66–70 year category, and the second would take the value 1 for patients in the 71–75 year category. Both indicator variables are added to the regression model as predictors.

In the example for GFR, the only noncontinuous predictor is sex. The category "male" was regarded as the reference category. Thus, the variable "sex" is an indicator for the female sex. It takes the value 1 if the patient is female and 0 if the patient is male. The regression coefficient corresponding to sex tells us that after adjusting for the effect of other predictor variables, female patients have a GFR that is 2.60 mL/min lower than that of male patients.

Inference for regression coefficients—Along with regression coefficients, we can report confidence intervals that give an idea of the uncertainty in estimating them. If the confidence interval corresponding to a predictor variable does not include 0, we conclude that it is statistically significant. Alternatively, we could perform a hypothesis test based on the *t* distribution and report a *p* value that tells us the probability of observing our estimated regression coefficient if its true value is 0. If the *p* value is much smaller than a predetermined level of significance (typically 0.05 or 0.01), we reject the null hypothesis that the regression coefficient is equal to 0. If there are *k* parameters in a model, the *p* value is obtained from the tables of the *t* distribution with $N - k$ degrees of freedom (*df*), where *N* is the sample size and *k* is the number of predictors in the regression model. In our example, we can deduce from the 95% confidence intervals that the regression coefficients corresponding to both 1 / serum creatinine and sex are significantly different from 0, and those corresponding to age and weight are not. A similar conclusion is obtained on the basis of the *p* values.

Model fit—The R^2 statistic introduced earlier can also be used to evaluate model fit for multiple variable linear regression models. The R^2 statistic is defined as the proportion of the variance in the outcome variable explained by the regression model. It ranges between 0% and 100%, with values closer to 100% indicating a better model fit. In our example for predicting GFR from age, weight, sex, and serum creatinine level, the R^2 statistic was quite low, meaning that the information obtained explained only 21% of the observed variation in GFR. A low value of R^2 is not unusual in real-life applications.

Model selection—When we have several candidate predictor variables, we are often faced with the challenge of choosing between different models that are based on different predictors. Besides assessing the fit of a model, the R^2 statistic may also be used to compare two different models for the same outcome. Table 2 lists R^2 values for different candidate multiple regression models with GFR as the outcome. The model with the highest value of R^2—that is, the model that best explains the observed variation in GFR—is the model with all four predictor variables included simultaneously. In interpreting these results, it must be noted that the R^2 statistic is influenced by the number of predictor variables in the model. Notice that in Table 2 the R^2 statistic increases with every additional predictor added to the model. Thus, when comparing two models, the R^2 statistic may simply favor the model with the greater number of predictors.

Besides the R^2 statistic, several other criteria have been proposed for model selection. One such criterion is the Bayesian information criterion (BIC). This criterion assesses model fit while simultaneously applying a penalty for every additional predictor added. Our interest is not in the actual value of the BIC for a given model, but rather the difference in the BIC between two models. The lower the BIC, the better the fit of the model. From Table 2, we see that according to the BIC criterion, adding age to the model worsens the model fit. Although criteria such as the R^2 and BIC may be used to assess model fit, the choice of which predictor variables go into a model depends also on their clinical relevance, their impact on the magnitude of regression coefficients associated with the remaining predictors, and their statistical significance.

Model validation—An important way to evaluate a model is to use it to predict the outcome in a data set that is independent of the one used to fit the regression model. This step is referred to as "model validation." Repeating the study to collect new data may not always be a feasible option because of the cost and time involved. Instead, if we have a sufficiently large sample, we may choose to split the data set into two parts—a model-building or training data set that is used to estimate the regression coefficients, and a validation data set. This is known as cross-validation [11]. The model-building data set needs to be sufficiently large to obtain the required precision in estimating the regression coefficients. If this is not possible with half the data, the model-building data set may be larger than the validation data set.

Confounding and effect modification—A multiple linear regression model allows us to study the relation between a primary predictor, *X* (e.g., the experimental treatment), and the outcome, *Y*, while adjusting for the effect of one or more secondary predictor variables (e.g., the patient's demographic characteristics). For illustration, we will consider only one secondary predictor, *Z*, but the concepts discussed here can be extended to the case of more than one secondary predictor. A variable Z is said to be a confounder if it is associated with both X and Y. The true relation between Y and X is not determined by Z. However, not including Z in the regression model results in an incorrect estimate of magnitude or direction of the regression coefficient of X. A variable Z is said to be an effect modifier if it affects the magnitude of the association between Y and X. To determine if Z is an effect modifier, we must add both Z and the product XZ to the regression model between Y and X. It is possible for a variable to be both a confounder and an effect modifier.

The difference between a confounder and an effect modifier is illustrated graphically in Figure 7. In this example, we are interested in studying the relation between the primary predictor variable, weight (kg), and the outcome variable, bone density (mass/volume units). In our hypothetical sample, the patient's sex is a variable that is associated with both the outcome (bone density) and the predictor (weight)—women tend to have a lower bone density and a lower weight than men. Figure 7A illustrates the case when sex is a confounding variable but not an effect modifier. Fitting a single regression line for both men and women that includes only weight as a predictor, we obtain a regression coefficient of 0.4 mass/volume units corresponding to weight. Fitting two separate regression lines—one among men and the other among women—we find that the slope of the two lines is the same and is equal to 0.2 mass/volume units (Fig. 7B). This is the correct value of the slope, which can also be obtained by fitting a single multiple variable regression model in the entire sample that includes both weight and sex as predictors, as follows:

$$\textit{Bone density} = 105 + 0.2\,\textit{weight} - 10\,\textit{sex},$$

where the predictor "sex" is an indicator variable for female sex.

Figure 7C illustrates the case when sex is an effect modifier of the relation between weight and bone density—that is, the strength of the association between weight and bone density is modified by the variable sex. This means the regression lines between bone density and weight among men and women have different slopes (see Fig. 7D). In our hypothetical example, bone density increases more rapidly with weight among women than among men. We can evaluate whether sex is an effect modifier using a single multiple variable regression model that includes weight, sex, and their product as predictors, as follows:

$$\textit{Bone density} = 105 + 0.2\,\textit{weight} - 10\,\textit{sex} - 0.15\,\textit{weight} \times \textit{sex}$$

From this single equation we can determine the different associations between bone density and weight among men and women. By setting sex = 0 in this equation, we find that the regression coefficient associated with

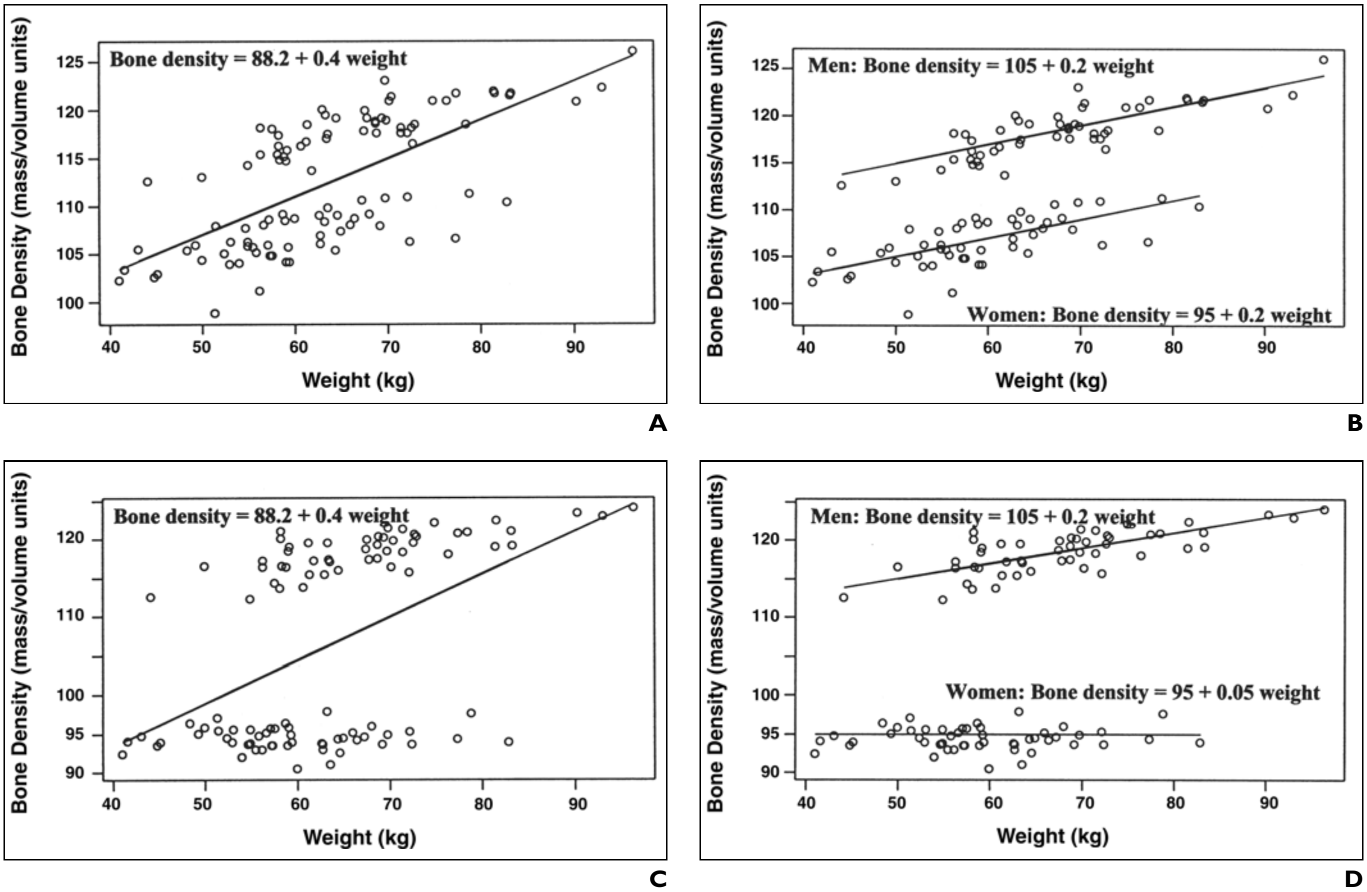


Fig. 7—Confounding and effect modification.
A–D, Graphs illustrate confounding (**A** and **B**) and effect modification (**C** and **D**).

weight is 0.2, the same as was obtained by fitting a separate linear regression model among men. Similarly, when setting sex = 1 in the equation, we find that the regression coefficient associated with weight = 0.2 – 0.15 = 0.05 mass/volume units, which is the same as the regression coefficient obtained when fitting the model among women alone. If the regression coefficient corresponding to the product term is significantly different from 0, we conclude that there is an interaction between weight and sex.

Logistic Regression

Logistic regression, like linear regression, can be used to relate a single outcome variable to one or more predictor variables. However, the outcome variable is dichotomous, having only two values (e.g., success or failure of an experimental treatment, survival or death at the end of a 10-year follow-up). One value of the dichotomous outcome variable must be designated as the outcome of interest—for example, success when the outcome has the values success or failure, or death if the outcome has the values death or survival. The odds of the outcome of interest are given by the ratio of the probability of observing the outcome of interest, to the probability of not observing it: probability of success / probability of failure, or probability of death / probability of survival. The logistic regression equation relates the logarithm of the odds of the outcome to the predictor variables.

In a hypothetical study, logistic regression was used to predict the extremely high breast density on mammography using information on a woman's parity (i.e., number of children), body mass index (BMI), and age. Extremely high breast density was defined as a dichotomous variable taking the value 1 when a woman's breast density was greater than or equal to 75%, and taking the value 0 when a woman's breast density was less than 75%. The resulting multiple logistic regression equation had the following form:

$$ln(\frac{\text{Probability of EHBD}}{1\text{- Probability of EHBD}})=$$
$$\alpha+\beta_1(nulliparous)+\beta_2(BMI)+\beta_3(age),$$

where *ln* is the logarithm to the natural base *e* and EHBD is extremely high breast density.

The predictor variables in a logistic regression equation may be continuous, nominal, or ordinal. As in the case of multiple linear regression, nominal and ordinal predictor variables are entered into the equation as indicator variables. In the logistic regression equation for extremely high breast density, BMI and age are both continuous variables, and nulliparous is an indicator that the woman is nulliparous.

The best estimates for the unknown parameters α, β_1, β_2, and β_3 may be obtained by a statistical method known as maximum likelihood. This method helps us identify the most likely value of the true parameters given the observed data and under the assumption that the number of patients with the outcome of interest follows a binomial distribution [2].

The relation between each predictor variable and the outcome in a logistic regression model is expressed in terms of an odds ratio (for more about odds ratios see the article by Blackmore and Cummings [4] in this series). When the predictor variable is ordinal or nominal, the odds ratio is a comparison between each indicator variable and the reference category. An odds ratio of 1 indicates there is no difference in the odds of the outcome of interest between the category associated with the indicator variable and the reference category. An odds ratio greater (lesser) than 1 indicates the outcome of interest is more (less) likely in the category associated with the indicator variable than in the reference category. Results for the extremely high breast density example are given in Table 3. The odds ratio of 5.53 corresponding to nulliparous tells us that the odds of extremely high breast density are (5.53 − 1) × 100 = 453% greater among women who are nulliparous compared with those who are not. For a continuous predictor variable, the odds ratio gives the relative increase (or decrease) in the odds of the outcome for a change of one unit of the predictor variable. For example, in Table 3, the odds ratio of 0.85 corresponding to BMI means that for a unit increase in the BMI, a woman's odds of extremely high breast density decrease by (1 – 0.85) × 100 = 15%. The odds ratios for all predictor variables are obtained by taking the exponent of the regression coefficient.

We can test whether each regression coefficient is different from 0 using a chi-square test with *N* – *k df*, where *N* is the sample size and *k* is the number of predictors in the regression model. By comparing the chi-square *p* values in Table 3 with the traditional level of significance of the null hypothesis of α = 0.05, we conclude that the predictors nulliparous and BMI are statistically significantly associated with an extremely high breast density. Alternatively, we can report a confidence interval for the odds ratio. If the confidence interval does not include 1, then the predictor is considered statistically significant. If the confidence interval includes 1, as in the case of the predictor age in Table 3, we conclude that it is not significantly associated with the outcome.

As in linear regression a logistic regression model can also be used to determine whether a particular predictor variable is a confounder or effect modifier. The fit of a logistic regression model may be assessed using the BIC or a statistic similar to the R^2 statistic.

Sample Size Determination

Any well-designed research study must begin with an idea of the sample size required. An insufficient sample size might leave us with important questions unanswered. On the other hand, too large a sample size might mean an unnecessarily expensive study. The sample size required for a study is calculated so that it provides sufficient evidence to make inferences about the primary parameter(s) of interest in the study. As mentioned throughout this series, there is an increasing emphasis in scientific journals on reporting of confidence intervals rather than *p* values. Thus, for this article we will limit ourselves to sample size formulae that are suitable for studies having the objec-

TABLE 3: Logistic Regression Model for Predicting Extremely High Breast Density

Predictor	Estimated Regression Coefficient	SE of Regression Coefficient	Chi-Square Value (*p*)[a]	Odds Ratio	95% CI for Odds Ratio
Nulliparous					
No (reference)	0	—	—	1	—
Yes	1.71	0.62	7.61 (0.006)	exp(1.71) = 5.53	(1.64, 20)
Body mass index	–0.16	0.07	5.22 (0.023)	exp(–0.16) = 0.85	(0.73, 0.98)
Age	–0.02	0.04	0.25 (0.599)	exp(–0.02) = 0.98	(0.89, 1.07)

Note—CI = confidence interval. Dash (—) indicates not applicable.

[a]Obtained by comparison with chi-square distribution with *N* – *k* = 102 – 3 = 99 degrees of freedom.

tive of reporting a confidence interval for the primary parameters of interest. Furthermore, we focus on sample size calculations for Pearson's correlation coefficient and simple linear regression. Sample size formulae for multiple variable linear regression and logistic regression are available but involve complex methods and are typically implemented by software programs [14]. These programs also provide calculations for studies in which the primary objective is to test a null hypothesis and report a *p* value.

General Concepts for Sample Size Calculation

Whatever the parameter of interest, certain concepts remain common to the exercise of sample size calculation.

First, the sample size calculation requires a guess value for the parameter of interest (e.g., correlation coefficient or the slope of a regression model) and parameters of its probability distribution (e.g., SE of the slope). This is rather paradoxical because the goal of the study is to find out more about this parameter. However, some reasonable range of guess values for the parameter can usually be found from the literature.

Second, identify a clinically meaningful range of values for this parameter.

Sample Size for Pearson's Correlation Coefficient

Assume we want to perform a study the goal of which is to measure the correlation between ratings of two experienced radiologists on a series of mammograms. Based on an earlier pilot study, our guess value for the correlation coefficient is $\rho_P = 0.85$. A sufficiently high correlation is deemed to be in the order of 0.8–0.9. Any value less than this is considered poor correlation. Ideally, we would like our research study to unequivocally determine whether the true correlation between the reviewers is sufficiently high. This means we would like our sample size to be large enough to ensure that the confidence interval lies entirely within or below the range 0.8–0.9—that is, the half-width of the confidence interval (or precision of our estimate) should be a maximum of 0.85 – 0.8 = 0.9 – 0.85 = 0.05. The calculation of the confidence interval requires the transformation of the correlation coefficient, ρ_P, into

$$Z_P = \frac{1}{2}\ln\left[\frac{1+\rho_P}{1-\rho_P}\right]$$

(see Appendix 1). Therefore, we need to determine the maximum permissible value of the confidence interval half-width on the transformed scale. To do this, we transform both the guess value of the correlation coefficient and the lower end of the confidence interval and calculate their difference. The maximum permissible half-width of the transformed confidence interval, is given by

$$w_z = \frac{1}{2}\ln\left[\frac{1+0.85}{1-0.85}\right] - \frac{1}{2}\ln\left[\frac{1+0.8}{1-0.8}\right] = 1.26 - 1.10 = 0.16.$$

The sample size required to obtain a $(1-\alpha)$% confidence interval is then calculated as

$$N = \left(\frac{Z_{1-\alpha/2}}{w_Z}\right)^2,$$

where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. Thus, to obtain a 95% confidence interval for our study, we would need a sample size of approximately

$$N = \left(\frac{Z_{1-\alpha/2}}{w_Z}\right)^2 = \left(\frac{1.96}{0.16}\right)^2 = 150.$$

Sample Size for the Slope of a Simple Linear Regression Model

Sample size calculation for the simple linear regression model typically focuses on determining whether the slope is different from 0. The required sample size can be obtained using the same approach as that given in this article for the correlation coefficient, by exploiting the fact that a slope of 0 in a simple linear regression equation is equivalent to a correlation of 0 between the predictor and outcome variables. Suppose we plan to study the relation between renal length as measured by sonography (predictor) and GFR (outcome) via simple linear regression. Suppose also that a smaller pilot study of the relation between these variables had reported a correlation coefficient of 0.3 (–0.2 to 0.8). To conclusively show a relation between the two variables, we would like the confidence interval to lie within 0.1–0.5 (i.e., to eliminate 0). The required sample size can be calculated using the methods described earlier for Pearson's correlation coefficient.

Conclusion

This article describes some of the most common statistical methods used by radiologists to evaluate the relation between variables. The article stresses the interpretation of these statistics and describes formulae to implement some of the simpler methods. Although it is unlikely that readers will actually perform these calculations by hand because they are all available in standard statistical packages, our aim in discussing them is to give the interested reader a better understanding of the motivation behind the statistical methods. Because of limited space we can only scratch the surface of many of the topics under regression models. More details on the topics discussed here may be found in introductory [7–9] and advanced [10–13] textbooks.

References

1. Karlik SJ. Exploring and summarizing radiologic data. *AJR* 2003; 180:47–54
2. Joseph L, Reinhold C. Introduction to probability theory and sampling distributions. *AJR* 2003; 180:917–923
3. Joseph L, Reinhold C. Statistical inference for continuous variables. AJR 2005; 184:1047–1056
4. Blackmore C, Cummings P. Observational studies in radiology. *AJR* 2004; 183:1203–1208
5. Hennekens CH, Buring E. *Epidemiology in medicine*. Boston, MA: Lippincott, Williams & Wilkins, 1987
6. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1:307–310
7. Moore DS, McCabe GP. *Introduction to the practice of statistics*, 3rd ed. New York, NY: Freeman, 1998
8. Glantz SA. *Primer of biostatistics*, 5th ed. New York, NY: McGraw-Hill, 2001
9. Dawson B, Trapp RG. *Basic and clinical biostatistics,* 3rd ed. New York, NY: McGraw-Hill Lange Medical Series, 2001
10. Harrell F. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, 1st ed. New York, NY: Springer-Verlag, 2001
11. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied regression analysis and multivariable methods*, 3rd ed. Pacific Grove, CA: Duxbury Press, 1998
12. Hosmer D, Lemeshow S. *Applied logistic regression*, 2nd ed. New York, NY: Wiley, 2000
13. Kleinbaum DG. *Logistic regression: a self-learning text*, 2nd ed. New York, NY: Springer-Verlag, 2002
14. Hintze JL. PASS [power analysis and sample size] user's guide. Kaysville, UT: NCSS [Number Cruncher Statistical System], 1996

Appendix 1 appears on the next page.

APPENDIX 1. Inference for Pearson's Correlation Coefficient (r_P)

p value

To calculate the *p* value, we need to transform r_P as follows:

$$Z_P = \frac{1}{2}\ln\left[\frac{1+r_p}{1-r_p}\right]$$

where *ln* is the natural logarithm. This transformation is required because even though *X* and *Y* may follow a normal distribution, r_P does not. However, Z_P is known to follow a normal distribution with a standard deviation

$$\sigma_Z = \sqrt{\frac{1}{n-3}}$$

making the calculation of the *p* value and confidence intervals easier. The remaining steps involved in calculating a *p* value are explained in the box below.

Compute the test statistic

$$z = \frac{Z_P - Z_0}{\sigma_Z}$$

The rule for estimating the *p* value depends on the alternative hypothesis H_A as follows (see [3] for more on hypothesis testing):

When $H_A : \rho_P > \rho_0$, the *p* value is given by the probability $P(Z \geq z)$.

When $H_A : \rho_P < \rho_0$, the *p* value is given by the probability $P(Z \leq z)$.

When $H_A : \rho_P \neq \rho_0$, the *p* value is given by the probability $P(Z \geq |z|)$.

The *p* value is calculated by comparing the test statistic with the tables of the normal distribution. Typically, if the *p* value is less than a predetermined level of significance, such as 0.05 or 0.01, the null hypothesis is rejected in favor of the alternative.

Recall that in our example of myocardial infarct volume and ejection fraction, the correlation coefficient for the entire sample of $n = 30$ patients was $r_P = –0.91$. To estimate the evidence in favor of the hypothesis "there is no relation between myocardial infarct volume and ejection fraction"—that is, H_0: $\rho_P = 0$—we begin by calculating the test statistic. First transform r_P into

$$Z_P = \frac{1}{2}\ln\left[\frac{1+r_p}{1-r_p}\right] = \frac{1}{2}\ln\left[\frac{1+(-0.91)}{1-(-0.91)}\right] = \frac{1}{2}\ln\left[\frac{1-0.91}{1+0.91}\right] = -1.53$$

Then transform ρ_0 into

$$Z_0 = \frac{1}{2}\ln\left[\frac{1+\rho_0}{1-\rho_0}\right] = \frac{1}{2}\ln\left[\frac{1+0}{1-0}\right] = 0$$

Finally, calculate the SD of Z_P as

$$\sigma_Z = \sqrt{\frac{1}{N-3}} = \sqrt{\frac{1}{30-3}} = 0.19$$

Using these three quantities, the test statistic can now be calculated as $z = (Z_P - Z_0) / \sigma_Z = (-1.53 - 0) / 0.19 = -8.05$. The evidence in favor of the null hypothesis against an alternative hypothesis of "there is a relation between myocardial infarct volume and ejection fraction"—that is, H_A: $\rho_P \neq 0$ is equal to $P(Z \geq |-8.05|)$. This is the probability that a variable following a standard normal distribution is less than –8.05 or greater than 8.05. From the normal distribution tables, we find that this probability is less than 0.0001. See module 10 in this series [2] for an explanation of how to use the tables of the normal distribution.

Confidence interval

As in the case of the *p* value, to construct a confidence interval for ρ_P we first need to transform r_P into Z_P. The upper (u_Z) and lower (l_Z) limits of the $(1-\alpha)\%$ confidence interval on the transformed scale are given by ($l_Z = Z_P - Z_{1-\alpha/2}\ \sigma_Z$, $u_Z = Z_P + Z_{1-\alpha/2}\ \sigma_Z$), where σ_Z is the previously defined SD of Z_P, and $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. The latter is the point below which the area under the normal distribution curve is equal to $1 - \alpha / 2$. We then retransform these limits to obtain the $(1-\alpha)\%$ confidence interval for ρ_P as ($l = [\exp(2l_Z) - 1] / [\exp(2l_Z) + 1]$, $u = [\exp(2u_Z) - 1] / [\exp(2u_Z) + 1]$). In our example of myocardial infarct volume and ejection fraction, we can use the previously calculated values of Z_P and σ_Z to obtain a 95% confidence interval on the transformed scale as ($l_Z = -1.53 - 1.96[0.19]$, $u_Z = -1.53 + 1.96[0.19]$) = (–1.90 to –1.16). The value $Z_{1-\alpha/2} = 1.96$ is obtained from the normal distribution table. On retransformation, we obtain the limits of the 95% confidence interval for ρ_P as

$$l = \frac{\exp(2l_z) - 1}{\exp(2l_z) + 1} = \frac{\exp(2(-1.90)) - 1}{\exp(2(-1.90)) + 1} = \frac{0.02 - 1}{0.02 + 1} = -0.96,$$

$$u = \frac{\exp(2lu_z) - 1}{\exp(2lu_z) + 1} = \frac{\exp(2(-1.16)) - 1}{\exp(2(-1.16)) + 1} = \frac{0.1 - 1}{0.1 + 1} = -0.81.$$

Appendix 2 appears on the next page.

APPENDIX 2. Inference for the Simple Linear Regression Model

Standard errors (SEs) for the intercept and slope of the simple linear regression model, and expressions for calculating the *p* value and confidence interval for these parameters are given in the box below:

	Intercept	Slope
Standard error:	$s_a = s\sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$	$s_b = \frac{s}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$
p value:	Null hypothesis: H_0: $\alpha = \alpha_0$ Test statistic: $t_a = \frac{a - \alpha_0}{s_a}$ when H_A: $\alpha \neq \alpha_0$ then *p* value = $P(t_{N-2} \geq \lvert t_a \rvert)$	Null hypothesis: H_0: $\beta = \beta_0$ Test statistic: $t_b = \frac{b - \beta_0}{s_b}$ when H_A: $\beta \neq \beta_0$ then *p* value = $P(t_{N-2} \geq \lvert t_b \rvert)$
	(where t_{N-2} denotes a standard *t* distribution with N–2 degrees of freedom)	
$(1-\alpha)$% confidence interval:	$a \pm t_{1-\alpha/2,N-2} \times s_a$	$b \pm t_{1-\alpha/2,\,N-2} \times s_b$
	(where $t_{1-\alpha/2,N-2}$ denotes the $(1-\alpha/2)$% quantile of the standard *t* distribution with N–2 degrees of freedom)	

Typically, we are more interested in the slope than in the intercept. A natural null hypothesis of interest is H_0: $\beta = 0$. The SE of the slope in our example is given by $s_b = 0.32$. See the table in this appendix for an illustration of how to calculate s_b in a smaller sample of five patients. Note the results there are slightly different from those in this section because they are based on a different sample. Using the formula in the box, the test statistic can be calculated as

$$t_b = \frac{b}{s_b} = \frac{-3.6}{0.32} = -11.5$$

As in the case of the correlation coefficient, the *p* value that we report depends on the direction of the alternative hypothesis. If the alternative hypothesis was H_A: $\beta \neq 0$, then the *p* value is given by $P(t_{N-2} \geq \lvert t_b \rvert)$—that is, the probability that the standard *t* distribution with $N - 2 = 28$ degrees of freedom (*df*) takes values less than or equal to $-\lvert t_b \rvert = -11.38$ or greater than or equal to $\lvert t_b \rvert = 11.38$. (Recall N = our sample size of 30. See [3] more details on the *t* distribution.) Looking up the *t* distribution tables corresponding to $N - 2 = 30 - 2 = 28$ *df*, we find that this probability is less than 0.001. Because this probability is much less than the traditional significance levels of 0.05 or 0.01, we reject the null hypothesis and conclude that there is a relation between ejection fraction and myocardial infarct volume.

Alternatively, we could construct a 95% confidence interval for the slope. As mentioned previously, this is more informative than simply reporting whether we did or did not reject a single null hypothesis. The term "$t_{1-\alpha/2,\,N-2}$" in the formula above denotes the $1-\alpha/2$ quantile of the *t* distribution with 28 *df* (i.e., the point on the standard *t* distribution below which there is a $1-\alpha/2$ probability). For a 95% confidence interval, we have $\alpha = 1 - 0.95 = 0.5$. The value of $t_{1-\alpha/2,\,N-2} = t_{0.975,28} = 2.05$. For our example, we have already calculated $b = -3.6\%$ and $s_b = 0.32$. Thus, the 95% confidence interval is given by

$$(b - t_{0.975,28} \times s_b \text{ to } b + t_{0.975,28} \times s_b)$$

$$= (-3.6 - 2.05 \times 0.32 \text{ to } -3.6 + 2.05 \times 0.32)$$

$$= (-4.3\% \text{ to } -2.9\%).$$

This interval gives us an idea of the range of values of the slope that is compatible with the data and cannot be rejected by a hypothesis test. Because the interval does not include 0, we can conclude that there is a negative relation between ejection fraction and myocardial infarct volume.

For a given value of myocardial infarct volume, our simple linear regression model may also be used to predict the ejection fraction for an average patient or to predict the ejection fraction for an individual patient. The SEs for the predicted mean ejection fraction and for an individual's ejection fraction are as follows:

SE for predicted mean outcome at x

$$s_{Mx} = s\sqrt{\frac{1}{N} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

SE for predicted individual outcome at x

$$s_{Ix} = s\sqrt{1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

TABLE: Calculating Pearson's Correlation Coefficient and the Simple Regression Equation Between Myocardial Infarct Volume and Ejection Fraction

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Patient number *(i)*	x_i (Myocardial infarct volume, mL)	y_i (Ejection fraction, %)	x_i- $\bar{x}$	y_i- $\bar{y}$	$(x_i$- $\bar{x}$ $)^2$	$(y_i$- $\bar{y}$ $)^2$	$(x_i$- $\bar{x}$ $)(y_i$- $\bar{y}$ $)$	$y_i - a - bx_i$	$(y_i - a - bx_i)^2$
1	2.5	65	-2.5	15	6.25	225	-37.5	2	4
2	3.75	55	-1.25	5	1.5625	25	-6.25	-1.5	2.25
3	5	50	0	0	0	0	0	0	0
4	6.25	40	1.25	-10	1.5625	100	-12.5	-3.5	12.25
5	7.5	40	2.5	-10	6.25	100	-25	3	9
	$\bar{x} = 5$	$\bar{y}$ =50			$\sum(x-\bar{x})^2$ = 15.625	$\sum(y-\bar{y})^2$ = 450	$\sum(x-\bar{x})(y-\bar{y})$ = -81.25		$\sum(y_i - a - bx_i)^2$ = 27.5

$$r_P = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}} = \frac{-81.25}{\sqrt{15.625\times 450}} = -0.97$$

$$b = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} = \frac{-81.25}{15.625} = -5.2$$

$$a = \bar{y} - b\bar{x} = 50 - (-5.2)(5) = 76$$

$$s^2 = \frac{\text{Residual sum of squares}}{N-2} = \frac{\sum_{i=1}^{N}(y_i - a - bx_i)^2}{N-2} = \frac{27.5}{3} = 9.17$$

Therefore, s = $\sqrt{9.17}$ = 3.03

$$s_a = s\sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}} = 3.03\sqrt{\frac{1}{5}+\frac{5^2}{15.625}} = 4.06$$

$$s_b = \frac{s}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}} = \frac{3.03}{\sqrt{250}} = 0.77$$

$$s_{M,2} = s\sqrt{\frac{1}{N} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}} = 3.03\sqrt{\frac{1}{5}+\frac{(2-5)^2}{15.625}} = 2.67$$

$$s_{I,2} = s\sqrt{1+\frac{1}{N} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}} = 3.03\sqrt{1+\frac{1}{5}+\frac{(2-5)^2}{15.625}} = 4.04$$

Note—For ease of illustration, we limit the sample to five patients. These results are slightly different from those reported in the text because they are based on a different sample. The mean myocardial infarct volume among these five patients is 5 mL, and the mean ejection fraction is 50%. First, subtract the mean infarct volume from each patient's infarct volume (see the column $x_i - \bar{x}$). For example, for patient 2, we have $x_i - \bar{x}$ = 3.75 – 5 = –1.25. Then take the square of this value for each patient (see the column $(x_i - \bar{x})^2$). For patient 2, this would be -1.25^2 = 1.5625. Do the same for ejection fraction. Finally, for each patient multiply $x_i - \bar{x}$ and $y_i - \bar{y}$. For patient 2, this is –1.25 × 5 = –6.25. In each of columns 6, 7, and 8 above, add the values across all patients. The correlation coefficient can then be calculated from the resulting sums. The slope and intercept of the regression model are calculated using columns 2, 3, 6, and 8. In column 10 we have the sum of the squared residuals across patients. This is used in the calculation of the SEs for the slope, intercept, predicted average ejection fraction, and predicted individual ejection fraction.

Notice that these two SEs are very similar except for the fact that an additional 1 appears in the term under the square root for the SE of the predicted outcome for an individual. This causes the SE of the predicted outcome for a single individual to always be greater than the predicted outcome for an average individual. This is because of the additional variance of the individual outcomes above the average outcome. In our example, $S_{M,2}$ = 1.14, and $S_{I,2}$ = 3.71. The predicted value of the outcome when the predictor is equal to x is denoted by $\hat{y}_x$. The predicted average ejection fraction corresponding to a myocardial infarct volume of 2 mL (denoted by $\hat{y}_2$) can be calculated using the regression equation as 70 – 3.6(2) = 62.8%. The expression for a (1–α)% confidence interval for the average ejection fraction is

$$\hat{y}_x - t_{1-\alpha/2,N-2} \times s_{M,x},\ \hat{y}_x + t_{1-\alpha/2,N-2} \times s_{M,\,x}$$

Recall that we had determined from the tables of the t distribution that $t_{0.975,28}$ is 2.05. Thus, the 95% confidence interval for the predicted mean ejection fraction when myocardial infarction volume = 2 mL is given by

$$= (\hat{y}_2 - t_{0.975,\,28} \times s_{M,\,2}, \hat{y}_2 + t_{0.025,\,28} \times s_{M,\,2}$$

$$= (62.8 - 2.05 \times 1.14 \text{ to } 23.6 + 2.05 \times 1.14)$$

$$= (60.5\% \text{ to } 65.1\%)$$

The confidence interval for an individual's ejection fraction when myocardial infarction volume is 2 mL is obtained by replacing the SE in the this expression by $s_{I,x}$—that is, by

$$= (\hat{y}_2 - t_{0.975,\,28} \times s_{I,\,2}, \hat{y}_2 + t_{0.975,\,28} \times s_{I,\,2})$$

$$= (62.8 - 2.05 \times 3.71 \text{ to } 23.6 + 2.05 \times 3.71)$$

$$= (55.2\% \text{ to } 70.4\%)$$

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005
16. Statistical Inference for Proportions, April 2005
17. Reader Agreement Studies, May 2005

Survival Analysis

Harald O. Stolberg[1,2]
Geoffrey Norman[3]
Isabelle Trop[4]

Received November 17, 2004; accepted November 23, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 19th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology.* The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Department of Radiology, McMaster University Medical Centre, 1200 Main St. W, Hamilton, ON, L8N 3Z5 Canada.

[2]Deceased.

[3]Department of Educational Research, Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada.

[4]Department of Radiology, Hospital St.-Luc, 1058 St-Denis St., Montreal, QC, H2X 3J4 Canada. Address correspondence to I. Trop.

AJR 2005;185:19–22

0361–803X/05/1851–19

The breadth of radiology research is expanding. Previously, a large proportion of radiology research projects were observational studies. Increasingly, research now involves groups of patients to whom specific interventions are administered in a randomized fashion. Analysis of data obtained from these experimental studies varies, depending on the end point of interest. Research protocols that are designed to evaluate the interval between entry of a patient into the study and the time until the event of interest are referred to as time-to-event studies, a form of follow-up study [1]. The event may be death in a diagnostic study of cancer or a progression of various chronic disease entities to a defined stage. In interventional studies, such as vascular and neuroradiologic procedures, the fate of grafts, stents, and other devices may be followed through time. Survival analysis, also called "life table" analysis, refers to the methodology of analysis of data gathered in such protocols. Survival analysis, then, is the topic of this article [2].

Overview

Under ideal circumstances, a study would enroll all its subjects simultaneously and follow them either for a fixed period of time or until they all reach some end point, such as recovery or death. However, more commonly, studies require a large number of subjects or look at relatively rare conditions, and so must enter subjects over a period of several months or even years. When the study finally ends, the subjects will have been followed for varying lengths of time, during which a number of different outcomes have to be considered: the event has not yet occurred (outcome A), some patients are lost to follow-up (outcome L), or the event has occurred (an example of the event or end point is death) (outcome D).

Figure 1 shows how we can illustrate these different outcomes, indicating what happened to the first 10 patients in a study. Subjects A, C, D, and F died during the trial; they are labeled "D" for dead. Subjects B, G, and I were lost to follow-up, hence the label "L," at various times after they started the drug. The other subjects, E, H, and J (labeled "C"), were still alive at the time the trial ended. These last three data points are called "right-censored." Subjects are considered "censored" when their data are incomplete. They are said to be right-censored because they have been followed to the end of the study (the "right-hand part" of the graph), but the outcome of interest has not occurred to them. To be more quantitative about the data, Table 1 shows how long each person was in the study and what the outcome was.

The Kaplan-Meier Approach to Survival Analysis

To do a survival analysis, we must figure out how many people survive for at least 1 year, for at least 2 years, and so on, in what

TABLE 1: Outcomes of the First 10 Subjects

Subject	Length of Time in Trial (months)	Outcome
A	61	Died
B	111	Lost
C	29	Died
D	46	Died
E	92	Censored
F	22	Died
G	37	Lost
H	76	Censored
I	14	Lost
J	45	Censored

Note—Reprinted with permission from [4].

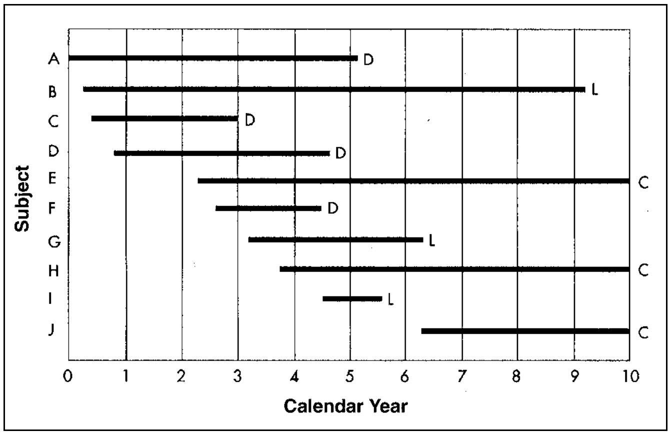


Fig. 1—Entry and withdrawal of subjects in a 10-year study. (Reprinted with permission from [4])

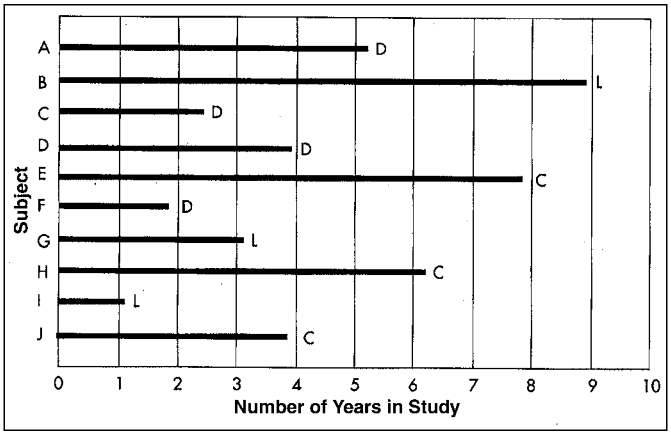


Fig. 2—Figure 1 redrawn so all subjects have a common starting date. (Reprinted with permission from [4])

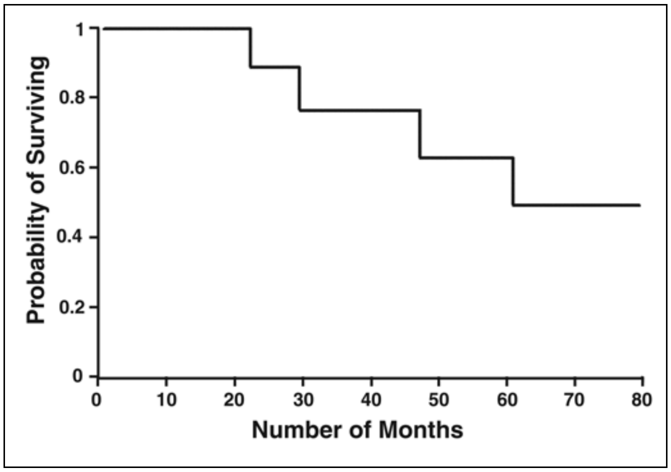


Fig. 3—Survival curve for data in Table 2.

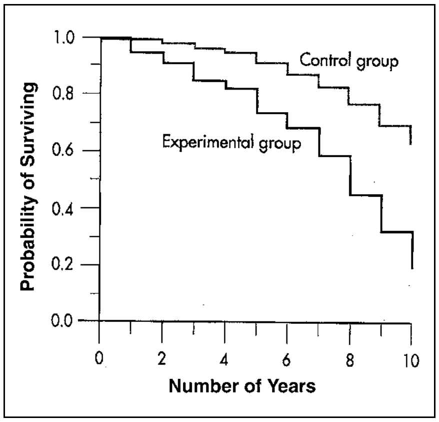


Fig. 4—Survival curves for both groups in study of patients with intramural hematoma of the aorta. (Reprinted with permission from [4])

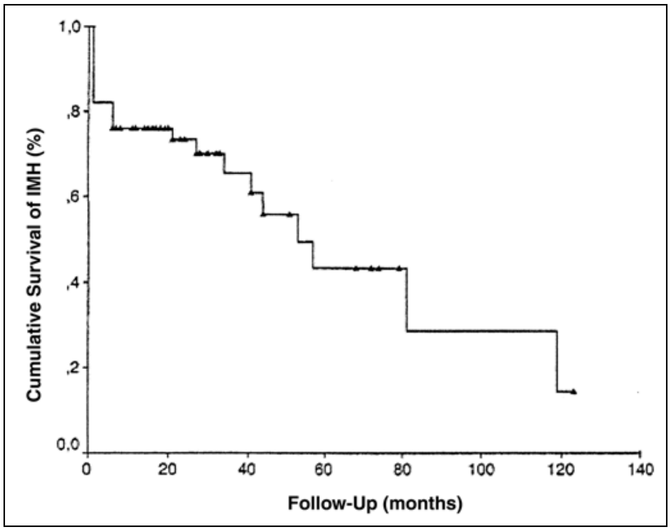


Fig. 5—Probability of survival after aortic intramural hematoma (IMH) in 66 study patients. Small triangles indicate censored cases.

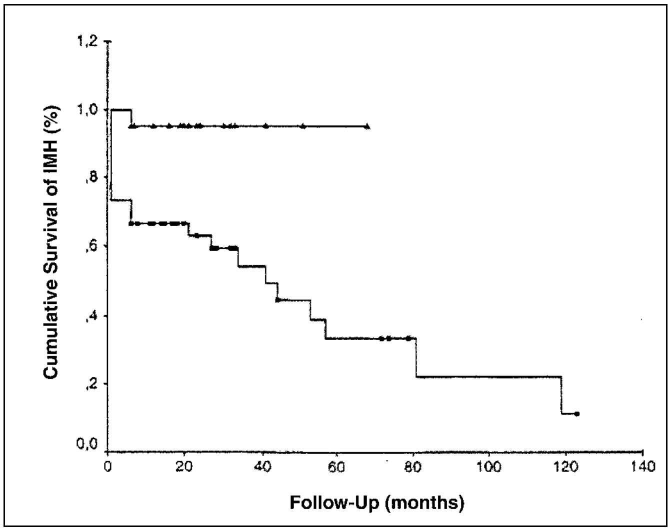


Fig. 6—Cumulative survival of patients with intramural hematoma (IMH) with (experimental group) and without (control group) treatment with β-blockers. Upper curve (*triangles*) indicates treated patients; small squares indicate censored cases. Difference between two subgroups was statistically significant ($p = 0.004$).

is called a “life table” technique. There are two ways to go about calculating a life table: the actuarial approach and the Kaplan-Meier approach [3]. The Kaplan-Meier approach is far more common in medical literature, so we will describe it.

The first step involves redrawing the graph, so that all the people appear to start at the same time. Figure 2 shows the same data as Figure 1; however, instead of the *x*-axis being Calendar Year, it is now Number of Years in Study. The lines are all the same length as in Figure 1; they have just been shifted to the left so that they all begin at time 0.

The Kaplan-Meier approach uses the exact time of death in the calculation of survival. It also computes the survival function only when an outcome occurs. To show how this is done, let us use the data for the 10 subjects in Table 1. The first step is to rank-order the length of time in the trial and flag which entries reflect the outcome of interest (death in this case) and which are due to withdrawal or censoring. We have done this by putting an asterisk after the data for subjects who were lost to follow-up or were censored by the termination of the study:

14* 22 29 37* 45* 46 61 76* 92* 111*

This data set would generate a life table (Table 2) with only four rows, one for each of the four patients who died.

One person was lost to follow-up before the first person died, so the number of remaining patients at risk at 22 months is only nine. Death rate, survival rate, or any other statistical estimate is calculated on the basis of the population at risk (Table 2). At 46 months, two people had died and three were lost to follow-up, so the number of patients at risk is five, and so on. This little data set would generate a survival curve like that shown in Figure 3 except for fewer steps.

Comparing Two (or More) Groups with the Log-Rank Test

Although the survival curve shown in Figure 3 tells us what happened to patients over time, we often want to compare two or more groups of patients—for example, patients with different kinds of stents, or patients who were screened (experimental group) versus patients who were not screened (control group). So we will create an expanded survival table with 250 experimental subjects and 250 control patients. These data are presented in Figure 4. This graph shows that the survival curve for the treatment group dropped at a faster rate than that for the control group. But is the difference statistically significant?

The best approach for evaluating whether the difference is indeed significant is to use the Mantel-Cox log-rank test, which is a modification of the Mantel-Haenszel chi-square test [4]. This test is a powerful method for analyzing data when the time to the outcome is important; it deals with censored data and differential length of follow-up of different subjects. As with most chi-square tests, the log-rank test compares the observed number of events with the number expected, under the assumption that the null hypothesis of no group differences is true. That is, if there were no differences between the groups, then at any interval, the total number of events should be divided between the groups roughly in proportion to the number of subjects at risk. The test determines how much the observed event rate differs from the expected rate.

The Cox Proportional Hazards Model

A more sophisticated method of analysis commonly used, which examines the difference in the survival curves while also accounting for other variables (covariates), is the Cox proportional hazards model [5]. Unlike the log-rank test, the proportional hazards model allows adjustment for any number of covariates, whether they are discrete (e.g., the technique used [CT or MRI]) or continuous (e.g., age or serum electrolyte level), and then computes a test for each, including, of course, a statistical test of the difference overall between the treatment and control groups. Both survival and hazard functions can refer to outcomes other than death. In the Cox model, this hazard is assumed to be separable into a product of one function that depends on time and another function that captures all the other variables including, specifically, the relative difference between treatment and control groups.

No matter which form of survival analysis statistical test is used, four assumptions must be met:

- Each person must have an identifiable starting point. All subjects should enter the trial at the same time in the course of their illness. Using diagnosis as an entry time can be problematic, because people may have had the disorder for varying lengths of time.
- A clearly defined and uniform end point is required. This is not a problem if the end point is death, but it can be a problem if the end point is recurrence of disease.
- The reasons that people drop out of the study cannot be related to the outcome. If persons have dropped out because they can no longer travel to their scheduled appointments as a result of the worsening of symptoms of the disease under study, the chances of survival could be seriously overestimated. Otherwise, any changes we see may be due to these secular changes, rather than the intervention.
- Diagnostic and treatment practices must not change over the life of the study.

We have said that survival or life table analysis allows us to look at how long people are in one state (e.g., life) followed by a discrete outcome (e.g., death). This analysis can handle situations in which the people enter the trial at different times and are followed up for varying periods; it also allows us to compare two or more groups [4]. The methods of life table (survival) analysis are increasingly used in diagnostic imaging research in recent years, and we therefore offer a recent review of a relevant research study [6].

This multicenter study evaluated patients with intramural hematoma of the aorta and hospital admission less than 48 hr after onset of initial symptoms. Patients were enrolled between January 1994 and December 2000 after confirmation of intramural hematoma on two imaging studies (transesophageal

TABLE 2: Kaplan-Meier Life Table Analysis of the Data in Table 1

Time (months)	No. at Risk	No. of Deaths	Death Rate	Survival Rate	Cumulative Survival Rate
t	R_t	D_t	q_t	p_t	P_t
22	9	1	0.1111	0.8889	0.8889
29	8	1	0.1250	0.8750	0.7778
46	5	1	0.2000	0.8000	0.6222
61	4	1	0.2500	0.7500	0.4667

Note—Reprinted with permission from [4].

echocardiography, CT, or MRI). Sixty-six patients were consecutively enrolled over the course of 7 years. They were subjected to medical treatment in an ICU setting and surgical treatment if indicated (criteria for surgical intervention are available in the original article). Follow-up of these patients ranged from 6 to 123 months and included outpatient visits and CT 6 months after the event and yearly thereafter.

From the raw data collected from 66 patients, a Kaplan-Meier curve was built (Fig. 5). Dissecting Figure 5, we obtain the following information: survival is set at 100% at the beginning of the study, when patients initially present to the emergency department. Each ladder step indicates a drop in survival—that is, the death of a patient because that was the event defined as the main outcome. A rapid decline ensues because close to 20% of patients die in the acute phase. The first loss of information occurs around 6 months, when the first follow-up is scheduled. The triangles indicate censored data, and the figure shows that at 20 months, 12 patients have already been censored. Figure 5 shows that the drop in survival is faster in the initial months after intramural hematoma: the curve drops faster between 20 and 60 months than later in the study.

Differential survival of subgroups of the study was assessed using the log-rank test. The resulting Kaplan-Meier curves obtained from comparison of patients who received oral β-adrenergic receptor blockers (experimental group) and those who did not (control group) are displayed in Figure 6. Visual analysis easily reveals that patients taking β-blockers (upper curve) enjoyed much greater survival than patients who did not receive the medication (lower curve). In fact, the upper curve shows that only one patient died early in the study, and that subsequently all patients from whom information is available are still alive. However, many censored data points are seen, but there is no reason to believe these patients have died without knowledge of the study's investigators, which would falsely lead to the conclusion that β-blockers have a protective effect. The log-rank test performed on these two subgroups of patients revealed important information that was embedded in the initial Kaplan-Meier curve (Fig. 5) and could not have been obtained had it not been for this separate analysis.

Conclusion

In this article, we address life table and survival analysis and describe life table techniques such as the Kaplan-Meier approach. For the comparison of two or more groups, we describe the Mantel-Cox log-rank test. Finally, we discuss the Cox proportional hazards model, which examines the difference in the survival curves and also accounts for other variables (covariates). These statistical methods allow one to work with nontraditional units of analysis: person–time rather than person only. These tools are seen increasingly in the research literature and are gaining popularity in radiology research. These methods of data analysis have potential applications in many fields of radiology, most notably in the analysis of screening techniques and interventional studies.

Acknowledgments

Sadly, Dr. Stolberg passed away last January. His determination and enthusiasm were key in seeing this project to completion, and we are indebted to him for all he accomplished.

We express our appreciation and gratitude to Monika Ferrier for her patience and support, which is always a difficult task with several authors. She has kept us on track and prepared the manuscript.

References

1. Norman GR, Streiner DL. *PDQ statistics*, 2nd ed. St. Louis, MO: Mosby, 1997
2. Altman DG, Machin D, Bagant TN, Gardner MJ. *Statistics with confidence*, 2nd ed. London, UK: BMJ Books, 2000
3. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Statist Assoc* 1958; 53:457–481
4. Norman GR, Streiner DL. *Biostatistics: the bare essentials*, 2nd ed. Hamilton, ON: B. C. Decker, 2000
5. Cox DR. Regression models and life tables. *J Roy Statist Soc* 1972; 34:187–220
6. Von Kodolitsch Y, Csosz SK, Koschyk DH, et al. Intramural hematoma of the aorta: predictors of progression to dissection and rupture. *Circulation* 2003; 107:1158–1163

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005
16. Statistical Inference for Proportions, April 2005
17. Reader Agreement Studies, May 2005
18. Correlation and Regression, July 2005

Multivariate Statistical Methods

Nancy A. Obuchowski[1]

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 20th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

Received November 17, 2004; accepted after revision November 23, 2004.

[1]Department of Quantitative Health Sciences and Department of Radiology, Rm. Wb4, The Cleveland Clinic Foundation, 9500 Euclid Ave., Cleveland, OH 44195. Address correspondence to N. A. Obuchowski.

AJR 2005; 185:299–309

0361–803X/05/1852–299

What Is "Multivariate"?

In radiology studies we often measure more than one end point, or outcome variable, on each patient. "Multivariate" means multiple outcome variables measured on the same patient. We might use multiple end points in a study for several reasons. In designing a study we might not know which end point is important, so we measure a variety of end points to find which ones are important. In other studies, we may have a set of variables that have been shown in the past to be important or that are important for clinical reasons, so we measure the set of variables.

Many examples of multivariate data occur in radiology studies. Consider the following five examples. A study was conducted to assess the effects of diagnostic imaging information on patients with lower back pain (Michael T. Modic, personal communication). Half the study patients were given the results of their imaging test; the other patients were not given their results. Six weeks later, the investigators recorded five variables (i.e., pain, function, absenteeism, quality of life, and self-efficacy) on each patient and compared the two groups. In a second study [1], mammographers were randomized to one of two groups: an intervention group to improve reviewer performance or a control (no intervention) group. The two groups interpreted mammograms before and after the intervention period. For this study there were two outcome variables: change in reviewer performance on mammograms with malignant lesions and change in reviewer performance on mammograms not containing malignant lesions. A third study compared the cardiovascular effects of an intensive, cholesterol-reducing diet with those of a standard diet (R. Brunken, C. Esselstyn, personal communication). Each patient in the two groups underwent PET before and after a short trial on the diets. Two variables were measured on the PET scans: the change in size and the change in severity of perfusion abnormalities. A fourth study was performed to assess the image quality of abdominal CT [2]. Patients' images were scored on 18 characteristics (i.e., 18 outcome variables): organ edge sharpness measured at six sites, visibility of 10 different vessels, and motion of the abdominal wall above and below the umbilicus. The final example comes from a study investigating the quantitative characteristics that can be used to distinguish benign and malignant breast lesions on MRI (Radhika Sivaramakrishna, personal communication). Four variables—margin fluctuation, tumor border roughness, entropy from 2D surface temperature, and a function of the convex hull area—were measured on each lesion and were used to distinguish the two types of lesions.

When Should We Use Multivariate Statistical Methods?

There are essentially three situations when multivariate statistical methods are needed [3]. This section describes each situation and provides examples. The appropriate multivariate statistical methods are applied later for each example.

First, multiple individual variables may be of interest to us, and we want to explore each one. The lower back pain study has five variables of interest (pain, function, absenteeism, quality of life, and self-efficacy), and we want to explore the effect of diagnostic imaging information on each. One common approach would be to test each variable and report the resulting *p* value; if a *p* value is less than the conventional level of 0.05, or 5%, then we might conclude that diagnostic information has an effect on this variable. Such an approach can provide misleading results. For example, suppose that diagnostic information really has no effect on any of the five variables. If we adopt a 5% significance level for each variable, then we have a 5% chance of incorrectly concluding that diagnostic information has an effect on the particular variable, and a 95% chance on each variable of correctly concluding that there is no effect. If the five variables act independently of each other, the probability of drawing the correct conclusion on all five variables is $(0.95)^5 = 0.77$, or only 77%. There is a 23% chance that we will make at least one mistake. This is known as the experiment-wise error rate [3]. Note that here we have assumed that the variables are independent; often, however, they are correlated in some way. This means that 0.23 is the upper limit on the probability of making at least one mistake (the good news), but now we cannot calculate the exact probability (the bad news) [3].

This example illustrates that *p* values from individual statistical tests (i.e., univariate analyses) are not necessarily significant just because they are less than 0.05 [4]. A simple solution, which is particularly useful when the variables are only loosely biologically related to one another, is to calculate and report adjusted *p* values [4]. Adjusted *p* values can be compared with 0.05, and if they are less than 0.05, then we can conclude that the variable is significant. With adjusted *p* values, if there really is no effect for any of the variables, then there is a 5% chance that we will make one or more mistakes and a 95% chance that we will make the correct conclusion on all the variables. We describe and illustrate this approach in the section on Adjusted *p* Values.

In the second situation, we have a set of variables that we are interested in examining as a set. In many situations the variables in the set are measured on two groups, and we are interested in the patterns of differences between the two groups for the set of variables. If differences between the groups are found for the set of variables, then we may want to explore which variable or group of variables is different for the two groups, but this is a secondary issue. It can happen that no one variable distinguishes the two groups, but the combination of variables in the set distinguishes the two groups well.

For this second situation, we present two examples that illustrate slightly different statistical methods. In the mammography study described earlier, the primary question focuses on the differences in the change in performance of the two groups of physicians; for this primary question, the change in performance on images containing a malignant lesion and the change in performance on images not containing a malignant lesion are treated as a single set. If a difference in this set is found between the two groups of physicians, then we might like to investigate whether a difference exists in performance just on images with a malignant lesion, just on images without a malignant lesion, or both. In this example it is not clear whether both measures of reviewer performance will be improved (or worsened) by the intervention, or whether one measure will be affected and the other measure will not be affected, or even whether the measures will be affected in opposite ways. For this example, we will apply a multivariate test, called the Pillai-Bartlett test [3], that looks for any type of difference between the two groups.

In the cardiovascular diet study, the primary question is whether diet affects perfusion abnormalities. If a difference is found in the perfusion abnormalities of patients with and those without the intensive diet, then we would like to investigate which variable is most affected by the diet; however, this is a secondary issue. The two outcome variables in this example (i.e., extent and severity of perfusion abnormalities) are closely biologically related. We expect them both to improve or neither to improve with the intensive diet. They may be improved by the diet to different degrees, or magnitudes, but we expect them to be affected in the same direction by the diet. For this example we will apply a multivariate test, a linear combination test that takes into account the close relationship of the variables and their consistent direction for change.

The third situation in which multivariate methods are needed is when we are not particularly interested in the raw variables themselves, but rather in the use of a combination or a subset of them. We again present two examples, each with different goals for the analysis. First, in the image quality study the 18 questions posed to the reviewers represented a list of important image quality characteristics, but none of the questions by themselves is of primary interest. Furthermore, with 18 variables and only 37 total patients in this study, we have far too many variables to investigate with this sample of patients. To reduce the number of variables, we could just discard some variables on the basis of some preliminary analyses; however, it would be better to keep all of the information if we could condense it into new, fewer variables. Multivariate methods such as cluster analysis can be used to identify similar groups, or clusters, of variables. Then, if the grouping of variables makes sense to us, we can create a new variable from the variables in each group. In this way we have reduced the data from 18 variables into however many groupings we think are appropriate, and we have created new variables that are functions of the old variables; no variables are omitted.

In the MRI breast imaging study, the goal was to identify the variable or set of variables that best distinguished known benign lesions from known malignant lesions. Once a variable or set of variables is found, then it can be used in the future to differentiate lesions of unknown status. The multivariate methods needed here are different from those needed in the previous example. In the CT image quality study, we wanted to group the variables, not the patients or lesions; we had no way of knowing if the groupings of variables were correct. In the MRI breast lesion study, we want to group the lesions into benign or malignant; because we know the pathology of each lesion, we know whether the groupings are correct. For this example, we will use multivariate methods such as discriminant analysis and multiple-variable logistic regression analysis to identify the best set of variables for grouping lesions of known status.

Adjusted *p* Values (Lower Back Pain Example)

In an ongoing study, patients with an acute episode of lower back pain were consented for the study and underwent MRI. Patients were randomized at presentation to one of two groups: diagnostic imaging information provided at presentation versus diagnostic imaging information not provided. Patients in the first group were told about the findings on their MRI examination, whereas patients in the latter group were not provided any information about the findings of their examination. Six weeks later, patients in both groups recorded their pain, function, quality of life, self-efficacy, and absenteeism using standardized questionnaires.

Because this is an ongoing study, we do not have raw data to present. For calculation of adjusted *p* values, however, we just need the *p* values from the univariate analysis of each variable (i.e., the unadjusted *p* values). Table 1 provides an illustrative example of the sort of findings that might be obtained. In the second column, *p* values are presented from Student's *t* tests on each of the five variables. Quality of life, function, and pain are all significant at the 0.05 level.

TABLE 1: Calculation of Adjusted *p* Values for Lower Back Pain Example

Variable (*i*)	Unadjusted *p* (p_i)	r_i	Conclusion	Adjusted *p*
Quality of life (1)	0.003	0.015	Reject	0.015
Function (2)	0.007	0.028	Reject	0.028
Pain (3)	0.048	0.144	NS	0.144
Self-efficacy (4)	0.070	0.140	NS	0.144[a]
Absenteeism (5)	0.145	0.145	NS	0.145

Note—NS = not significant.

[a] Adjusted *p* values are usually just the r_i values. However, because the adjusted *p* values must be sequentially ordered, the adjusted *p* value for self-efficacy takes on the value of the previous adjusted *p* value.

There are several methods for calculating adjusted *p* values [4]; we describe and illustrate one simple method [5] here.

Step 1

Order the unadjusted *p* values from smallest to largest, so that $p_1 < p_2 < p_3 \ldots < p_i \ldots < p_n$, where p_i is the *i*-th variable and *n* is the total number of outcome variables. The unadjusted *p* values for the lower back pain study have been ordered in this way in Table 1. Note that $n = 5$ in this example.

Step 2

Compute the value of r_i as $(n - i + 1)p_i$. For example, $r_1 = np_1$ and $r_2 = (n - 1)p_2$. The values of r_i for the lower back pain study are given in the third column of Table 1.

Step 3

Compare r_1 with the planned type 1 error rate (usually 0.05). If r_1 is less than or equal to the planned type 1 error rate, then conclude that the variable is statistically significant and continue to step 4. If r_1 is greater than the planned type 1 error rate, then we conclude that none of the *n* variables is statistically significant. In the lower back pain example, r_1 equals 0.015 and is less than 0.05. So we conclude that quality of life is affected by diagnostic imaging information.

Step 4

Continue to compare r_i with the planned type 1 error rate, starting with $i = 2$ and continuing to $i = n$. If r_i is less than or equal to the planned type 1 error rate and if the previous variable (i.e., $i - 1$) is determined to be statistically significant, then conclude that the *i*-th variable is also statistically significant. As soon as a variable is determined not to be statistically significant, then all remaining variables are also considered not statistically significant. In the lower back pain study, the first two variables (i.e., quality of life and function) are statistically significant. The variable "pain" is not statistically significant (i.e., $r_3 > 0.05$); thus, none of the remaining variables (i.e., self-efficacy and absenteeism) is considered to be statistically significant.

On the basis of the unadjusted *p* values, we would have concluded that quality of life, function, and pain are all affected by diagnostic information. However, we know that the experiment-wise error rate (i.e., the overall significance level) greatly exceeded 5%. On the basis of the adjusted *p* values, we conclude that quality of life and function (not pain) are affected by diagnostic information. Because these are adjusted *p* values, the overall significance level has been maintained at equal to or less than 5%.

Pillai-Bartlett Statistic (Mammography Example)

Pepe et al. [1] describe a study design to test whether a specific intervention (i.e., an educational program) improves the performance of mammographers. Radiologists are randomly assigned to either the intervention group or a control (i.e., no intervention) group. The radiologists in both groups first interpret a common set of images. The performance of each reviewer for images with and without breast cancer (e.g., sensitivity and false-positive rate [FPR], respectively) is recorded. After the intervention period, a second set of images is interpreted by the same radiologists. The authors want to test whether the radiologists' performances are altered by the intervention.

(Note that for convenience, we will use the terms "sensitivity" and "false-positive rate" to denote the two measures of reviewer performance. In this example, however, we are emphasizing the performance of a sample of reviewers on sets of fixed images; we are deemphasizing the sampling of the patients for the study. A variety of statistical methods are available [6–8] for characterizing and comparing diagnostic accuracy that take into account the sampling of both patients and reviewers.)

Table 2 summarizes a set of fictitious data (i.e., no actual data were reported by Pepe et al. [1]). The first two columns are the changes in sensitivity and FPR for the intervention and control groups for the 14 reviewers (seven per group). The sensitivity changes are illustrated in Figure 1A, and the FPR changes are illustrated in Figure 1B. Note that in both figures there is considerable overlap—that is, physicians in the two groups have similar increases in sensitivity and similar increases in FPR. In fact, *t* tests (i.e., univariate analysis) on the changes in sensitivity and FPR indicate no statistically significant differences between the control and intervention groups (last column of Table 3).

Figure 1C illustrates, simultaneously, the changes in sensitivity and FPR. The figure shows two distinctly separate groups of data points—that is, physicians in the control group have changes toward the lower left, whereas physicians in the intervention group have changes toward the upper right; the distinction is not apparent from the univariate displays of the data.

Clearly, we want a test statistic that takes both measures of performance into account simultaneously. There are four well-known and related test statistics that can be applied here: the Pillai-Bartlett trace (also called the Pillai-Bartlett or Bartlett statistic), Wilks' lambda (also called the likelihood ratio test statistic), the Hotelling-Lawley trace, and Roy's largest eigenvalue statistic (also called Roy's maximum characteristic root or the union-intersection statistic). Most statistics packages will output the results of all four. They require certain assumptions about the basic data distributions—that is, that the data follow a multivariate normal distribution and that the variances and covariances in each group are identical (i.e., homoskedastic). Many different methods are available for assessing the multivariate normality and homogeneity of variance and covariance assumptions. Some simple methods are described and illustrated as follows:

Assessing Multivariate Normality Assumption

The following are the steps for assessing the multivariate normality assumption [9, 10].

Step 1—For each (treatment) group and each outcome variable, test that the data follow a univariate normal distribution. This is best accomplished by calculating the Shapiro-Wilk W test and examining the statistical measures called skewness (i.e., symmetry) and kurtosis (i.e., peakedness). Most standard statistical packages can do this for you. In SAS [11], the code for our mammography example is *proc univariate normal; by trt; var sen fpr;*. (Note that sen and fpr are the variable names for the change in sensitivity and change in FPR; trt is the variable name for the treatment effect: 1 = intervention, 0 = no intervention.) The *p* values for the Shapiro-Wilk test are 0.766 and 0.857 for sensitivity and FPR for the control group, and 0.278 and 0.100 for sensitivity and FPR for the intervention group; because these *p* values exceed 0.05, the univariate assumption is reasonable. The skewness values are –0.07 and 0.26 for sensitivity and FPR for the control group, and –1.09 and –0.70 for sensitivity and FPR for the intervention group. The skewness values should be near 0 for univariate normality. For this small sample size, these values are not unreasonable. (Note that for large sample sizes, we expect the values to be much closer to 0.) Finally, the kurtosis values are –1.31 and –0.79 for sensitivity and FPR for the control group, and 0.72 and –1.29 for sensitivity and FPR for the intervention group. The kurtosis values should be near 3 for univariate normality, but SAS outputs the kurtosis values minus 3. Thus, in our example we examine the kurtosis values to see if they are near 0; again, for this small sample size, these values are not unreasonable.

Step 2—For each (treatment) group, compute all of the principal components and then compute the skewness and kurtosis measurements for each principal component. Most standard statistical packages compute principal components. The SAS code for our mammography example is *proc princomp out = prin; by trt; var sen fpr; proc univariate; by trt; var prin1 prin2;*. In our mammography example, the skewness values are –1.02 and –0.36 for the first and second principal components for the control group, and 0.01 and 0.43 for the first and second principal components for the intervention group. If the data follow a multivariate normal distribution, then (*n* / 6)(sum of the square of the skewness measurements) should follow a chi-square distribution with *p* degrees of freedom, where *p* is the number of outcome variables. In our example, $p = 2$, and the test statistic for the control group is $(7 / 6)(1.04 + 0.13) = 1.37$; the test statistic for the intervention group is $(7 / 6)(0.0001 + 0.18) = 0.21$. The critical value from a chi-square distribution with 2 degrees of freedom is 5.99; because 1.37 is less than 5.99 and 0.21 is less than 5.99, the multivariate normal assumption is reasonable. The kurtosis values are –0.06 and –1.55 for the first and second principal components for the control group, and –1.07 and 0.06 for the first and second principal components for the intervention group. If the data follow a multivariate normal distribution, then $(n / 24)^{1/2}$(sum of the kurtosis measurements) should follow a normal distribution. In our example, the test statistic for the control group is $(7 / 24)^{1/2}(–0.06 + –1.55) = –0.87$; the test statistic for the intervention group is $(7 / 24)^{1/2}(–1.07 + –0.06) = –0.55$. Because both test statistics have values between –1.96 and 1.96, we conclude that the multivariate normal distribution is reasonable.

Steps for Assessing Variance and Covariance Homogeneity Assumption

Step 1—For each variable, test if the variances of the groups are the same. When there are two groups, this can be done easily with most statistical packages. In SAS, a test for homogeneity of variances is performed when the *t* test procedure is executed. The SAS code for our example is *proc ttest; class trt; var sen fpr;*. After the *t* test results, SAS outputs the results of the test of the hypothesis of equal variances. For sensitivity, the *p* value of this test is 0.577, indicating that we can assume equal variances in the two groups for this variable. For FPR, the test for equal variances gives a *p* value of 0.445, indicating that we can assume equal variances in the two groups for this variable, as well.

Step 2—For each group, examine the correlation coefficient(s) between the variables. In SAS we can obtain Pearson's correlation coefficient between sensitivity and FPR for the two groups by executing the following code: *proc corr; by trt; var sen fpr;*. For the control group, the Pearson's correlation coefficient

TABLE 2: Changes in Reviewer Performance During the Intervention Period (Fictitious Data)

Change In		Sensitivity		FPR	
Sensitivity	FPR	Before Intervention	After Intervention	Before Change	After Change
Control group (*n* = 7 reviewers)					
0.10	0.05	0.80	0.90	0.04	0.09
0.12	0.06	0.75	0.87	0.06	0.12
0.08	0.07	0.88	0.96	0.05	0.12
0.13	0.08	0.77	0.90	0.08	0.16
0.07	0.11	0.90	0.97	0.10	0.21
0.14	0.08	0.66	0.80	0.07	0.15
0.10	0.10	0.59	0.69	0.12	0.22
Intervention group (*n* = 7 reviewers)					
0.16	0.05	0.70	0.86	0.05	0.10
0.16	0.06	0.75	0.91	0.06	0.12
0.15	0.09	0.80	0.95	0.08	0.17
0.12	0.10	0.85	0.97	0.10	0.20
0.14	0.12	0.76	0.90	0.11	0.23
0.11	0.12	0.88	0.99	0.09	0.21
0.07	0.12	0.85	0.92	0.13	0.25

Note—Changes are defined as performance after intervention minus performance before intervention. FPR = false-positive rate.

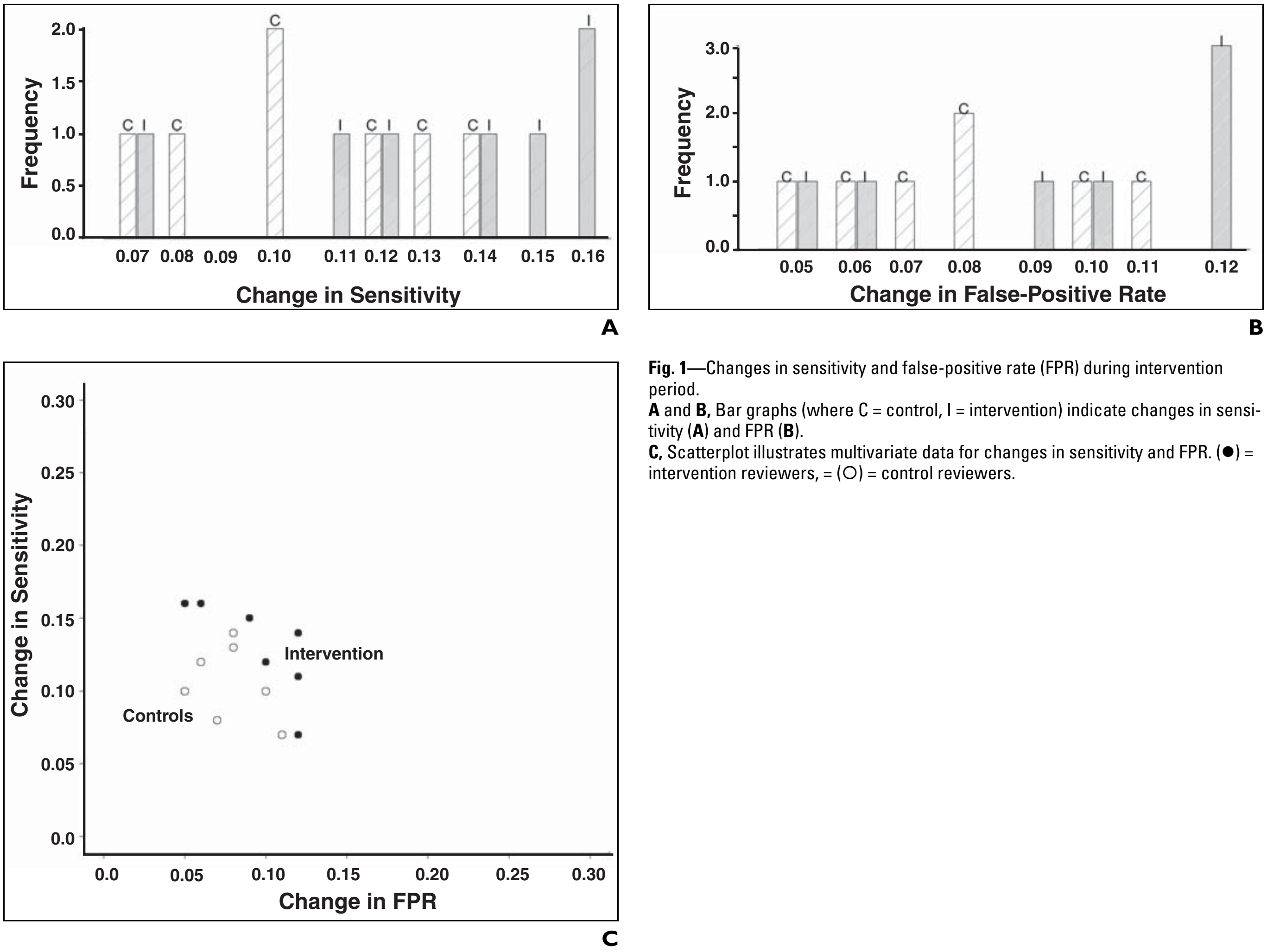


Fig. 1—Changes in sensitivity and false-positive rate (FPR) during intervention period.
A and **B,** Bar graphs (where C = control, I = intervention) indicate changes in sensitivity (**A**) and FPR (**B**).
C, Scatterplot illustrates multivariate data for changes in sensitivity and FPR. (●) = intervention reviewers, = (○) = control reviewers.

TABLE 3: Mean Changes (SDs) in Sensitivity and FPR for Control Group and Intervention Group

Outcome Variables	Control	Intervention	*t* Statistic	*p*
Sensitivity	0.11 (0.03)	0.13 (0.03)	1.55	0.148
FPR	0.08 (0.02)	0.09 (0.03)	1.15	0.273

Note—FPR = false-positive rate.

between sensitivity and FPR is –0.32; the same correlation for the intervention group is –0.75. The two correlations differ in magnitude, but with a small sample size this is not unreasonable. So we conclude that the homogeneity assumption is reasonable.

Several authors [3, 12] have suggested that for general use, the first statistic—the Pillai-Bartlett statistic—should be used. Their rationale is based on several aspects of the tests' performance, including the fact that the Pillai-Bartlett statistic performs well even when the multivariate normality and homogeneity of variances and covariances assumptions are not entirely met. When two groups are being compared, as in the mammography example, or if there is only one outcome variable (i.e., the univariate case), these four statistics give the same result anyway. The code for SAS to produce these test statistics for the mammography example is *proc glm; class trt; model sen fpr = trt; manova h = trt;*. For our example data, the F-statistic is 4.10 and has an associated *p* value of 0.047. Thus, we reject the null hypothesis that the intervention had no effect and conclude that mammographers' performances are affected by the intervention.

In the multivariate setting, we need to construct simultaneous confidence intervals for the outcome measures. If there are *k* outcome measures (in our example, $k = 2$), then the confidence statements for all the *k* outcome measures hold simultaneously with a specified high probability (usually, 0.95). In other words, it is a guarantee of a specified probability against any of the *k* statements being incorrect. The formula for constructing a simultaneous confidence interval for the difference between two populations is given in the appendix. For our example, the 95% simultaneous confidence intervals for the difference in the change in sensitivity and FPR between the two groups of physicians are [–0.03 to 0.07] and [–0.03 to 0.05], respectively.

So far we have investigated only the changes in sensitivity and FPR between the two groups; however, the actual sensitivities and FPRs before and after intervention (last four columns of Table 2) may provide other useful information. For example, the change in sensitivity is negatively correlated to the preintervention sensitivity (Pearson's correlation coefficient is $r = -0.49$, $p = 0.073$); in contrast, the change in FPR is positively correlated to the preintervention FPR (Pearson's correlation coefficient is $r = 0.90$, $p < 0.001$). This suggests that reviewers with higher sensitivities before intervention experience a smaller increase in sensitivity after the intervention period, and reviewers with higher FPRs before intervention experience a larger increase in FPR. Reviewers' preintervention sensitivities and FPRs do not appear to be related to one another (Pearson's correlation coefficient, $r = 0.03$, $p = 0.911$). We might ask if, by chance, the reviewers in the intervention group tended to have lower preintervention sensitivities and higher preintervention FPRs than those in the control group; if true, this would suggest that the intervention does not affect performance after all. A simple comparison of the mean preintervention measures between the two groups does not support this, but this is one reason that baseline values (e.g., preintervention measures) should be evaluated.

With a larger sample, we could fit a model describing a reviewer's postintervention performance (i.e., the dependent variable) as a function of baseline performance and treatment group. With data from only 14 reviewers, however, we probably don't want to fit this complicated a model. Rather, we will use and report the results of the Pillai-Bartlett statistic based on the changes in performance, along with this simple analysis of the baseline performances.

Linear Combination Test (PET Perfusion Imaging Example)

In a pilot study, patients with coronary artery disease were randomized to either an intensive, lipid-lowering, plant-based diet ($n = 5$) or a standard diet of 30% calories as fat ($n = 4$). Patients' hearts were studied by rubidium-82 PET perfusion imaging 3 weeks after beginning the diet; the 3-week images were compared with PET images taken at baseline (prediet). The changes in extent (i.e., size) and severity of perfusion abnormalities were recorded.

The data are given in Table 4 and illustrated in Figure 2; note that the data have been modified for proprietary reasons. The mean changes and SDs are summarized for the two groups in Table 5. The figure and the means in Table 5 suggest that the intensive diet may improve the extent and severity of perfusion abnormalities; however, the p values from the univariate analysis (last column of Table 5) are not significant at the 0.05 level. The sample sizes in the two groups are quite small, so we can conclude little about the effect of the diet.

TABLE 4: Changes in PET Findings After 3 Weeks (Modified Data)

Intensive Diet ($n = 5$)		Standard Diet ($n = 4$)	
Extent	Severity	Extent	Severity
8.54	2.97	4.99	–1.68
16.57	5.84	6.85	2.36
1.72	10.25	–19.27	0.0
0.55	–2.04	–1.36	–7.46
6.85	34.23		

Note—Positive values for change indicate improvements (i.e., reduction in size or severity) in perfusion abnormalities at 3 weeks.

The Pillai-Bartlett trace test for these data gives an F-value of 1.97, with an associated p value of 0.220. Thus, on the basis of this general purpose multivariate test, we still find insufficient evidence to reject the null hypothesis.

We now perform a multivariate method [13] that takes advantage of the common direction we expect the outcome variables to take. It is a linear combination test, which means that it combines, in a linear fashion, the univariate test statistics, taking into account their correlation. This analysis will test the null hypothesis that diet has no effect on the two imaging variables, versus the alternative hypothesis that the diet affects the variables in the same direction. The same assumptions about the data—that is, that the data follow a multivariate normal distribution and that the variances and covariances in each group are identical—are again required. These assumptions appear reasonable for the data (analysis of assumptions follows the same steps as described in the previous example, but the details are not shown here).

To perform the linear combination test, we need the value of the univariate test statistics—that is, 1.47 and 1.61 (from the fourth column of Table 5), and the value of the correlation between the outcome variables. We will take an average of the Pearson's correlation coefficients from the two groups, but because the sample sizes in the two groups are different, we will use an average weighted by the sample sizes. Pearson's correlation coefficient between the changes in extent and severity for the intensive diet group is 0.06, based on five patients; the correlation for the control group is 0.02, based on four patients. The weighted average, or pooled correlation, is 0.04.

Pocock et al. [13] give the formula for the test statistic in matrix form, which can be simplified when there are only two outcome measures (see Appendix 2). The value of the test statistic for our example is 2.14. Pocock et al. report that for two outcome variables, their test statistic has an approximate t distribution with degrees of freedom equal to total sample size minus 4; for our example the degrees of freedom is 5; thus, the associated p value is 0.085. We do not reject the null hypothesis at the 0.05 level; however, this result might be

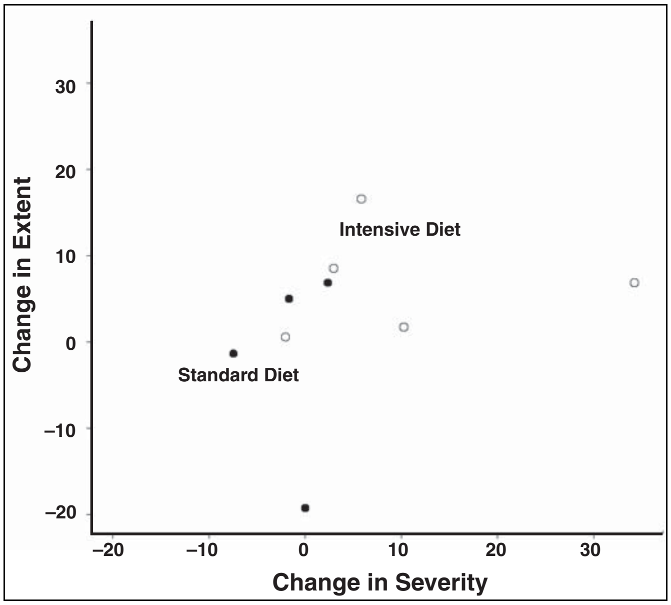


Fig. 2—Scatterplot illustrates data from study of PET perfusion imaging of patients with coronary artery disease who were randomized to an intensive diet (○) or to a standard diet (●).

persuasive enough to encourage us to perform the full-scale study.

Cluster Analysis (CT Image Quality Example)

In this study, Herts et al. [2] compared the image quality of helical CT of the abdomen for two scanning times: 0.75 versus 1 sec per revolution. Three radiologists each evaluated 18 separate image quality variables for 37 patients: 17 patients at 0.75 sec and 20 patients at 1.0 sec. They used a 10-point scale to evaluate each variable: 1 (blurry) to 10 (sharp) to describe organ edge sharpness at six sites; 1 (poorly visualized, unenhanced) to 10 (well visualized, enhanced) to describe vessel visibility at 10 sites; and 1 (frequent or large-scale) to 10 (none detected) to describe motion of the abdominal wall above and below the umbilicus. With 37 patients evaluated by three reviewers on each of 18 image quality questions (i.e., 1,998 total observations), the data are best obtained by downloading them from the ACR Web site where these modules are described.

The analysis plan for this example is as follows: First, find suitable clusters of like variables (the number of clusters should be substantially less than the number of original variables); second, from each cluster, create a new variable from the original variables in the cluster; and third, use these new variables for comparing the two scanning times. Note that this is only one type of cluster analysis. In particular, in this example we are looking for clusters of similar variables. In other situations we might be looking for clusters of similar patients. Although they are not illustrated in this paper, a variety of approaches can be used for this sort of cluster analysis, including simple graphical methods and hierarchical methods (e.g., nearest neighbor, average distance, and minimum variance approaches) [11]. The various options are available in many statistical software packages, including SAS.

Before performing the cluster analysis, we must first examine the correlations between the original variables. If there are no large negative correlations, then we can proceed with the cluster analysis. However, if some of the variables are highly negatively correlated, then the cluster analysis will see those variables as being very dissimilar, when in fact they are highly similar, just their scale of measurement is reversed. In this situation, these variables need to be transformed before the cluster analysis [11]. In the CT image quality study, there are no large negative correlations between any of the variables.

We now perform cluster analysis to see if the data suggest any groupings of like variables.

Statistical packages usually offer several options for the analysis. First, the groupings can be based on either the correlations or covariances between the variables. If you want all the variables to be given equal importance, then use the correlation option; if you want variables with larger variances to have more importance, then use the covariance option. (Note that in situations in which the variables are measuring different things with different units of measurement, the correlation option is usually most appropriate.) Second, the variables in each cluster can be either the optimized weighted average of the variables (the first principal component option) or the unweighted average of the standardized (the centroid option based on correlations) or nonstandardized (centroid option based on covariances) variables. Note that a standardized variable is just the value of the variable minus its mean value, all divided by the SD.

The SAS code for performing cluster analysis on the CT image quality example, with the correlation and first principal component options (defaults in SAS), is *proc varclus; var quest1-quest18;*. The cluster analysis produced six clusters from the original 18 variables, as listed in Table 6.

In contrast, using the centroid option based on covariances (SAS code is *proc varclus centroid cov; var quest1-quest18;*), the analysis produced 11 clusters from the 18 original variables. These 11 clusters were just further divisions of the six clusters in Table 6; no new grouping of variables was identified. The authors of the study [2] examined the six clusters in Table 6 and determined that the six clusters made biologic sense. Furthermore, for analysis purposes we prefer fewer clusters. The six clusters were subsequently labeled liver and spleen edge sharpness, renal edge sharpness and abdominal wall motion, portal vein and intrahepatic vessels, celiac axis and common hepatic artery, superior mesenteric vessels and mesenteric branches, and renal artery origin.

TABLE 5: Mean Changes (SDs) in Extent and Severity of Perfusion Abnormalities According to Diet

Outcome Variables	Intensive Diet	Standard Diet	*t* Statistic	*p*
Extent	6.85 (6.39)	−2.20 (11.91)	1.47	0.185
Severity	10.25 (14.13)	−1.70 (4.18)	1.61	0.150

TABLE 6: Results of Cluster Analysis for CT Image Quality Example

Cluster No.	Quest	Original Variables
1	1	Organ edge sharpness of anterior right lobe of liver
	2	Organ edge sharpness of anterior left lobe of liver
	3	Organ edge sharpness of posterior left lobe of liver
	4	Organ edge sharpness of splenic margin
2	5	Organ edge sharpness of anterior and posterior renal margins
	6	Organ edge sharpness of medial and lateral renal margins
	17	Motion of anterior abdominal wall above umbilicus
	18	Motion of anterior abdominal wall below umbilicus
3	7	Vessel visibility and enhancement of main portal vein
	8	Vessel visibility and enhancement of main portal bifurcation
	9	Vessel visibility and enhancement of intrahepatic portal and hepatic veins
4	10	Vessel visibility and enhancement of celiac axis
	11	Vessel visibility and enhancement of common hepatic artery
5	12	Vessel visibility and enhancement of origin of superior mesenteric artery
	13	Vessel visibility and enhancement of superior mesenteric artery and vein at pancreatic head
	14	Vessel visibility and enhancement of mesenteric branch vessels
6	15	Vessel visibility and enhancement of left origins of renal artery
	17	Vessel visibility and enhancement of right origins of renal artery

When using the first principal component option, the statistical packages will provide coefficients for each variable in a cluster. These coefficients can then be multiplied by the standardized variables before summing the scores in each cluster. The sums of the scores for each cluster become the new variables for analysis. The authors of this study [2], however, took the sum of the scores of the original variables in each cluster (i.e., the principal component coefficients were not used). They divided the sum of each cluster by the number of original variables in the cluster so that the scale would be the same as the original variables. They used these newly generated six variables to compare the 0.75- and 1.0-sec scans. Such an approach is easy to explain and interpret and, in this example, produced similar results.

The means of these six new variables for the two scanning times are given in Table 7. Various methods can now be used to compare the two scanning times. Here, we will use the Pillai-Bartlett statistic to test the hypothesis that the two scanning times differ for one or more of the six new image quality variables. We perform a separate analysis for each of the three reviewers. The results of the Pillai-Bartlett trace test are as follows: reviewer 1, $p = 0.119$, reviewer 2, $p = 0.769$, and reviewer 3, $p = 0.907$, suggesting that there is insufficient evidence to conclude that the image quality of the two scanning times is different.

Multiple-Variable Logistic Regression Analysis (MRI Breast Lesion Example)

Quantitative border measurements were made from the MRI images of 42 benign lesions and 47 malignant lesions from 89 total patients. The status of the lesions was known from biopsy results. Four border measurements were computed: margin fluctuation (MF), tumor border roughness (TBR), entropy from 2D surface temperature (GST), and a function of the convex hull area (FCHA). The goal of the study was to determine which variable or set of variables best distinguishes benign from malignant lesions. The data, modified for proprietary reasons, can be obtained by downloading them from the ACR Web site where these modules are described.

The means and SDs of the four border measurements for benign and malignant lesions are summarized in Table 8. To suit the goals of this study, we should think of the dependent variable as whether the lesion is benign or malignant, and the independent, or predictor, variables as the four border measurements. We begin the analysis by first assessing the importance of each predictor variable alone, without consideration of the other predictor variables. We use univariate logistic regression, which is a convenient way to model a binary dependent variable as a function of an independent variable [14]. Because the dependent variable is binary, in logistic regression the dependent variable is represented by the natural log of the quantity: the expected value of the dependent variable divided by one minus its expected value; this is called the "logit transformation." Then the logit transformation of the dependent variable is modeled as a linear function of the predictor variable. The SAS code for the univariate logistic regression analysis with TBR is *proc logistic; model l_type = tbr;* (where "l_type" stands for lesion type and is coded as 1 = malignant and 0 = benign). *Proc logistic* in SAS outputs a useful measure called the "*c* statistic," which has 1.0 as a maximum value and 0.5 as its effective lower value. For the TBR measure, the value of the *c* statistic is 0.787. The interpretation, which is analogous to the interpretation of the ROC curve area [15], is as follows: if presented with a randomly chosen benign lesion and a randomly chosen malignant lesion, the probability of correctly distinguishing the two, by calling the lesion with the higher TBR value "malignant" and the lesion with the lower TBR value "benign," is 0.787, or 78.7%. In Table 8, all the predictors have *c* statistics greater than 0.5, suggesting that all four border measurements have some ability to distinguish benign and malignant lesions. (Note that SAS does not provide standard errors for the *c* statistic, so without additional calculations we cannot determine if the *c* statistics are significantly better than 0.5.)

Discriminant analysis is a powerful multivariate method for separating units (lesions in our example) into two or more populations and allocating units whose population membership is unknown into one of these populations [11]. For the method to work properly, however, the data must follow a multivariate normal distribution. In our MRI breast lesion example we have continuous-type data, but the data do not follow a multivariate normal distribution (this is evident from the first step in assessing the multivariate normality assumption described previously). In other situations, all the predictor variables might not be the continuous type. Some examples of noncontinuous variables that are often used as predictor variables are sex, which is a binary variable; level of pain, which is often rated on an ordinal scale from 1 to 10; and employment status, which is often categorized as employed, homemaker, retired, student, and unemployed.

As an alternative to discriminant analysis, multiple-variable logistic regression is often used [11, 14]. This is an extension of the univariate logistic regression analysis. The dependent variable in the model is again the lesion type, and the border measurements are considered simultaneously as the predictor

TABLE 7: Means (SDs) of Six New CT Image Quality Variables

New Image Quality Variables	0.75 sec	1.0 sec
Liver and spleen edge sharpness	6.78 (0.17)	7.09 (0.28)
Renal edge sharpness and abdominal wall motion	5.81 (1.07)	6.28 (1.02)
Portal vein and intrahepatic vessels	7.47 (0.34)	7.33 (0.45)
Celiac axis and common hepatic artery	6.96 (0.30)	7.00 (0.24)
Superior mesenteric vessels and branches	7.10 (0.24)	7.19 (0.18)
Renal artery origin	4.23 (1.12)	5.04 (1.31)

Note—Means are computed over patients and reviewers. SDs describe the variability among the means of the three reviewers.

TABLE 8: Means (SDs) of Border Measurements in MRI of Breast Lesions

Border Measurements	Benign	Malignant	*c* Statistic
MF	0.0015 (0.0011)	0.0010 (0.0008)	0.668
TBR	8.3404 (10.1768)	22.0823 (15.8907)	0.787
GST	1.6987 (0.2540)	1.8682 (0.2914)	0.707
FCHA	0.8540 (0.0549)	0.8899 (0.0425)	0.693

Note—MF = margin fluctuation, TBR = tumor border roughness, GST = entropy from 2D surface temperature, FCHA = a function of the convex hull area.

variables. As with any statistical modeling, we must be careful not to overfit the model (i.e., include more predictor variables than can be supported by the number of observations in the study). A general rule of thumb with logistic regression analysis is that you need at least 10–15 observations (here, patients) of each type (here, type is patients with a particular lesion pathology) for each predictor variable in the model [16–18]. If we want to assess a model with all four border measurements, then we would need 40–60 patients of each type (i.e., 40–60 with benign lesions and 40–60 with malignant lesions). Our sample size is just adequate for assessing this model.

The SAS code for fitting the model for our example is *proc logistic; model l_type = mf tbr gst fcha/backward lackfit;*. SAS first fits a model with all four border measurements. Then, because we included the "*backward*" option, SAS will drop from the model the predictor variable that is contributing the least to the model. SAS will continue to drop predictor variables until the remaining ones are all statistically significant at the 0.05 level. The "*lackfit*" option tells SAS that we want it to print the results of a test (called the Hosmer and Lemeshow goodness-of-fit test [14]) to see if the model is a good representation of the data. If it is not a good representation of the data, then we will not use the model.

The results of the multiple-variable logistic regression analysis are as follows: The model was first fit with all four border measurements. MF was not statistically significant and contributed least to the model, so it was dropped. Then GST was removed from the model, as well as FCHA.

The final model includes TBR as the only predictor of whether a lesion is benign or malignant. In other words, once you have the TBR value of a lesion, MF, GST, and FCHA do not provide any additional help in distinguishing the lesion as benign or malignant. The *p* value for the model fit is 0.299, indicating that the model is a reasonable fit for the data. Because we want to use the model in the future to predict the status of lesions of unknown type, we need to examine the model. SAS provides an estimate of the model's intercept, -1.0958, and the regression coefficient for TBR, 0.0872. From these values, we can estimate the probability that a lesion is malignant by substituting the TBR value into this equation:

$$\text{Prob(malignant)} = 1 / \{1 + \exp[-(-1.0958 + 0.0872 \times \text{TBR})]\}.$$

For example, if the TBR value of an unbiopsied lesion is 30, then, based on this model, the probability that the lesion is malignant is:

$$1 / \{1 + \exp[-(-1.0958 + 0.0872 \times 30)]\} = 0.82, \text{ or } 82\%.$$

Figure 3 illustrates the probability of a malignant lesion as a function of the TBR value. Clearly, there is considerable overlap in the TBR values of benign and malignant lesions and in their probabilities of being malignant.

As with all statistical modeling, we must remember that the model may perform well for the data used to create the model, but the model may not perform as well with new observations. Thus, before using the model in clinical practice, we must test its performance using different observations. If we have a large sample size, then sometimes we can split the data into a training data set and a testing data set. The training set is used to create the model; the testing set is used to determine how well the model performs. In the MRI breast lesion study, however, the sample size was barely adequate for training.

Another important point is that our model was created on the basis of TBR values between 0.0 and 65.96 (i.e., these are the minimum and maximum TBR values in our sample). We do not know what the relationship is between the probability that a lesion is malignant and TBR values less than 0.0 or TBR values greater than 65.96. Although we can plug any TBR value into our model and get back a value for the probability that the lesion is malignant, this is not advisable. Rather, when using our model to predict whether a lesion is benign or malignant, we should consider only TBR values from the range of TBR values used to create the model.

Discussion

Multivariate statistical methods have many applications in radiology studies. They are particularly useful for controlling the type 1 error rate in a study, and they sometimes provide insight into the multidimensional patterns in the data that would be overlooked with univariate analyses.

As with all statistical analysis, we recommend that an analysis plan be prepared at the start of a study so that the results of the data do not drive the methods used. This can be a particular problem when there are multiple nonsignificant end points. It is sometimes tempting to not report the nonsignificant end points and report only the statistically significant ones. This strategy, however, can lead to serious misinterpretations of the data because the type 1 error rate is not properly controlled. Other good-practice strategies include plotting or otherwise summarizing the raw data so that the results of the statistical analysis can be verified with the raw data, and evaluation of the validity of any assumptions required in the statistical analysis.

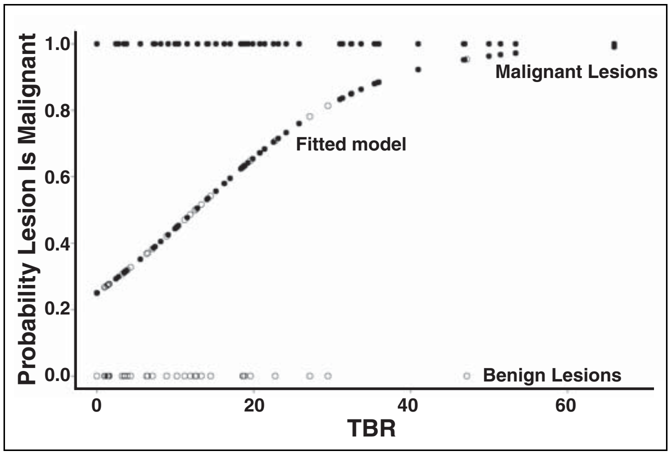


Fig. 3—Graph shows probability of malignant lesion as function of tumor border roughness (TBR) value. Open circles at bottom of figure show probability of malignant lesion for those lesions that we know to be benign (probability = 0). Solid circles at top of figure show probability of malignant lesion for those lesions that we know are malignant (i.e., probability = 1.0). Set of points in middle of figure represents probability of malignant lesion, based on model, for each lesion in data set. Note considerable overlap in TBR values of benign and malignant lesions and in their probabilities of being malignant.

Acknowledgments

I appreciate the contributions of data from several investigators: Radhika Sivaramakrishna, Brian Herts, Richard Brunken, and Caldwell Esselstyn, and the helpful suggestions made on an earlier draft by Craig Beam and Michael Lieber.

References

1. Pepe MS, Urban N, Rutter C, Longton G. Design of a study to improve the accuracy in reading mammograms. *J Clin Epidemiol* 1997; 50:1327–1338
2. Herts BR, Baker ME, Davros WJ, et al. Helical CT of the abdomen: comparison of image quality between scan times of 0.75 and 1 sec per revolution. *AJR* 1996; 167:58–60
3. Hand DJ, Taylor CC. *Multivariate analysis of variance and repeated measures: a practical approach for behavioural scientists.* London: Chapman & Hall, 1987
4. Wright SP. Adjusted *p*-values for simultaneous inference. *Biometrics* 1992; 48:1005–1013
5. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979; 6:65–70
6. 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
7. Obuchowski NA. Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. *Acad Radiol* 1995; 2[suppl 1]:S22–S29, S57–S64, S70–S71
8. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Acad Radiol* 2000; 7:341–349
9. Srivastava MS. A measure of skewness and kurtosis and a graphical method for assessing multivariate normality. *Statistics and Probability Letters* 1984; 2:263–267
10. Looney SW. How to use tests for univariate normality to assess multivariate normality. *The American Statistician* 1995; 49:64–70
11. Khattree R, Naik DN. *Multivariate data reduction and discrimination with SAS software*. Cary, NC: SAS Institute Inc., 2000
12. Olson CL. Comparative robustness of six tests in multivariate analysis of variance. *J Am Stat Assoc* 1974; 69:894–908
13. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; 43:487–498
14. Hosmer DW, Lemeshow S. *Applied logistic regression.* New York, NY: Wiley, 1989
15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29–36
16. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modeling strategies for improved prognostic prediction. *Stat Med* 1984; 3:143–152
17. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treat Rep* 1985; 69:1071–1077
18. Smith LR, Harrell FE, Muhlbaier LH. Problems and potentials in modeling survival. In: Grady ML, Schwartz HA, eds. *Medical effectiveness research data methods* (summary report). Rockville, MD: U.S. Department of Health and Human Services, Agency for Health Care Policy and Research (pub. no. 92-0056), 1992:151–159

APPENDIX 1: Formula for Constructing a Simultaneous Confidence Interval for the Difference Between Two Populations

The formula for constructing a simultaneous confidence interval for the difference between two populations for the *i*-th outcome measure is

$$(\bar{x}_{1i} - \bar{x}_{2i}) \pm c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_{ii}} \qquad (1)$$

where $\bar{x}_{1i}$ is the sample mean of the *i*-th outcome measure for the first population (e.g., control group), $\bar{x}_{2i}$ is the sample mean of the *i*-th outcome measure for the second population (e.g., intervention group), and n_1 and n_2 are the sample sizes from the first and second populations. The value of c is given by

$$c = \sqrt{\frac{(n_1 + n_2 - 2)k}{n_1 + n_2 - k - 1}}\sqrt{F_{k,\, n_1 + n_2 - k - 1(\alpha)}} \qquad (2)$$

where $F_{k,\, n_1 + n_2 - k - 1}(\alpha)$ is the upper (100α)th percentile of the F distribution with numerator degrees of freedom equal to *k* and denominator degrees of freedom equal to $(n_1 + n_2 - k - 1)$, and *k* is the total number of outcome measures. s_{ii} is the sample variance for the *i*-th outcome measure, pooled over the two populations:

$$s_{ii} = \frac{(n_1 - 1)s_{1i} + (n_2 - 1)s_{2i}}{n_1 + n_2 - 2} \qquad (3)$$

where s_{1i} and s_{2i} are the sample variances for the *i*-th outcome measure for the two populations.

APPENDIX 2: Formula for Linear Correlation Test

Pocock et al. [13] give the following formula for the linear combination test for outcome variables with unknown, but equivalent, variance-covariance matrix:

$$\frac{J'S^{-1}t}{(J'S^{-1}J)^{1/2}} \qquad (4)$$

where *J'* is a 1xk vector of all 1's (i.e., 1, 1, …, 1), *k* is the number of outcome measures, *S* is the estimate of the kxk correlation matrix for the *k* outcome measures, and *t* is the kx1 vector of univariate *t* statistics for the *k* outcome measures.

When there are only two outcome measures (i.e., $k = 2$), the numerator of formula (4) can be written simply as $[1 / (1 - r^2)] \times (t_1 \times [1 - r] + t_2 \times (-r + 1))$, where *r* is the estimated correlation between the two outcome variables, and t_1 and t_2 are the univariate *t* statistic values for the two outcome variables. The denominator is simply the square root of $[1 / (1 - r^2)] \times (2 - 2r)$.

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005
16. Statistical Inference for Proportions, April 2005
17. Reader Agreement Studies, May 2005
18. Correlation and Regression, July 2005
19. Survival Analysis, July 2005

Decision Analysis and Simulation Modeling for Evaluating Diagnostic Tests on the Basis of Patient Outcomes

Sylvia K. Plevritis[1]

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 21st article in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

Received November 17, 2004; accepted after revision November 23, 2004.

[1]Department of Radiology, LUCAS Center P267, Stanford University, Stanford, CA 94305. Address correspondence to S. K. Plevritis.

AJR 2005; 185:581–590

0361–803X/05/1853–581

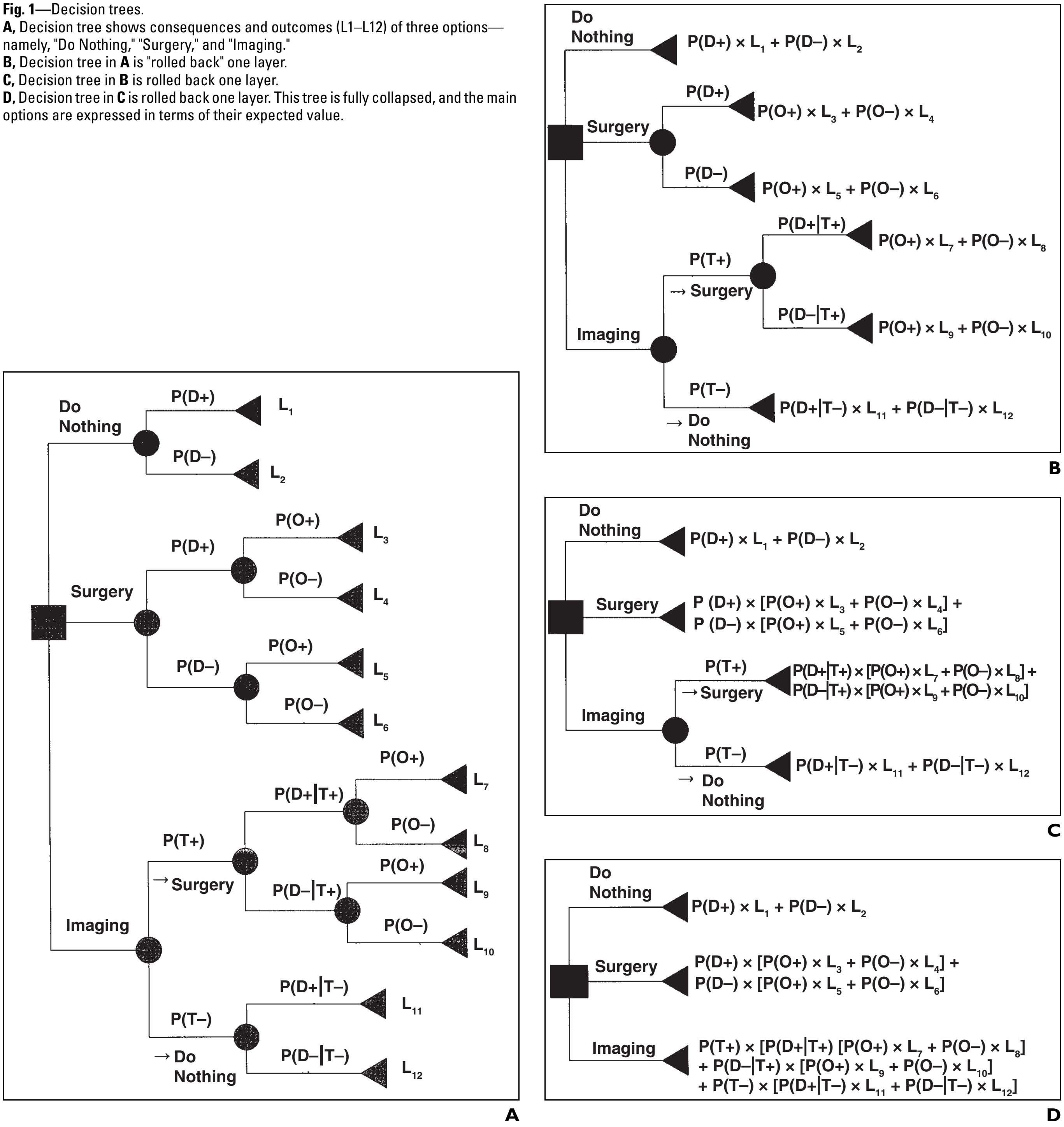
An imaging test with highest diagnostic accuracy is not necessarily the test of choice in clinical practice. The decision to order a diagnostic imaging test needs to be justified by its impact on downstream health outcomes. Decision analysis is a powerful tool for evaluating a diagnostic imaging test on the basis of long-term patient outcomes when only intermediate outcomes such as test sensitivity and specificity are known. The basic principles of decision analysis and “expected value” decision making for diagnostic testing are introduced. Markov modeling is shown to be a valuable method for linking intermediate to long-term health outcomes. The evaluation of Markov models by algebraic solutions, cohort simulation, and Monte Carlo simulation is discussed. Finally, cost-effectiveness analysis of diagnostic testing is briefly discussed as an example of decision analysis in which long-term health effects are measured both in life-years and costs.

The emergence of evidence-based medicine has handed radiologists the challenge of evaluating the impact of a diagnostic imaging test on a patient’s long-term outcome, often measured by overall survival and total health care expenditures [1]. This new challenge represents a significant departure from traditional evaluations of diagnostic examinations in which the main end points are intermediate ones—namely, test sensitivity and specificity. The shift from intermediate technology-specific to long-term patient-specific outcomes is being driven by the fact that a test with the highest diagnostic accuracy may not necessarily be the test of choice in clinical practice [2]. When making the decision to order a diagnostic imaging test, a clinician considers the health outcomes downstream from the imaging examination. For example, the health risks of interventions resulting from false-positive (FP) and false-negative (FN) findings should be compared with the health benefits associated with true-negative (TN) and true-positive (TP) findings. Increasingly, the cost of the diagnostic test, including the downstream costs generated as a result of imaging, are also factored in the decision-making process.

Few radiologists would argue against the importance of measuring the impact of diagnostic tests on long-term outcomes, but many are concerned with the feasibility of evaluating long-term outcomes through traditional clinical trials. Except when evaluating the impact of an imaging test for an acute state that is life-threatening, evaluating the impact of an imaging test in the adult population in terms of overall survival can require follow-up of 10 years or more. In children, even longer follow-up periods could be required. Long follow-up times compete with demands to diffuse promising technologies quickly and increase the risk of delaying technologic innovations. For a disease with a low risk of death, an economically unfeasible sample size may be required to detect a survival benefit due to diagnostic testing.

Linking the intermediate outcomes (such as TPs, TNs, FPs, and FNs) to long-term outcomes (such as survival) without requiring a clinical trial is sometimes possible. This link is often made when existing clinical data (usually collected for different purposes) can be extrapolated to address the problem of interest. Often the data are extrapolated through a number of assumptions that are formulated into a mathematic model in which the link between intermediate and long-term outcomes is expressed in terms of probabilistic events [3]. The Markov model, described later in this article, is an example of the methods commonly used in this extrapolation process. When reliable models can be generated, the opportunity arises to evaluate a variety of hypothetical clinical

Fig. 1—Decision trees.
A, Decision tree shows consequences and outcomes (L1–L12) of three options—namely, "Do Nothing," "Surgery," and "Imaging."
B, Decision tree in **A** is "rolled back" one layer.
C, Decision tree in **B** is rolled back one layer.
D, Decision tree in **C** is rolled back one layer. This tree is fully collapsed, and the main options are expressed in terms of their expected value.



paradigms that would not be economically feasible or practical to analyze experimentally via traditional clinical trials. The process of choosing from a number of hypothetical clinical paradigms by comparing them in terms of model-based probabilistic outcomes is often referred to as decision analysis.

This article will focus on the basic principles of decision analysis and "expected value" decision making. Emphasis is placed on evaluating diagnostic testing on the basis of long-term patient outcomes given only knowledge of the test sensitivity and specificity. Markov models will be briefly introduced because they are typically incorporated into decision analysis models to provide the link between intermediate and long-term outcomes. Cost-effectiveness analysis, which is a type of decision analysis in which the health effects and costs are tracked simultaneously, will also be briefly discussed. Finally, the major strength and weakness of decision analysis will be summarized.

Decision Analysis

Decision analysis is a deductive reasoning process that enables a decision maker to choose from a well-defined set of options on the basis of the systematic model-based analysis of all the probable outcomes [4–6]. Every outcome has a known probability of occurrence and a numeric value (i.e., life expectancy). The purpose of decision analysis is to quantify each option in terms of its expected (or average) value. A rational decision maker would choose the option that provides the greatest expected value. For example, if the outcome of the decision is measured in terms of life expectancy, the decision maker would choose to maximize the expected value; if the outcome is measured in costs, the decision maker would choose to minimize the expected value.

The critical components underlying decision analysis include clarifying the decision and the value used to measure the success of the decision, identifying the options, formulating every possible outcome from every possible decision, assigning a probability to each possible chance event, and assigning a value to each possible outcome. Once these components are determined, computing the expected values for each option can be straightforward.

Consider a generic clinical problem that involves the optional use of a diagnostic imaging test:

A patient presents with clinical symptoms of a life-threatening disease that requires surgery. What should the clinician recommend, knowing that surgery carries a risk of death? If the patient's probability of having the disease is low relative to the risk of surgery-related fatality, the clinician may recommend "Do Nothing" to avoid the risk of death due to the surgery. If the patient's probability of disease is high, the clinician may recommend "Surgery" immediately on the premise that the risk of death from the disease is higher than the risk of death from surgery. Now suppose a diagnostic imaging test becomes available with known sensitivity and specificity for the disease of interest. The clinician may choose to order the imaging test and then recommend Do Nothing if the imaging finding is negative or Surgery if the imaging finding is positive. Should the clinician order the imaging test, or make the recommendation of Do Nothing or Surgery, without the findings from the imaging test?

To answer this question, the decision maker, who is the clinician in the our example, needs to define a value on which to base the decision. If the value is life expectancy, then the clinician would want to know if ordering the diagnostic test will increase the patient's life expectancy. Decision analysis reveals how the patient's life expectancy depends on the choice made by the decision maker and events that are governed by chance—namely, the patient's probability of having the disease before getting the imaging results (i.e., the pretest probability or disease prevalence), the patient's life expectancy if the disease is present and untreated, the expected survival gain from a successful surgery, the risk of death from surgery, and the sensitivity and specificity of the imaging test.

Decision Trees

Decision analysis is aided by the use of a decision tree [4, 5]. A decision tree is a graphic model that represents the consequences for each possible decision through a sequence of decision and chance events [7]. A decision tree is constructed with three types of nodes: decision nodes, chance nodes, and terminal nodes, commonly represented as squares, circles, and triangles, respectively. A decision node is a branching point in the tree where several options are available to the decision maker for his or her choosing. A chance node is a branching point from which several outcomes are possible, but they are not available to the decision maker for his or her choosing. Instead, at a chance node, the outcome is randomly drawn from a set of possible outcomes (this is equivalent to saying that they are governed by chance). A chance event could be, for example, that a patient presenting with symptoms for a disease actually has the disease. At a chance node, every outcome is assigned a probability of occurrence, which is often estimated from a clinical trial or observational data. The decision tree is typically drawn by starting at the far left with a decision node and continuing from left to right through a sequence of decision and chance nodes. Every possible pathway through the decision tree ends at the far right with a terminal node. Every terminal node is assigned a value.

A simple decision tree associated with the clinical problem described previously is given in Figure 1A. This decision tree has one decision node that illustrates three possible options: Do Nothing, meaning the patient is sent home; Surgery, meaning the patient undergoes immediate surgery; and Imaging, meaning the patient undergoes a diagnostic imaging test and then surgery if the imaging findings are positive.

The Do Nothing option yields two chance events: the patient has the disease, with probability P(D+), and is assigned a life expectancy of L1 years; or the patient does not have the disease, with probability P(D–), and is assigned a life expectancy of L2 years.

The Surgery options yields four chance events: the patient has the disease with probability P(D+), experiences fatal surgical complications with probability P(O+), and has a life expectancy of L3 years; the patient has the disease with P(D+), undergoes successful surgery with probability P(O–), and has a life expectancy of L4 years; the patient does not have the disease with probability P(D–), experiences complications due to surgery with probability P(O+), and has a life expectancy of L5 years; and the patient does not have the disease with probability P(D–), does not experience complications due to surgery with probability P(O–), and has a life expectancy of L6 years.

The Imaging option yields six chance events. In four chance events, the imaging findings are positive, with probability P(T+), and the patient undergoes surgery; then the patient who has the disease, with conditional probability P(D+|T+), experiences fatal surgical complications with probability P(O+) and is assigned a life expectancy of L7 years; the patient who has the disease, with conditional probability P(D+|T+), has a successful surgery, with probability P(O–), and is assigned a life expectancy of L8 years; the patient who does not have the disease, with conditional probability P(D–|T+), but experiences fatal surgical complications, with probability P(O+), and is assigned a life expectancy of L9 years; the patient who does not have the disease, with conditional probability P(D–|T+), has successful surgery, with probability P(O–), and is assigned a life expectancy of L10 years. In two chance events the imaging findings are negative, with probability P(T–); and either the patient has the disease, with probability P(D+|T–), and is assigned a life expectancy of L11 years; or the patient does not have the disease, with probability P(D–|T–), and is assigned a life expectancy of L12 years.

All the probabilities populating the decision tree are summarized in Table 1.

To evaluate the Imaging option, the probabilities P(D+|T+), P(D–|T+), P(D+|T–),

TABLE 1: Probability Notation and Base Case Values

Notation	Meaning	Base Case Value
P(D+)	Probability disease is present = prevalence = pretest probability	**0.10**
P(D–)	Probability disease is not present = [1–P(D+)]	0.90
P(T+|D+)	Sensitivity = probability of a positive test given disease is present = probability of a true-positive	**0.90**
P(T–|D+)	(1 – sensitivity) = probability of negative test given that the disease is present = probability of a false-negative	0.10
P(T–|D–)	Specificity = probability of negative test given that the disease is present = probability of a true-negative	**0.80**
P(T+|D–)	(1 – specificity) = probability of positive test given that the disease is not present = probability of a false-positive	0.20
P(D+|T+)	Probability disease is present given that test is positive = PPV	0.33
P(D–|T+)	Probability disease is not present given that test is positive = [1 – PPV]	0.67
P(D+|T–)	Probability disease is present given that test is negative = [1 – NPV]	0.01
P(D–|T–)	Probability disease is not present given that test is negative = NPV	0.99
P(T+)	Probability of a positive test	0.27
P(T–)	Probability of a negative test = [1 – P(T+)]	0.73
P(O+)	Probability of surgery-related death	**0.05**
P(O–)	Probability of successful surgery = [1 – P(O+)]	0.95

Note—Bold base case values are assigned, nonbold values are derived. PPV = positive predictive value, NPV = negative predictive value.

P(D–|T–), P(T+), and P(T–) must be evaluated. The probability that the patient has the disease given a positive imaging finding, denoted as P(D+|T+), is commonly referred to as the positive predictive value of the test. The probability that the patient does not have the disease given a negative imaging finding, denoted as P(D–|T–), is commonly referred to as the negative predictive value of the test. These probabilities can be derived from the pretest probability of the disease and the sensitivity and specificity of the test using Bayes' theorem:

$$P(D+|T+) = P(T+|D+)\ P(D+) / P(T+)$$
$$P(D-|T-) = P(T-|D-)\ P(D-) / P(T-),$$

where P(T+|D+) is defined as the sensitivity and P(T–|D–) is defined as the specificity [8]. The probability of a positive and negative test can be computed as:

$$P(T+) = P(T+|D+)\ P(D+) + P(T+|D-)\ P(D-)$$
$$P(T-) = 1-P(T+).$$

Therefore, incorporating an imaging test into the decision tree simply requires knowledge of the test's sensitivity and specificity and the patient's pretest probability of disease.

Expected Value Decision Making

Decision analysis operates on the principle that a rational choice from a set of options is the one with the greatest expected value [4, 5]. It is possible that a "good" decision leads to a "bad" outcome because chance is involved. The likelihood of a bad outcome is minimized when the decision is the one with the greatest expected value. This principle is often referred to as Bayes' Decision Rule and is credited to the Reverend Thomas Bayes, an 18th century minister, philosopher, and mathematician who formulated Bayes' theorem.

Computing the expected value of each option is accomplished by "rolling back" or "averaging" the decision tree. The process of rolling back the decision tree for the clinical example illustrated in Figure 1A is shown in Figures 1B–1D. In each figure we progressively roll back the right-most layer of terminal branches to their originating node and assign an expected value to that node, in effect turning the originating node into a new terminal node. If the originating node is a chance node, then the expected value is calculated as the weighted average of the expected values of its possible outcomes, where the weights are the probabilities that each outcome will occur. If the originating node is a decision node, the outcome is the one with the best expected value. This process is continued until the decision node at the far left-most part of tree is the only remaining decision node in the tree. The decision tree is said to be "fully collapsed." An example of a fully collapsed tree is given in Figure 1D.

For the Do Nothing option, the life expectancy is P(D+) × L1 + P(D–) × L2 years.

For the Surgery option, the life expectancy is P(D+) × [P(O+) × L3 + P(O–) × L4] + P(D–) × [P(O+) × L5 + P(O–) × L6].

For the Imaging option, the life expectancy is P(T+) × {P(D+|T+) × [P(O+) × L7 + P(O–) × L8] + P(D–|T+) × [P(O+) × L9 + P(O–) × L10]} + P(T–) × [P(D+|T–) × L11 + P(D–|T–) × L12].

To evaluate and compare the life expectancies for each of the three options, the probabilities and life expectancies L1 through L12 must be assigned. Consider the example in which a 60-year-old patient presents with symptoms indicative of a specified disease that has poor prognosis. Probability values for the chance events are given in Table 1. These values can be derived from the following three assumptions: patient's pretest probability for the disease is 0.10; the probability of surgery-related death is 0.05; and the diagnostic test has a sensitivity of 0.90 and specificity of 0.80. To compute the expected value of each option, we also need to assign a value to each possible outcome. Table 2 lists all the possible intermediate outcomes (column 1) and their associated life expectancies (column 6). For example, if the clinician recommends Do Nothing and the intermediate outcome is that the patient has the disease (D+), then the patient's life expectancy is L1 = 65 years. If the patient does not have the disease, his or her life expectancy is 80 years. If the patient experiences operative death, then his or her life expectancy is 60 years. We assume successful surgery is not curative but extends the patient's life expectancy to 72.5 years. Later we will show how these life expectancies can be estimated with a Markov model that links the intermediate health outcomes to overall survival.

TABLE 2: Transition Probabilities and Long-Term Outcomes

Option and Intermediate Outcome[a]	Transition Probabilities[b]				Long-Term Outcome Expressed as Life Expectancy (yr)[c]
	$p_1 = \mathrm{P}(\mathrm{DSD}_n \mid \mathrm{Alive}_{n-1})$	$p_2 = \mathrm{P}(\mathrm{DOC}_n \mid \mathrm{Alive}_{n-1})$	$p_3 = \mathrm{P}(\mathrm{DS}_n \mid \mathrm{Alive}_{n-1})$	$p_4 = \mathrm{P}(\mathrm{Alive}_n \mid \mathrm{Alive}_{n-1})$	
Do Nothing					
D+	0.15	0.05	—	0.80	L1 = 65
D–	—	0.05	—	0.95	L2 = 80
Surgery					
D+, O+	—	—	1.00	—	L3 = 60
D+, O–	0.03	0.05	—	0.92	L4 = 72.5
D–, O+	—	—	—	1.00	L5 = 60
D–, O–	—	0.05	—	0.95	L6 = 80
Imaging					
T+, D+, O+	—	—	1.00	—	L7 = 60
T+, D+, O–	0.03	0.05	—	0.92	L8 = 72.5
T+, D–, O+	—	—	—	1.00	L9 = 60
T+, D–, O–	—	0.05	—	0.95	L10 = 80
T–, D+	0.15	0.05	—	0.80	L11 = 65
T–, D–	—	0.05	—	0.95	L12 = 80

Note—D+ indicates patient has disease, D– indicates patient does not have disease, O+ indicates patient experienced operative death, O– indicates patient underwent successful surgery, T+ indicates positive imaging finding, T– indicates negative imaging finding.

[a] Intermediate outcomes depend on options illustrated in decision tree in Figure 1A.

[b] Probabilities correspond to Markov model in Figure 5. DSD$_n$, DOC$_n$, and DS$_n$ indicate patient enters Disease-Specific Death, Death from Other Causes, Death from Surgery, respectively, at cycle number n. Alive$_n$ and Alive$_{n-1}$ indicate patient is alive at cycle n and cycle n–1, respectively. Cycle period is 1 year. Dash (—) indicates 0.

[c] Output of Markov model when the patient is 60 years old at initiation.

Given the probabilities assigned to the chance events (Table 1, column 3) and the expected values of each possible outcome (Table 2, column 6), the life expectancies for the options Do Nothing, Surgery, and Imaging are 78.5 years, 78.3 years, and 78.9 years, respectively. Because the maximum life expectancy is associated with the Imaging option, the clinician would recommend surgery on the basis of positive imaging findings.

Factors that were not considered in the decision process that could change the clinician's recommendation include the invasiveness of the imaging test; quality of life while living with the symptoms; utilities derived from TP, TN, FP, and FN findings on imaging; and the possibility of delaying the surgery. However, the general ideas presented here can be extended to include these factors. In addition, this general approach can be used to consider more complex decisions that may involve more than one imaging test ordered sequentially or in parallel.

Sensitivity Analysis

Sensitivity analysis is a necessary component of decision analysis that is used to evaluate the robustness of the decision to variations in model assumptions. In decision trees, probabilities of chance events and the values at terminal nodes may not be known. Under these circumstances, the values assigned may be reflective of an expert's best guess. The possibility exists that by varying the dubious model inputs, the expected values will not be affected greatly or will be affected but not enough to change the ranking of the options in order of expected value. Under either of these scenarios, a decision maker would be more confident in implementing the option with the greatest expected value. However, when changing an input affects the ranking of the options, the decision maker would be less certain about proceeding without clarifying the value of that input.

N-way (or multivariate) sensitivity analyses refer to the process of varying *N* parameters in a model simultaneously while all other parameters remain constant. The most simple and common example is one-way (or univariate) sensitivity analysis, in which one model parameter is varied in a range between an upper and lower bound, while all the other parameters are kept constant. A series of one-way sensitivity analyses is the easiest way to identify which parameters have the strongest effect on the optimal decision. For the example just given, a one-way sensitivity analysis on the pretest probability is shown in Figure 2 using the remaining parameters in Tables 1 and 2. For P(D+) less than or equal to 0.03, the Do Nothing option has the greatest life expectancy; for P(D+) greater than 0.03 but less than 0.54, the Imaging option has the greatest life expectancy; and for P(D+) greater than or equal to 0.54, the Surgery option has the greatest life expectancy. The point at which the decision shifts from one alternative to another is often referred to as the crossover point or the threshold.

Although a one-way sensitivity analysis is computationally easy, the outcomes may not be representative of realistic clinical situations. For example, changing test sensitivity without changing test specificity is usually not possible. In a two-way sensitivity analysis, two parameters, such as sensitivity and specificity, are varied at the same time, preferably choosing paired values of sensitivity and specificity along a receiver operating characteristic (ROC) curve [9]. A two-way sensitivity analysis on test sensitivity and test specificity is shown in Figure 3. Here the sensitivity and specificity are not varied along an ROC curve. Instead, the sensitivity was varied continuously from 0 to 1, and for each value of the sensitivity, the life expectancy was evaluated at four discrete values of specificity: namely, 0.3, 0.5, 0.7, and 0.9. The op-

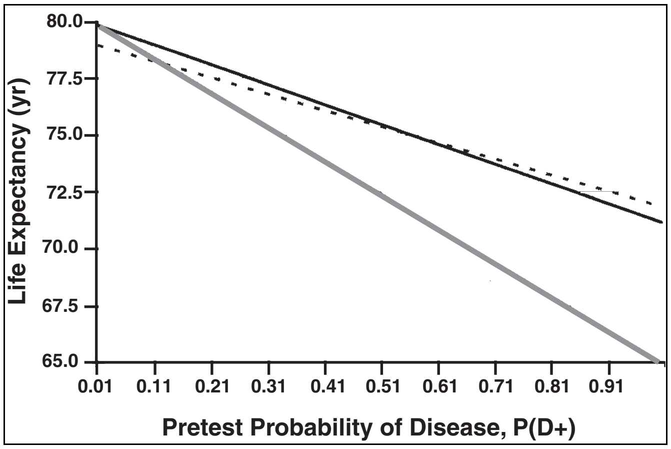


Fig. 2—One-way sensitivity analysis shows impact of changes in pretest probability of disease on life expectancy for three options, "Do Nothing" (*gray line*), "Surgery" (*dotted line*), and "Imaging" (*black line*).

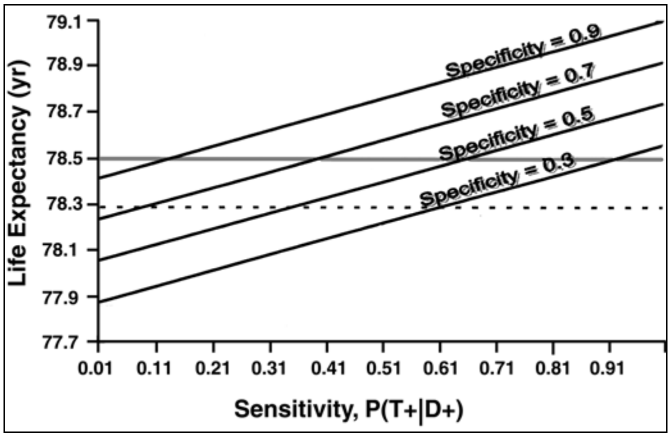


Fig. 3—Two-way sensitivity analysis shows impact of changes in test sensitivity and test specificity on life expectancy for "Imaging" (*black lines*) option. Life expectancies for "Do Nothing" (*gray line*) and "Surgery" (*dotted line*) options are included for comparison. Do Nothing option dominates Surgery option.

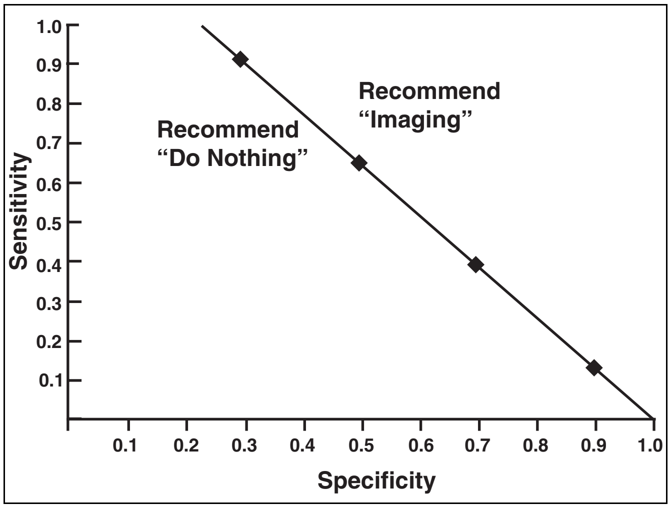


Fig. 4—Two-way sensitivity analysis illustrates optimal decision as a function of sensitivity and specificity of diagnostic test. "Imaging" option would be recommended for values above line, and "Do Nothing" option would be recommended for values below.

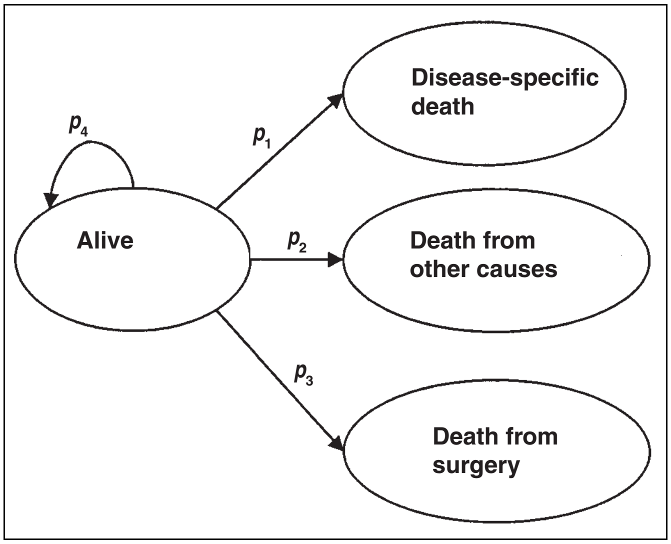


Fig. 5—Markov model with four health states ("Alive," "Disease-Specific Death," "Death from Other Causes," "Death from Surgery") and four transition probabilities (p_1, p_2, p_3, and p_4).

timal decision is either Do Nothing or Imaging, depending on the test's sensitivity and specificity. The Do Nothing option dominates the Surgery option, meaning that the optimal choice between Do Nothing and Surgery would be Do Nothing, as was observed in Figure 2 for a pretest probability of 0.1. The Imaging option dominates the Surgery option under the following conditions: when the specificity is 0.9 and the sensitivity is greater than 0.13, when the specificity is 0.7 and the sensitivity is greater than 0.4, when the specificity is 0.5 and the sensitivity is greater than 0.65, and when the specificity is 0.3 and the sensitivity is greater than 0.91. The optimal decision in "ROC Space" (i.e., for all values of sensitivity and specificity) is shown in Figure 4. Once the specificity is less than 0.24, the Do Nothing option is recommended for all values of sensitivity.

Markov Models for Estimating Life Expectancy

Markov models are commonly used in medical decision analysis for estimating life expectancy [10, 11]. In the previous example, the life expectancy for every possible outcome was known, but this information is usually not available. As discussed in the introduction, the challenge in basing decisions on maximizing life-years lies in finding a model that links the known intermediate health states to survival. A Markov model may be an appropriate tool for establishing this link when it is possible to represent a patient's life history from a known intermediate health to death through a series of transitions among a finite set of health states that have been observed elsewhere.

A simple Markov model for the clinical example described is composed of four health states: "Alive," "Disease-Specific Death," "Death from Other Causes," and "Death from Surgery." This model is shown in Figure 5. Each oval represents a health state. The arrows represent the possibility of transition from one state to another. The arrow that points back into the health state Alive indicates that the patient can remain in the health state Alive after a given cycle. Transitions between health states occur in a designated time period, known as the cycle period of the model. The cycle period for chronic diseases is typically 1 year, whereas the cycle period for acute diseases is often shorter—that is, months or even days. The probability that the patient will move from one health state to another in a given cycle is referred to as the transition probability. The life expectancy is the average length of time spent in all health states other than death.

The transition probabilities for the Markov model shown in Figure 5 are as follows:

$$p_1 = \text{P(Disease-Specific Death at cycle number } n \mid \text{Alive at cycle number } n-1\text{)},$$

$$p_2 = \text{P(Death from Other Causes at cycle number } n \mid \text{Alive at cycle number } n-1\text{)},$$

$$p_3 = \text{P(Death from Surgery at cycle number } n \mid \text{Alive at cycle number } n-1\text{), and}$$

$$p_4 = \text{P(Alive at cycle number } n \mid \text{Alive at cycle number } n-1\text{)},$$

$$\text{where } p_1 + p_2 + p_3 + p_4 = 1.$$

Note that the health state at cycle number n is conditioned on the health state at cycle number n–1 and is independent of the health state before cycle number n–1. This property is the defining property of Markov models of this type, which are referred to as Markov chain models.

The transition probabilities for the example just described are given in Table 2, assuming a cycle period of 1 year. All these transition probabilities can be derived from the following three assumptions: if the patient has the disease, the probability of transitioning from Alive to Disease-Specific Death in 1 year's time is 0.15 if the patient does not undergo surgery and 0.03 if the patient undergoes successful surgery; the probability of transitioning from Alive to Death from Other Causes in 1 year's time is 0.05 if the patient does not experience surgery-related death; and surgery-related death is immediate.

Once the health states, allowed transitions between health states, and transition probabilities are identified, the life expectancy can be calculated using an algebraic solution, a cohort simulation, or a Monte Carlo simulation. All three approaches will be illustrated for estimating the life expectancy L1 in Figure 1A, where a 60-year-old patient presents with clinical symptoms, the decision is made to Do Nothing, and the patient actually has the disease (D+). In this case, $p_1 = 0.15$, $p_2 = 0.05$, $p_3 = 0$, and $p_4 = 0.80$, as shown in Table 2.

Algebraic Solution

If the transition probabilities are constant over time, then a closed-form algebraic solution exists for estimating the life expectancy. In the simple example above, the patient's life expectancy L1 is calculated as follows:

$$LE = \text{present age} + 1 / (p_1 + p_2 + p_3) = 60 + 1 / (0.15 + 0.05 + 0) = 65 \text{ years.}$$

In more complex Markov chain models with numerous transient, recurrent, and absorbing states, a matrix formalism may be necessary to evaluate the model using the closed-form, algebraic approach.

Cohort Simulation

If the transition probabilities are not constant over time, simulating the outcomes of a cohort of patients is commonly implemented. This simulation process is initiated by distributing a cohort among the health states. For the above example, the entire cohort begins in the Alive state. At each cycle the cohort is redistributed among the states, depending on the transition probabilities. Markov cohort simulation for estimating the life expectancy L1 is illustrated in Table 3. The initial cohort size is 10,000. At the start of the simulation, the 10,000 patients are in the Alive state. By the end of the first cycle, $p_1 \times 10{,}000 = 0.15 \times 10{,}000 = 1{,}500$ patients enter Disease-Specific Death and $p_2 \times 10{,}000 = 0.05 \times 10{,}000 = 500$ patients enter Death from Other Causes, leaving 10,000 – 1,500 – 500 = 8,000 patients in the Alive state for the start of the second cycle. By the end of the second cycle, an additional $p_1 \times 8{,}000 = 0.15 \times 8{,}000 = 1{,}200$ patients enter the Disease-Specific Death, and $p_2 \times 8{,}000 = 0.05 \times 8{,}000 = 400$ patients enter Death from Other Causes, leaving 8,000 – 1,200 – 400 = 6,400 in the Alive state. The cumulative number of patients in each of the three states for the first 40 cycles of the Markov process is shown in the Table 3. Each row totals 10,000 patients. Life expectancy is calculated as the average amount of time a patient is in the Alive state. For this example, a 60-year-old patient remains in the Alive state for 20 years on average, making his or her life expectancy L1 = 60 + 20 = 80 years.

Monte Carlo Simulation

If complex dependencies exist in the state transition model, an intensive computer simulation procedure called Monte Carlo simulation may be needed to compute the life expectancy. In Monte Carlo simulation, patients traverse the health states one at a time, with a random number generator (RNG) determining what happens to an individual at each cycle of the process. An RNG is a computer algorithm that produces sequences of numbers that on aggregate have a specified probability distribution and individually possess the appearance of randomness.

To estimate L1 in the above example via Monte Carlo simulation, an RNG samples a uniform distribution from 0 to 1. When the RNG produces a number in the range 0–0.05, the patient is assigned to Disease from Other Causes. This will happen 5% of the time, which corresponds to the transition probability from Alive to Disease from Other Causes. When the RNG produces a number greater than 0.05 but less than or equal to 0.20, the patient is assigned to Disease-Specific Death. This will happen 15% of the time, which corresponds to the transition probability from Alive to Disease-Specific Death. Finally, when the RNG produces a number greater than 0.2 but less than or equal to 1.0, the patient remains in the Alive state, which will happen 80% of the time. In this simple example, only one random number needs to be generated at every cycle. Once the patient enters a death-related state, the life history of that patient is terminated and a new run begins that traces the life history of the next patient. The process is repeated until a large number of runs (typically 10,000) are performed. There is no formula specifying the exact number of runs needed, but the number should increase with

TABLE 3: Markov Cohort Simulation: Distribution of a 10,000-Patient Cohort at End of Each 1-Year Cycle

			Cumulative No. in State	
Cycle No.	Age of Alive Population	No. in Alive State	Disease-Specific Death	Death from Other Causes
0	60	10,000.00	0	0
1	61	8,000.00	1,500.00	500.00
2	62	6,400.00	2,700.00	900.00
3	63	5,120.00	3,660.00	1,220.00
4	64	4,096.00	4,428.00	1,476.00
5	65	3,276.80	5,042.40	1,680.80
6	66	2,621.44	5,533.92	1,844.64
7	67	2,097.15	5,927.14	1,975.71
8	68	1,677.72	6,241.71	2,080.57
9	69	1,342.18	6,493.37	2,164.46
10	70	1,073.74	6,694.69	2,231.56
11	71	858.99	6,855.75	2,285.25
12	72	687.19	6,984.60	2,328.20
13	73	549.76	7,087.68	2,362.56
14	74	439.80	7,170.15	2,390.05
15	75	351.84	7,236.12	2,412.04
16	76	281.47	7,288.89	2,429.63
17	77	225.18	7,331.12	2,443.71
18	78	180.14	7,364.89	2,454.96
19	79	144.12	7,391.91	2,463.97
20	80	115.29	7,413.53	2,471.18
21	81	92.23	7,430.82	2,476.94
22	82	73.79	7,444.66	2,481.55
23	83	59.03	7,455.73	2,485.24
24	84	47.22	7,464.58	2,488.19
25	85	37.78	7,471.67	2,490.56
26	86	30.22	7,477.33	2,492.44
27	87	24.18	7,481.87	2,493.96
28	88	19.34	7,485.49	2,495.16
29	89	15.47	7,488.39	2,496.13
30	90	12.38	7,490.72	2,496.91
31	91	9.90	7,492.57	2,497.52
32	92	7.92	7,494.06	2,498.02
33	93	6.34	7,495.25	2,498.42
34	94	5.07	7,496.20	2,498.73
35	95	4.06	7,496.96	2,498.99
36	96	3.25	7,497.57	2,499.19
37	97	2.60	7,498.05	2,499.35
38	98	2.08	7,498.44	2,499.48
39	99	1.66	7,498.75	2,499.58
40	100	1.33	7,499.00	2,499.67
41	101	1.06	7,499.20	2,499.73
42	102	0.85	7,499.36	2,499.79
43	103	0.68	7,499.49	2,499.83
44	104	0.54	7,499.59	2,499.86
45	105	0.44	7,499.67	2,499.89

Note—Life expectancy is computed as average amount of time a patient remains in Alive state.

TABLE 4: Monte Carlo Simulation of a Markov Process

	Run 1		Run 2		Run 3		Run 4		Run 5		…	Run 10,000	
Cycle No.[a]	RNG	State	RNG	State	RNG	State	RNG	State	RNG	State		RNG	State
1	0.63	Alive	0.75	Alive	0.23	Alive	0.93	Alive	0.55	Alive		0.92	Alive
2	0.83	Alive	0.91	Alive	0.32	Alive	0.59	Alive	0.15	DSD		0.64	Alive
3	0.56	Alive	0.03	DOC	0.69	Alive	0.30	Alive				0.48	Alive
4	0.20	Alive			0.33	Alive	0.60	Alive				0.82	Alive
5	0.09	DDS			0.63	Alive	0.19	DSD				0.61	Alive
6					0.30	Alive						0.86	Alive
7					0.39	Alive						0.86	Alive
8					0.83	Alive						0.07	DSD
9					0.52	Alive							
10					0.68	Alive							
11					0.27	Alive							
12					0.14	DSD							
…													

Note—Each run represents life history of a single individual. At each cycle in a given run, a random number generator (RNG) outputs a number from a uniform distribution between 0 and 1. If RNG produces a number ≤ 0.05, patient is assigned to "Death from Other Causes" (DOC); > 0.05 and ≤ 0.20, "Disease-Specific Death" (DSD); or > 0.20 and ≤ 1,"Alive." Once patient enters a death-related state, life history of that patient is terminated and a new run begins until maximum number of runs is completed.

[a]Cycle period = 1 year.

the complexity of the model to reduce simulation variability in the result.

Six sample runs of a Monte Carlo simulation are shown in Table 4. In each run, the patient is initiated in the Alive state at age 60 and ages 1 year in every cycle. Table 4 shows that in run 1, the patient dies of the disease after 5 years (at age 65); in run 2, the patient dies of other causes after 3 years (at age 63). The runs are repeated 10,000 times. The life expectancy is the average age at death. A valuable output of Monte Carlo simulation is a histogram of age at death, so that measures of variability in the life expectancy are easy to calculate [12, 13].

Markov models have much broader applicability than estimating life expectancy. They are used in a variety of fields to represent processes that evolve over time in a probabilistic manner. The article by Kuntz and Weinstein [14] is recommended further reading on Markov modeling in medical decision analysis.

Cost-Effectiveness Analysis

Cost-effectiveness analysis is a type of decision analysis in which both health and economic outcomes are considered simultaneously in making a decision [15, 16]. The decision analysis example described previously focused on maximizing life expectancy (LE). Although maximizing life expectancy is a reasonable value, it may not necessarily be the basis for a preferred decision. If the difference in life expectancy between an existing clinical protocol and a new clinical protocol is small but the difference in costs is large, it may be more prudent to follow the existing protocol and invest health care dollars in another clinical problem for which the incremental life expectancy is higher for the same health care expenditures.

In cost-effectiveness analysis, the expected value is reported as the marginal cost per year of life saved (MCYLS) [17]. When the decision tree is rolled back, the average cost is evaluated in parallel with the life expectancy. Dominant options are ranked in terms of incremental cost-effectiveness ratios.

The value of diagnostic testing is put to the greatest challenge in cost-effectiveness analysis. Often diagnostic testing increases both life expectancy and health care costs. The application of cost-effectiveness analysis to diagnostic testing is introduced in an article by Fryback [18] and discussed in more detail in an article by Singer and Applegate [19]. More general discussions on the role of cost-effectiveness analysis and recommendations for reporting results are found in other articles [20–22].

Summary

Decision analysis is a multifaceted concept. Underlying the decision analytic process is clarification of the decision and values for making a good decision, integration of data from multiple data sources, and mathematic modeling. The necessary steps for any decision analysis are summarized in Appendix 1.

The major strength of decision analysis is that the process offers an explicit and systematic approach to decision making based on uncertainty. The major weakness of decision analysis lies with the decision analyst who uses data to populate a model without understanding the biases in the data and therefore does not fully explore their impact on the decision [23, 24]. This problem is minimized when the decision analyst is fully knowledgeable of both clinical domain–specific and methodology-specific issues.

This article has focused the basic ideas of decision analysis toward the problem of evaluating a diagnostic imaging test on the basis of long-term patient outcomes when only the test's sensitivity and specificity are known. Markov models were introduced as means of linking intermediate to long-term outputs. Even when the inputs and structure of the decision analysis model may be incompletely supported by data, the decision analysis process itself can be valuable in identifying important areas of uncertainty and directing the investment of resources toward acquiring information needed to address the question of interest. Such analyses may be warranted before resources are committed to large-scale, costly clinical trials.

References

1. Thornbury JR. Why should radiologists be interested in technology assessment and outcomes research? *AJR* 1994; 163:1027–1030
2. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991; 11:88–94
3. Ramsey SD, McIntosh M, Etzioni R, Urban N. Simulation modeling of outcomes and cost effectiveness. *Hematol Oncol Clin North Am* 2000; 14:925–938
4. Weinstein MC, Fineberg HV, Elstein AS, et al. *Clinical decision analysis*. Philadelphia, PA: Saunders, 1980
5. Pauker SG, Kassirer JP. Decision analysis. *N Engl J Med* 1987; 316:250–258
6. Sox H, Blatt MA, Higgins MC, Marton KI. *Medical decision making*. Boston, MA: Butterworths, 1988
7. Fineberg HV. Decision trees: construction, uses, and limits. *Bull Cancer* 1980; 67:395–404
8. Schulzer M. Diagnostic tests: a statistical review. *Muscle Nerve* 1994; 17:815–819
9. Metz CE. Basic principles of ROC analysis. *Semin*

Nucl Med 1978; 8:283–298

10. Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Med Decis Making* 1993; 13:322–338
11. Beck JR, Pauker SG. The Markov process in medical prognosis. *Med Decis Making* 1983; 3:419–458
12. Tambour M, Zethraeus N. Bootstrap confidence intervals for cost-effectiveness ratios: some simulation results. *Health Econ* 1998; 7:143–147
13. Critchfield GC, Willard KE. Probabilistic analysis of decision trees using Monte Carlo simulation. *Med Decis Making* 1986; 6:85–92
14. Kuntz KM, Weinstein MC. Life expectancy biases in clinical decision modeling. *Med Decis Making* 1995; 15:158–169
15. Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-effectiveness in health and medicine*. Oxford, England: Oxford University Press, 1996
16. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med* 1977; 29:716–721
17. Detsky AS, Naglie IG. A clinician's guide to cost-effectiveness analysis. *Ann Intern Med* 1990; 113:147–154
18. Fryback DG. Technology evaluation: applying cost-effectiveness analysis for health technology assessment. *Decisions in Imaging Economics* 1990; 3:4–9
19. Singer ME, Applegate KE. Cost-effectiveness analysis in radiology. *Radiology* 2001; 219:611–620
20. Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. Panel on Cost-Effectiveness in Health and Medicine. *JAMA* 1996; 276:1172–1177
21. Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA* 1996; 276:1253–1258
22. Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. Panel on Cost-Effectiveness in Health and Medicine. *JAMA* 1996; 276:1339–1341
23. Sheldon TA. Problems of using modelling in the economic evaluation of health care. *Health Econ* 1996; 5:1–11
24. Buxton MJ, Drummond MF, Van Hout BA, et al. Modelling in economic evaluation: an unavoidable fact of life. *Health Econ* 1997; 6:217–227

APPENDIX 1: Necessary Steps for Any Decision Analysis

1. Identify the clinical problem and targeted patient population.
2. Identify clinical options.
3. Identify the decision maker.
4. Identify the outcomes associated with each clinical option.
5. Identify the value on which the decision will be based.
6. Assign a value to each terminal branch of the decision tree. This step may include additional modeling, such as Markov models, to link known intermediate health states to long-term outcomes.
7. Assign probabilities to each chance event.
8. Compute the expected value of each decision by averaging out the decision tree.
9. Perform a sensitivity analysis.
10. Report the model assumptions, inputs, and results.

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005
16. Statistical Inference for Proportions, April 2005
17. Reader Agreement Studies, May 2005
18. Correlation and Regression, July 2005
19. Survival Analysis, July 2005
20. Multivariate Statistical Methods, August 2005

Radiology Cost and Outcomes Studies: Standard Practice and Emerging Methods

William Hollingworth[1]

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 22nd and final article in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology.* The series has been designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

DOI:10.2214/AJR.04.1780

Received November 17, 2004; accepted after revision November 23, 2004.

[1]Department of Radiology, University of Washington, Harborview Medical Center, 325 Ninth Ave., Box 359960, Seattle, WA 98104. Address correspondence to W. Hollingworth (willh@u.washington.edu).

AJR 2005;185:833–839

0361–803X/05/1854–833

Cost and outcomes research has become an integral part of radiology since the pioneering randomized controlled trials (RCTs) of radiographic screening for lung and breast cancer in the 1960s and 1970s [1, 2]. The impetus for radiologists to become involved in this technology assessment process is likely to continue to increase in the foreseeable future. Medical expenditures are at an all time high; in 2002 the United States spent just under 14% of its gross domestic product on health care. This equates to about $4,900 per capita annually, more than double the amount spent by other industrialized countries such as Sweden, Australia, and Japan [3]. The source of the increase in spending is certainly multifactorial, including the need to provide care to an aging population, which places ever higher expectations on the capabilities of medicine. However, the public and health care professionals alike perceive that medical technologies also drive expenditure. A survey of health economists revealed that 81% identified technologic change as the primary reason for the increase in health sector spending [4]. Purchases of expensive diagnostic imaging equipment are particularly visible; 68% of respondents to a U.S. community survey thought that the increase in diagnostic procedures played a large or very large role in increasing health care costs [5].

The introduction of noninvasive angiography using MRI or MDCT to replace catheter angiography provides one of several examples in which diagnostic imaging advances have the potential to simultaneously reduce costs and benefit patients [6, 7]. However, this will not always be the case. Newer imaging technology may increase costs for any of the following reasons: if it is an adjunct rather than a replacement for existing imaging methods; if it has a higher unit cost than existing imaging; or if, by making the imaging process more convenient, the threshold for imaging is lowered [8]. In these situations the onus will continue to be on radiologists to provide evidence that newer imaging techniques improve diagnostic and therapeutic decision making and thereby benefit patients.

This article has three objectives: first, to identify factors that, in combination, make radiology cost and outcomes studies unique; second, to review standard methods for measuring the cost and outcomes of diagnostic imaging; and third, to describe emerging methods that will help radiologists conduct and interpret cost and outcomes studies in future years.

What Are the Factors That Make Cost and Outcomes Research in Radiology Unique?

The Gap Between Diagnosis and Outcome

The fundamental distinction between outcomes research in radiology and other areas of medicine, such as surgery and pharmaceutics, is the distance between cause and effect. That is, the chain of events that separates the immediate aim of radiology, which is to make an accurate diagnosis, from the ultimate goal, which is to improve patient health and life expectancy at an affordable cost. The links in this chain have been formalized in the hierarchy originally developed by Fineberg et al. [9] and adapted by others [10, 11]. The first two levels of this hierarchy depend on the capability of the imaging technology to depict normal and abnormal anatomy and function (level 1) and the ability of radiologists to use the images to make accurate diagnoses (level 2). Beyond these initial two levels, the value of diagnostic imaging is dictated by factors that are not under the control of radiology. The referring clinician must be convinced by the imaging results to change the working diagnosis (level 3) and therapy (level 4) for the patient. Effective therapeutic options must be available if the change in therapy is to benefit patients (level

5). Finally, the net cost of diagnosis and treatment must be justified by improvements in patients' health (level 6). Failure at any one of the latter four levels will undermine the value of even the most accurate diagnostic test.

The Size of the Study

One upshot of this hierarchy of events is that imaging, particularly when used to screen asymptomatic populations, is likely to directly benefit only a small subgroup of recipients. This is in contrast to therapeutic interventions, in which all patients have the potential to benefit. For example, in many breast cancer screening programs, fewer than 1% of mammograms result in a confirmed case of cancer [12]. The health of the remaining 99% of women is unlikely to be directly affected beyond reassurance provided by a negative result or anxiety raised by false-positive findings. Consequently, most studies of screening are large trials recruiting thousands of patients, or decision analyses based on hypothetical models of diagnostic accuracy and therapeutic effectiveness. Large trials are needed to detect with statistical accuracy health effects in the small proportion of the population with the disease.

The Intrinsic Value of Diagnostic Information

Even diagnostic imaging of symptomatic patients may not radically alter treatment for many recipients. For example, in a study comparing MRI and arthrography for patients with shoulder pain and suspected full-thickness rotator cuff tears, Blanchard et al. [13] found that preimaging management plans changed in 36% and 25% of patients, respectively. Although imaging may not always trigger a change in therapy, diagnostic information may still have intrinsic value. In 1994, Mushlin et al. [14] found that patients with suspected multiple sclerosis became less anxious after a positive MRI diagnosis, even though they faced a chronic disease with, at that time, few therapeutic options. A negative test result may also be beneficial if it reassures the patient that nothing is seriously wrong. However, this is not a predictable effect; indeed, in some patients, negative test findings can heighten anxiety about the cause of ongoing symptoms [15]. These intrinsic effects emphasize the importance of assessing patients' perceptions of their physical and mental health after imaging.

Standard Methods in Cost and Outcomes Research

The diverse nature of cost and outcomes research makes it difficult to be prescriptive in defining best practice. However, as research methods have evolved there have been a number of landmark publications that have defined a methodologic blueprint for research. The Consolidated Standards of Reporting Trials (CONSORT) statement provides a checklist of items considered essential for the clear presentation of RCT results [16]. Similar guidelines have been developed for nonrandomized studies [17], economic evaluations [18], and decision analysis models [19]. In addition, a number of excellent articles apply general cost and outcomes methods to radiology [20, 21].

The purpose of this section is to briefly recapitulate the standard methodologic issues, with the expectation that readers who require more detail will turn to the citations listed in the text.

Study Design

Blackmore et al. [22] identified 238 radiology cost and outcome studies conducted over a 40-year period. Most studies presented primary data from observational cohort or case-control studies (59%) or RCTs (18%), and the remaining studies used secondary data available in the medical literature to build decision analysis models. RCTs are thought to be the best method of providing unbiased evidence on the costs and effectiveness of alternative imaging technologies [23]. The process of randomly allocating patients to receive one of the two or more putative technologies makes it probable that any differences observed in subsequent outcomes will be truly due to the imaging strategy and not caused by the myriad of individual patient characteristics that confound the interpretation of nonrandomized studies. However, RCTs do have drawbacks and are not necessary to answer all radiology outcomes research questions [24]. Most notably, rigorous RCTs require a substantial commitment of time and money. Moreover, often only a select subset of patients enrolls in trials, making the extrapolation of trial results to real-world clinical practice problematic. Despite these caveats, for the most important questions, RCTs should continue to spearhead the push toward the rational use of diagnostic imaging.

Choosing the Perspective of the Study

Innovations in imaging rarely affect all elements of society, such as physicians and insurers, equally. The value of imaging will depend on the viewpoint, or perspective, of the analyst. By stating the perspective of the study, the researcher predetermines the relevant costs that ought to be included in the analysis. For example, a recent trial compared the cost of abdominal CT with 120 mL of nonionic contrast versus the same technique with 100 mL of the same contrast material pushed with 40 mL of saline [25]. From the perspective of the hospital and society as a whole, the small cost reduction of the saline flush method is relevant because it might generate substantial savings in the long run. However, from the perspective of third-party insurers, who pay a fixed reimbursement rate for contrast-enhanced CT, the cost reduction is of no immediate relevance or value. Therefore, an explicit statement of the perspective of the study is a vital, although often overlooked [26], part of a cost and outcomes study.

Current guidelines recommend that the default study perspective should be societal [18]. This is the broadest perspective and includes the costs borne by individuals and public and private organizations within society.

Measuring Costs

Table 1 provides examples of the costs and costing methods that might be used for diagnostic technology assessment from the point of view of four commonly encountered perspectives. Importantly, the cost of medical care to society is not equivalent to the charge billed by the provider. Charges incorporate both costs and a profit margin. From the perspective of society, profit merely represents the transfer of money from one member of society (the payer) to another (the provider), no resources are depleted, and society as a whole is neither richer nor poorer. Therefore, charges tend to overestimate cost.

Costs can either be calculated directly using activity-based costing (ABC) methods or indirectly using proxies for cost based on third-party insurer reimbursement rates or cost-to-charge ratios. The ABC method, also referred to as microcosting, is the more accurate and laborious. It is usually reserved for elements of cost likely to be most influential for the study results. Nisenbaum et al. [27] used ABC methods to calculate the costs of 17 CT procedures performed at a university hospital. Each element of resource use is identified, measured, and valued. For example, the CT machine cost per examination is a function of the purchase cost, maintenance and upgrade costs, machine life expectancy, yearly hours of machine operation, and the number of minutes spent imaging each patient. Using this detailed approach, a cost for all elements of the CT procedure, including

TABLE 1: Costs Under Alternative Perspectives

Resource Item	Perspective			
	Societal	Hospital & Care Provider	Third-Party Insurer	Patient & Family
Diagnostic imaging	Cost of equipment, consumables, overhead, and personnel[a]	Cost of equipment, consumables, overhead, and personnel[a]	Reimbursement rate and administrative costs for covered items	Out of pocket expenses (e.g., charge, copayment)
Medication	Cost of developing, manufacturing, and marketing the drug[b]	Negotiated price of medication	Reimbursement rate and administrative costs for covered drugs	Out of pocket expenses (e.g., charge, copayment)
Outpatient and office-based therapy	Cost of equipment, consumables, overhead, and personnel[a]	Cost of equipment, consumables, overhead, and personnel[a]	Reimbursement rate and administrative costs for covered items	Out of pocket expenses (e.g., charge, copayment)
Hospital admission	Cost of equipment, consumables, overhead, and personnel[c]	Cost of equipment, consumables, overhead, and personnel[c]	Reimbursement rate and administrative costs for covered items	Out of pocket expenses (e.g., charge, copayment)
Patient time and money spent receiving care	Cost of transportation, parking, etc.; opportunity cost of time[d]	Not included	Not included	Cost of transportation, parking, etc.; opportunity cost of time[d]
Patient time off work due to illness	Opportunity cost of time[d]	Not included	Not included	Opportunity cost of time[d]
Informal care giving	Opportunity cost of care givers' time[d]	Not included	Not included	Opportunity cost of care givers' time[d]

[a] In situations in which cost cannot be directly calculated, Medicare reimbursement rates (including both professional and technical components for diagnostic tests) are often used as a proxy for cost [46].
[b] In situations in which cost cannot be directly calculated, average wholesale price, which approximates prices in discount pharmacies, is often used as a proxy for cost [46].
[c] In situations in which cost cannot be directly calculated, Medicare Prospective Payment System or cost-to-charge ratios are often used as a proxy for cost [46].
[d] Hourly wage that patient or care giver would have been earning is often used to estimate cost of time lost due to illness [46].

consumables (e.g., contrast material and film) and radiologist, technologist, administrative, and overhead (e.g., rent) costs, is developed.

The intricate ABC approach is not always feasible, and simpler methods are often sufficient. For example, the Centers for Medicare and Medicaid Services has made extensive efforts to implement a resource-based relative value scale (RBRVS) of reimbursement. This system provides reimbursement for each radiology procedure based on the perceived complexity and resource utilization required to perform that procedure. One advantage of this system is that it is standardized at a national level. Nevertheless, recent work has indicated that substantial inaccuracies may still exist in reimbursement rates, resulting in poorly (e.g., radiography and interventional) and favorably (e.g., sonography, MR, and CT) reimbursed techniques [28]. Other authors have used cost-to-charge ratios to estimate cost by removing the element of profit in the charges billed for medical procedures [29]. The cost-to-charge ratio is the ratio of annual departmental expenditure to revenue. However, because the profit margin may vary widely among imaging examinations, the devaluation of charges based on uniform departmental-level cost-to-charge ratios provides only a crude estimate of the cost of individual imaging examinations. Therefore, overreliance on reimbursement rates or cost-to-charge ratios may distort the cost analysis. In practice, there is a trade-off between the accuracy and the feasibility of costing methods. Many studies use a combination of ABC methods for key cost elements, such as the initial imaging, and cost proxies for other costs, such as subsequent medications and inpatient and outpatient care.

All cost data should be standardized and updated to reflect current costs. Often, because of the scarcity of cost information, analysts draw on cost data from several years. In these circumstances, historical cost data are inflated to current values using the medical care component of the consumer price index. On a similar theme, current U.S. guidelines recommend that future costs, savings, and health outcomes be discounted at a rate of 3% per year [18]. Therefore, a screening test in 2004 that prevented \$1,000 of treatment costs in 2006 would receive credit for saving only \$943 (i.e., $\$1,000 / [1 + 0.03]^2$). The rationale for discounting is based on evidence that people prefer to have resources now rather than in the future for several reasons, including the opportunity to profitably invest current funds. Controversially, discounting lowers the estimated efficiency of screening interventions, in which costs occur immediately but benefits are delayed.

Choosing the Type of Economic Evaluation and Measuring Outcomes

Although there are four types of economic evaluation commonly defined in the literature (Table 2), most health care studies can be classified as one of two types. Currently, the most prevalent method is cost-effectiveness analysis, accounting for more than 80% of published analyses [30]. The distinguishing feature of cost-effectiveness analysis is that the outcome measure used reflects only a limited aspect of health. This primary outcome can be a clinical measure such as mortality, bone density, or exercise tolerance, or a patient-reported measure such as pain or quality of life. For example, in an RCT comparing coronary interventions guided by intravascular sonography or angiography, Mueller et al. [31] used 2-year major cardiac event-free survival to determine whether either imaging method improved patient outcomes. Cost-effectiveness analysis works well in situations in which imaging is expected to improve one predominant aspect of health. However, if imaging is likely to affect more than one element of health or

TABLE 2: Types of Cost and Outcomes Studies

Type of Evaluation	Cost Measure	Outcome Measure
Cost-minimization analysis (CMA)	Dollars	Assumed or known to be the same for both imaging strategies
Cost-effectiveness analysis (CEA)	Dollars	Any of various intermediate (e.g., number of cases detected), clinical (e.g., mortality), or patient-reported (e.g., pain) outcomes
Cost–utility analysis (CUA)	Dollars	Quality-adjusted life years (QALYs)
Cost–benefit analysis (CBA)	Dollars	Dollars

longevity, then the more inclusive quality-adjusted life year (QALY) outcome measure used in cost-utility analysis is recommended.

Cost-utility analysis measures outcomes by weighting years of life by a factor (Q) that represents the patient's health-related quality of life. Q is anchored at 1 (perfect health) and 0 (a health state considered to be as bad as death) and is estimated for all health states between these extremes. A QALY is simply the number of years that a patient spends in each health state multiplied by the quality of life weight, Q, of that state. For example, a patient who spends 2 years in an imperfect health state, where $Q = 0.75$, would achieve 1.5 QALYs (0.75 × 2). The quality weight, Q, can be elicited directly from patients using methods such as the visual analogue scale, time trade-off, and standard gamble; these methods have been described in detail elsewhere [21, 32]. Alternatively, in an increasing number of studies, Q is estimated indirectly via a quality of life questionnaire such as the EQ-5D [33] or the Health Utilities Index [34]. The questionnaire asks the patient to categorize current health in various dimensions—for example, physical functioning, pain, and mental health. Every possible combination of questionnaire responses is associated with a quality weight, Q, from a catalog or algorithm provided by the questionnaire creators. The weights in this catalog are based on prior surveys of the general public's preferences for the health states described by the questionnaire. This indirect approach to estimating Q is currently being used in a trial comparing duplex sonography with clinical surveillance after femoral vein bypass [35]. In this trial, imaging influences medical therapy for ischemia or surgical decisions to amputate and therefore affects several aspects of health, including mobility, self-care, and pain. These researchers chose the EQ-5D questionnaire, which incorporates all of these dimensions of health.

The QALY provides a universal outcome measure that could be used in all clinical trials. Therefore, the efficiency of femoral vein sonography from the trial just described could, in theory, be compared with any other medical intervention in which cost-utility analysis data are available. For this reason, current guidelines favor cost-utility analysis as the most useful method for policy makers [18]. However, some authors are skeptical of the QALY method [36], and it is likely that cost-effectiveness analysis will remain a popular method of economic evaluation in the near future.

The benefits of screening, diagnosis, and preventive treatment may influence the entire course of patients' lives. Therefore, cost and outcomes studies should strive to measure the lifetime impact of imaging. However, in prospective studies it is not practical to follow up patients indefinitely. Therefore, analysts often report the primary results after the first few years of follow-up and extrapolate any differences in cost and outcomes data over the remaining life expectancy of patients [37].

Analysis Methods

The incremental cost-effectiveness ratio (ICER) is conventionally used to summarize the relative efficiency of medical procedures. The ICER is calculated as follows:

$$ICER = (\bar{C}_1 - \bar{C}_0)/(\bar{E}_1 - \bar{E}_0) = (\Delta\bar{C}/\Delta\bar{E})$$

where $\bar{C}_1$, $\bar{C}_0$, $\bar{E}_1$, and $\bar{E}_0$ are the mean cost and effectiveness of the two imaging strategies being compared, and $\Delta\bar{C}$ and $\Delta\bar{E}$ are the difference between the mean costs and mean effectiveness of the two strategies, respectively.

Therefore, a screening strategy that increases costs by an average of $500 per patient and improves life expectancy by an average of 0.04 QALYs per patient, has an ICER of $12,500 per QALY saved. Typically, less cost-effective imaging strategies will have higher, positive, ICER values. However, no consensus exists on an exact threshold that would distinguish efficient from inefficient health care interventions. In reality, this threshold will vary over time and according to many other factors, including the amount of money available to fund health care.

The ICER statistic has several weaknesses. Most important, the meaning of a negative ICER statistic is ambiguous and open to misinterpretation. For example, an efficient imaging strategy that is both cheaper (–$1,000) and more effective (0.1 QALYs) than the strategy with which it is being compared has an ICER of –$10,000. Likewise, an inefficient imaging strategy that is both more expensive ($500) and less effective (–0.05 QALYs) than the strategy with which it is being compared also has the same ICER value, –$10,000. The policy implications of these two scenarios are diametrically opposed, yet the ICER is identical. Furthermore, merely presenting the ICER estimate without quantifying the surrounding confidence interval is of limited value. Unfortunately, however, the ICER has an undefined variance; this complicates even simple statistical tasks such as hypothesis testing and confidence interval calculation [38].

In recognition of these weaknesses, newer methods are emerging, such as the net benefit statistic and cost-effectiveness acceptability curves, and are being used to complement or supplant the ICER statistic in economic analyses. These emerging methods will be discussed in the final section of this article.

It is often difficult to generalize cost-effectiveness results observed in one imaging center to other settings. For example, a survey of 26 Canadian MRI centers concluded that the average operating time per week was 64 hr (range, 25–113 hr) [39]. It would be unreasonable to assume that the cost of MRI equipment per examination is identical for centers at opposite ends of this spectrum. Therefore, sensitivity analysis is frequently used to judge whether study conclusions might be reversed by plausible deviations in parameters, such as the intensity of MRI machine utilization, that underpin cost and efficacy estimates. In the example given, the sensitivity analyst might vary the mean capital cost of MRI by ± 60% to simulate the plausible variation in operating hours and to judge whether a particular application of MRI is likely to be efficient even in centers with low patient throughput. Sensitivity analysis takes many forms, including one-way, multiway, and threshold analyses. These methods have been described in detail in a previous article in this series [40].

Emerging Analytic Methods

Evaluating the Imaging Process from the Patient's Perspective

In many clinical applications there are now a multitude of highly accurate imaging alternatives available. It is frequently impossible to differentiate between two imaging techniques purely on the basis of their impact on patient health or medical care costs. In these circumstances, researchers have begun to formally assess patients' views on the desirability of competing imaging procedures. For example, Blanchard et al. [41] found that 26% of patients undergoing shoulder MRI reported it to be unpleasant or extremely unpleasant compared with 7% undergoing arthrography, although most patients would allow either test to be repeated [41]. Swan et al. [42] developed a method for further quantifying the strength of patient preferences. They report that, on average, patients with peripheral vascular disease would be willing to wait an extra 6 weeks for imaging results and treatment if they could avoid the discomfort and risk of X-ray angiography. By comparison, patients would wait just more than 2 weeks to avoid the MR angiography procedure [42].

Net Benefits

Presenting cost-effectiveness results using the net benefit statistic resolves many of the problems associated with incremental cost-effectiveness ratios [43]. The net benefit statistic is calculated as follows:

$$\overline{NB} = \lambda(\Delta\bar{E}) - (\Delta\bar{C})$$

where λ is the amount that society is willing to pay for an improvement in health.

Therefore, continuing the previous example, if society is willing to pay $100,000 per QALY gained, then our hypothetical screening strategy that increased mean QALYs by 0.04 and increased mean costs by $500 would have a net benefit of $3,500 ([$100,000 × 0.04] – $500). Unlike the ICER, the interpretation of the net benefit statistic is clear-cut; a positive value indicates a cost-effective imaging strategy in which the net costs are more than justified by the net benefits, whereas a negative value indicates the opposite. The larger the net benefit statistic, the more cost-effective the imaging strategy and the more highly it should be prioritized. Furthermore, in large samples the mean net benefit statistic is normally distributed; therefore, hypothesis testing and confidence interval calculation are straightforward [43].

One potential limitation of the net benefit approach is that λ, the value that society is willing to pay for improved health, must be explicitly quantified and embedded in the net benefit calculation. In general, λ is not accurately known and will vary from setting to setting. To address this limitation, many authors now present their results across the spectrum of λ values. These values range from $0, implying that society cannot afford or is not willing to pay anything for improved health and will simply choose the cheapest option, through to millions of dollars, implying that society wishes and is able to pay handsomely for even the most meager health improvements. Using resampling or simulation methods [44], the probability that the net benefit statistic is positive (i.e., the intervention is cost-effective) can be calculated for each value of λ and presented as a cost-effectiveness acceptability curve (CEAC).

Cost-Effectiveness Acceptability Curves

The CEAC describes the probability that an imaging intervention is cost-effective at different willingness-to-pay thresholds. Figure 1 shows the information provided by the CEAC from a randomized trial comparing rapid MRI with radiography as the initial imaging test in patients with lower back pain [45]. The primary finding of this trial was that costs were slightly (≈ $300), but not statistically significantly, higher in patients initially imaged with rapid MRI and that there was no clinically or statistically important difference in physical function outcomes. In this trial, the ICER alone is difficult to interpret because it is negative and has an undefined confidence interval. The CEAC provides more useful information. In this case, the curve crosses the *y*-axis, where society places no value on improvements in back-related function, at 0.16 (Fig. 1). This confirms that, on the basis of the trial data, a 16% probability still exists that rapid MRI is the cheapest strategy. Therefore, more data are required to state with certainty that the rapid MRI strategy is more expensive than radiography. As we move right along the *x*-axis, the probability that rapid MRI is cost-effective increases. This reflects the fact that the more society is willing to pay for improvements in physical function, the more likely it is that the extra cost of rapid MRI will be justified by

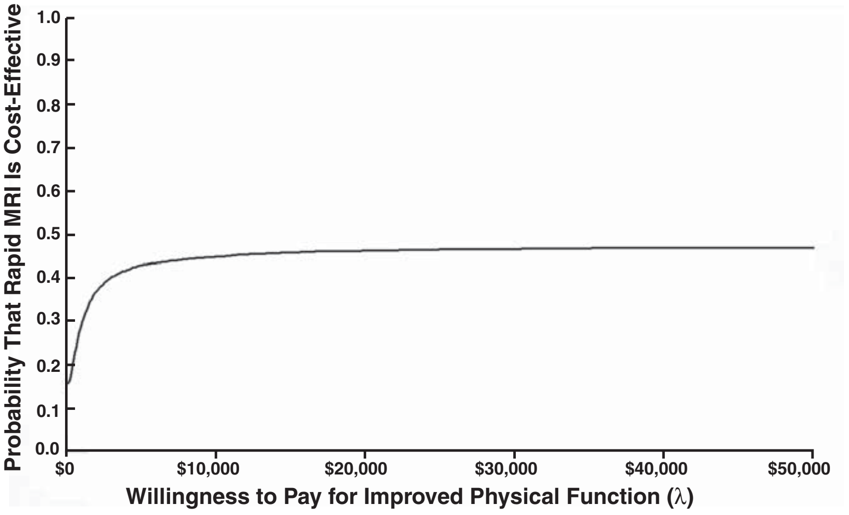


Fig. 1—Graph of cost-effectiveness acceptability curve shows probability that rapid MRI cost-effectiveness increases as society is willing to pay more for improvements in physical functioning.

small improvements in function. However, in this example, the probability curve flattens quickly and never rises above 0.50. This happens because the trial data provide no substantive evidence that the rapid MRI strategy is either more or less effective than radiography. Therefore, even if society is willing to pay excessively for improved health, a 50% probability still exists that rapid MRI is not the most effective strategy. This graph informs the decision maker that it is probable, but not certain, that rapid MRI is currently not a cost-effective initial imaging tool for improving the function of patients with lower back pain.

Conclusions

This article provides a starting point for radiologists and allied health professionals who have an interest in conducting or applying the results of health services research. By its very nature, health services research is multispecialty research because the diagnostic information provided by radiology must be combined with the therapeutic expertise of other clinical specialties to improve the health of patients. This fact, coupled with the large sample sizes needed to provide a definitive answer to some screening questions, can make this type of research seem daunting. However, there are now numerous examples where simple observational studies [12, 13] and compact randomized trials [25, 45] have been used to elucidate the links between diagnostic imaging and the ultimate goal of better health for patients. It seems inevitable that the frequency and importance of these cost and outcomes studies will continue to increase in the future.

Acknowledgement

The author thanks Jeffrey G. Jarvik, MD, MPH, for his useful comments on an earlier version of this manuscript.

References

1. Shapiro S, Strax P, Venet L. Evaluation of periodic breast cancer screening with mammography: methodology and early observations. *JAMA* 1966; 195:731–738
2. Taylor WF, Fontana RS, Uhlenhopp MA, Davis CS. Some results of screening for early lung cancer. *Cancer* 1981; 47:1114–1120
3. Organisation for Economic Co-Operation and Development. *Health at a glance*: *OECD indicators 2003*. Paris, France: OECD, 2003
4. Fuchs VR. Economics, values, and health care reform. *The American Economic Review* 1996; 86:1–24
5. Lindenthal JJ, Lako CJ, van der Waal MA, Tymstra T, Andela M, Schneider M. Quality and cost of healthcare: a cross-national comparison of American and Dutch attitudes. *Am J Manag Care* 1999; 5:173–181
6. Budoff MJ, Achenbach S, Duerinckx A. Clinical utility of computed tomography and magnetic resonance techniques for noninvasive coronary angiography. *J Am Coll Cardiol* 2003; 42:1867–1878
7. U-King-Im JM, Hollingworth W, Trivedi RA, et al. Contrast-enhanced MR angiography vs intra-arterial digital subtraction angiography for carotid imaging: activity-based cost analysis. *Eur Radiol* 2004; 14:730–735
8. Stevens A, Milne R, Burls A. Health technology assessment: history and demand. *J Public Health Med* 2003; 25:98–101
9. Fineberg HV, Bauman R, Sosman M. Computerized cranial tomography: effect on diagnostic and therapeutic plans. *JAMA* 1977; 238:224–227
10. Jarvik JG. The research framework. *AJR* 2001; 176:873–878
11. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991; 11:88–94
12. Brown J, Bryan S, Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ* 1996; 312:809–812
13. Blanchard TK, Bearcroft PW, Constant CR, Griffin DR, Dixon AK. Diagnostic and therapeutic impact of MRI and arthrography in the investigation of full-thickness rotator cuff tears. *Eur Radiol* 1999; 9:638–642
14. Mushlin AI, Mooney C, Grow V, Phelps CE. The value of diagnostic information to patients with suspected multiple sclerosis. Rochester-Toronto MRI Study Group. *Arch Neurol* 1994; 51:67–72
15. Lucock MP, Morley S, White C, Peake MD. Responses of consecutive patients to reassurance after gastroscopy: results of self administered questionnaire survey. *BMJ* 1997; 315:572–575
16. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001; 285:1987–1991
17. Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health* 2004; 94:361–366
18. Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. Panel on Cost-Effectiveness in Health and Medicine. *JAMA* 1996; 276:1339–1341
19. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices–Modeling Studies. *Value Health* 2003; 6:9–17
20. Sunshine JH, Applegate KE. Technology assessment for radiologists. *Radiology* 2004; 230:309–314
21. Singer ME, Applegate KE. Cost-effectiveness analysis in radiology. *Radiology* 2001; 219:611–620
22. Blackmore CC, Black WC, Jarvik JG, Langlotz CP. A critical synopsis of the diagnostic and screening radiology outcomes literature. *Acad Radiol* 1999; 6[suppl 1]:S8–S18
23. Hunink MG, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. *Radiology* 2002; 222:604–614
24. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996; 312:1215–1218
25. Schoellnast H, Tillich M, Deutschmann HA, et al. Abdominal multidetector row computed tomography: reduction of cost and contrast material dose using saline flush. *J Comput Assist Tomogr* 2003; 27:847–853
26. Severens JL, van der Wilt GJ. Economic evaluation of diagnostic tests: a review of published studies. *Int J Technol Assess Health Care* 1999; 15:480–496
27. Nisenbaum HL, Birnbaum BA, Myers MM, Grossman RI, Gefter WB, Langlotz CP. The costs of CT procedures in an academic radiology department determined by an activity-based costing (ABC) method. *J Comput Assist Tomogr* 2000; 24:813–823
28. Saini S, Seltzer SE, Bramson RT, et al. Technical cost of radiologic examinations: analysis across imaging modalities. *Radiology* 2000; 216:269–272
29. Subramanian S, Spies JB. Uterine artery embolization for leiomyomata: resource use and cost estimation. *J Vasc Interv Radiol* 2001; 12:571–574
30. Nixon J, Stoykova B, Glanville J, Christie J, Drummond M, Kleijnen J. The U.K. NHS economic evaluation database: economic issues in evaluations of health technology. *Int J Technol Assess Health Care* 2000; 16:731–742
31. Mueller C, Hodgson JM, Schindler C, Perruchoud AP, Roskamm H, Buettner HJ. Cost-effectiveness of intracoronary ultrasound for percutaneous coronary interventions. *Am J Cardiol* 2003; 91:143–147
32. Morimoto T, Fukui T. Utilities measured by rating scale, time trade-off, and standard gamble: review and reference for health care professionals. *J Epidemiol* 2002; 12:160–178
33. Brooks R. EuroQol: the current state of play. *Health Policy* 1996; 37:53–72
34. Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Ann Med* 2001; 33:375–384
35. Kirby PL, Brady AR, Thompson SG, Torgerson D,

Davies AH. The Vein Graft Surveillance Trial: rationale, design and methods. VGST participants. *Eur J Vasc Endovasc Surg* 1999; 18:469–474

36. Schwappach DL. Resource allocation, social values and the QALY: a review of the debate and empirical evidence. *Health Expect* 2002; 5:210–222
37. Ramsey SD, Berry K, Etzioni R, Kaplan RM, Sullivan SD, Wood DE. Cost effectiveness of lung-volume-reduction surgery for patients with severe emphysema. *N Engl J Med* 2003; 348:2092–2102
38. O'Brien BJ, Briggs AH. Analysis of uncertainty in health care cost-effectiveness studies: an introduction to statistical issues and methods. *Stat Methods Med Res* 2002; 11:455–468
39. Rankin RN. Magnetic resonance imaging in Canada: dissemination and funding. *Can Assoc Radiol J* 1999; 50:89–92
40. Plevritis SK. Decision analysis and simulation modeling for evaluating diagnostic tests on the basis of patient outcomes. *AJR* 2005; 185:581–590
41. Blanchard TK, Bearcroft PW, Dixon AK, et al. Magnetic resonance imaging or arthrography of the shoulder: which do patients prefer? *Br J Radiol* 1997; 70:786–790
42. Swan JS, Fryback DG, Lawrence WF, Sainfort F, Hagenauer ME, Heisey DM. A time-tradeoff method for cost-effectiveness models applied to radiology. *Med Decis Making* 2000; 20:79–88
43. Zethraeus N, Johannesson M, Jonsson B, Lothgren M, Tambour M. Advantages of using the net-benefit approach for analysing uncertainty in economic evaluation studies. *Pharmacoeconomics* 2003; 21:39–48
44. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves: facts, fallacies and frequently asked questions. *Health Econ* 2004; 13:405–415
45. Jarvik JG, Hollingworth W, Martin B, et al. Rapid magnetic resonance imaging vs radiographs for patients with low back pain: a randomized controlled trial. *JAMA* 2003; 289:2810–2818
46. Gold M, Siegel J, Russell L, Weinstein M. *Cost-effectiveness in health and medicine*. New York, NY: Oxford University Press, 1996

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005
16. Statistical Inference for Proportions, April 2005
17. Reader Agreement Studies, May 2005
18. Correlation and Regression, July 2005
19. Survival Analysis, July 2005
20. Multivariate Statistical Methods, August 2005
21. Decision Analysis and Simulation Modeling for Evaluating Diagnostic Tests on the Basis of Patient Outcomes, September 2005

Meta-Analysis of Diagnostic and Screening Test Accuracy Evaluations: Methodologic Primer

Constantine Gatsonis[1]
Prashni Paliwal[1,2]

OBJECTIVE. Interest in evidence-based diagnosis is growing rapidly as diagnostic and screening techniques proliferate. In this article we provide an overview of systematic reviews of diagnostic performance and discuss in detail statistical methods for the most common variant of the problem: meta-analysis of studies in which a pair of estimates of sensitivity and specificity is reported. The need to account for possible variations in threshold for test positivity across studies led to the formulation of the Summary ROC (SROC) curve method. We discuss graphical and model-based ways to estimate, summarize, and compare SROC curves, and we present an example from a meta-analysis of data on techniques for staging cervical cancer. We also present a brief survey of the methodologic literature for addressing heterogeneity, correlated data, multiple thresholds per study, and systematic reviews of ROC studies. We conclude with a discussion of the significant methodologic challenges that continue to face investigators in this area of diagnostic medicine research.

CONCLUSION. Systematic reviews of diagnostic performance are a rigorous approach to examining and synthesizing evidence in the evaluation of diagnostic and screening tests. The information from such reviews is needed by clinicians, health policy makers, researchers in diagnostic medicine, developers of diagnostic techniques, and the general public. However, despite progress in study quality and reporting and in methodologic development, major challenges confront investigators undertaking these reviews.

The need for systematic reviews of diagnostic and screening tests has grown markedly in recent years as technologic advances have brought forth a vast array of such techniques. Patients, physicians, and policy makers all need information on the reliability and performance of tests and the interpretation of results. In addition, the increased availability of a plethora of diagnostic and screening techniques has meant increased use of tests and a dramatic increase in health care costs.

As evidence-based medicine expands from therapy to diagnosis, the role of systematic reviews acquires added importance [1]. The information from systematic reviews of diagnostic and screening tests is necessary for the following purposes: determination of the proper and efficacious use of diagnostic and screening tests in the clinical setting; decision making about health care policy and financing; evaluation of the performance and status of a diagnostic technique to determine areas for further research, development, and evaluation; and evaluation of the quality and scope of available primary studies of diagnostic and screening techniques and thus development of information necessary for determining directions of future research in diagnostic medicine.

A taxonomy of the important aspects of evaluation of diagnostic and screening tests would distinguish three broad areas of end points: the diagnostic performance of the test, assessed with measures of test accuracy and predictive value; the impact of the test on the process of care, assessed by metrics of the effect of the test on subsequent diagnostic and therapeutic decision making; and the impact of the test on patient-level outcomes, including mortality, morbidity, satisfaction and health-related quality of life, health care utilization, and cost [2–4].

It is also possible, although not formally practiced, to distinguish developmental levels for a technique, following the trajectory from early development to broad dissemination. For example, a four-stage categorization would include stage 1 (discovery), in which the technical parameters and diagnostic crite-

Series editors: Craig A. Beam, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 23rd and final artricle in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology.* The series has been designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

Keywords: diagnostic accuracy, evidence, meta-analysis, statistical methods, Summary ROC curve, systematic reviews

DOI:10.2214/AJR.06.0226

Received February 13, 2006; accepted after revision February 20, 2006.

[1]Center for Statistical Sciences, Brown University, Box G-H, Providence RI, 02912. Address correspondence to C. Gatsonis (gatsonis@stat.brown.edu).

[2]Present address: Department of Psychiatry, Yale University, New Haven, CT 06511.

AJR 2006; 187:271–281

0361–803X/06/1872–271

ria of a technique are established; stage 2 (introduction), in which diagnostic performance is assessed and fine tuning of the technology is performed in single-institution studies; stage 3 (maturity), in which the technique is evaluated in comparative, multicenter, prospective clinical studies (efficacy); and stage 4 (dissemination), in which the technique is evaluated as used by the community at large (effectiveness) [3].

Appropriate end points can be selected for each developmental level of a technique. In general, however, evaluation of diagnostic performance is a relevant end point for studies at any stage. The most commonly used metric of diagnostic performance, and the one discussed in detail in this primer, is the pair of estimated sensitivity and specificity values for a test. Others include receiver operating characteristic (ROC)–based measures and measures of the predictive value of a test.

This primer focuses exclusively on systematic reviews of the diagnostic performance of tests. We provide a brief description of the main steps in conducting systematic reviews, from formulating the research question through primary study retrieval and data collection to data analysis and interpretation of results. We also discuss statistical methods for deriving summaries of diagnostic performance data and give an example of an application to meta-analysis of the diagnostic accuracy of tests in the detection of lymph node involvement in women with cervical cancer. The article considers methods for meta-analysis of studies in which a single pair of sensitivity and specificity estimates is reported. Extensions of the basic method are described, and a brief guide to the methodologic literature is provided. We summarize our recommendations and discuss methodologic and subject-matter challenges in the last section.

Overview of Systematic Reviews of Diagnostic Accuracy

The conduct of a systematic review of diagnostic test accuracy proceeds through the following major steps [5, 6]:

1. Definition of the objectives of the review.
2. Literature search and retrieval of studies.
3. Assessment of study quality and applicability to the clinical problem at hand.
4. Extraction of data.
5. Statistical analysis.
6. Interpretation of results and development of recommendations.

Each of the six steps in the process involves its own challenges and can be further refined with more detailed flowcharts [7]. We provide a brief description of the tasks involved in each step.

Definition of the Objectives of the Review

A systematic review of diagnostic accuracy begins with defining the clinical context and developing a precise description of the diagnostic question for which test accuracy is to be assessed. This part of the process is similar to the development of the protocol for a primary study. It includes specification of the clinical question giving rise to the potential use of the test or tests under investigation, the technical characteristics of the tests, the conditions under which the tests are interpreted, and the reference information used in the assessment of test accuracy [8]. Because systematic reviews of diagnostic accuracy are called on to inform the use of diagnostic tests in clinical care, comparisons of alternative tests are most valuable.

Literature Search and Retrieval of Studies

Although on search strategies extensive literature for studies of therapy is available, the corresponding body of literature on diagnostic test evaluation is relatively small. Deville et al. [9] and Bachmann et al. [10] discuss strategies relating to diagnostic and screening tests.

The search for appropriate studies must be comprehensive, objective, and reproducible, and the searcher must consider all available evidence. The search should not simply be for documents in English and should cover publications beyond journals, such as conference proceedings and other reports. Hand searching through publications, reference checking, and searching for unpublished reports often is necessary, especially to assess the extent of publication bias. Finally, it is important to document the process and the outcome of each search.

Assessment of Study Quality and Applicability

The scope of assessment of study quality is broad and not generally well defined. In the context of studies of diagnostic performance, assessment of quality has to consider the important features of the design and execution of the study, including factors such as definition of the research question and clinical context, specification of appropriate patient population, description of the diagnostic techniques under study and their interpretation, detailed accounting of how the reference standard information was defined and obtained, and any other factors that can affect the integrity of the study and the generalizability of the results.

Methods of quality assessment may focus on the absence or presence of key qualities in the study report (checklist approach), use scores developed for this purpose (scale approach), or use the levels-of-evidence methods by which a level or grade is assigned to studies fulfilling a predefined set of criteria. The literature on assessment of the quality of therapy studies is extensive, at least in comparison with the literature on diagnostic test evaluations [11, 12]. Two developments in the diagnostic area are the Standards for Reporting of Diagnostic Accuracy (STARD) checklist for reporting of studies of diagnostic accuracy [4, 13, 14] and the quality assessment tool for diagnostic accuracy (QUADAS) for assessing the quality of studies of diagnostic accuracy [12, 15]. The former may be beneficial in improving the quality of published reports and, indirectly, in improving the quality of primary studies. The latter is a rigorously constructed tool that can be used by investigators undertaking new systematic reviews.

Incorporation of quality assessment results into meta-analysis is a matter of debate. A simple and perhaps draconian approach is to exclude studies of poor quality. A less drastic alternative is to use quality scores as weights in the statistical analysis. However, the exact definition of the weights is often a matter of disagreement, and the statistical rationale for their use is shaky. Another alternative, which we recommend to investigators, is to conduct sensitivity analysis. The goal of sensitivity analysis is to assess the contribution of poor-quality studies to the results of the full meta-analysis. The assessment is made by comparing the results from the statistical analysis with the results of the specific studies included and excluded. Sensitivity analysis also can be used to assess the effect on diagnostic accuracy of a study characteristic or a combination of study characteristics.

Extraction of Data

In studies of imaging techniques, test results are most commonly reported as binary (yes or no) or ordinal categoric. An example of the latter often used in ROC studies is a five-category scale for degree of suspicion about the presence of a target condition. The categories are commonly described as follows: 1 = definitely normal, 2 = probably normal, 3 = equivocal, 4 = probably abnormal, and 5 = definitely abnormal. In recent years, degree of suspicion assessments also have been made on nearly continuous scales, for example, scales from 1 to 100. Continuous test results are typically reported in the evaluation of laboratory tests, such as the concentration of a substance.

A binary test result is typically obtained by dichotomizing a test outcome measured on a continuous scale. The continuous scale can be observed directly, as is the case with many laboratory tests. As an alternative, the scale can be a latent, unobservable one, as is the case with the observer's degree of suspicion in ROC studies. In either case, the binary test result is obtained by application of a threshold for test positivity. The presence of such a threshold is a fundamental theme in the evaluation of diagnostic and screening tests.

In this primer, as in most published work on diagnostic and screening test evaluation, disease status is assumed to be binary. Thus, for a particular threshold of test positivity, the study results can be presented in the familiar two-by-two table showing cross classification of disease status and test outcome (Table 1).

Although it may seem reasonable to expect that obtaining an appropriate two-by-two table from a published study should be rather straightforward, practical experience suggests that this is not always the case. Investigators need to consider carefully the data report and may also need to contact the authors of the report to obtain the necessary information.

Measures of test performance are defined either conditionally on disease status (sensitivity, specificity) or conditionally on test result (predictive value). Commonly used metrics include test sensitivity = $P(T+|D+)$; specificity = $P(T-|D-)$; positive predictive value = $P(D+|T+)$; and negative predictive value = $P(D-/T-)$, where $P(\ldots)$ is the probability of the event in parentheses, T is the test result, and D is the true disease status. In addition, studies may report other metrics, such as diagnostic odds ratio (OR): *sens spec* / (1 – *sens*)(1 – *spec*); positive likelihood ratio: *LR*+ = $P(T+|D+)$ / $P(T+|D-)$ = *sens* / (1 – *spec*); and negative likelihood ratio: LR– = $P(T-|D+)$ / $P(T-|D-)$ = (1 – *sens*) / *spec*. See also the recent article in the *AJR* by Weinstein et al. [16].

TABLE 1: Two-by-Two Table of Binary Test Results Versus Disease Status

Test Result (T)	True Status (D)		
	Nondiseased	Diseased	Total
Negative	*a*	*c*	*a* + *c*
Positive	*b*	*d*	*b* + *d*
Total	*a* + *b*	*c* + *d*	*N*

This primer is concerned mainly with meta-analysis of studies reporting estimates of pairs of sensitivity and specificity. The methods discussed in the next section assumes the availability of a single two-by-two table from each study. However, the results of some studies are reported with more than one threshold of test positivity and even more than one definition of disease status. It is important for investigators to record all the information on alternative thresholds reported in retrieved studies and to determine which of the thresholds of test positivity is the most relevant for the purposes of the systematic review. The methods for combining data when several thresholds are used in each study is beyond the scope of this primer but is discussed briefly later in the Other Methods section.

Statistical Analysis

Because binary test outcomes are defined on the basis of an explicit or implicit threshold for test positivity, it follows that measures of binary test performance depend on the particular threshold used to generate the binary test outcomes. This dependence is a fundamental aspect of diagnostic test evaluation. In the case of test sensitivity and specificity, dependence on the threshold induces a tradeoff between the two quantities as the threshold for positivity is moved across all possible values. The curve of all pairs of sensitivity and specificity values achieved by moving the threshold across its possible range is the ROC curve [17, 18].

Comparison of tests on the basis of ROC curves takes into consideration the actual curves and is aided by summary measures that have been proposed in the literature. The area under the curve (AUC) is the most commonly used summary and can be interpreted as average sensitivity for the test, taken over all specificity values. Strictly speaking, the AUC is equal to the probability that if a pair of diseased and nondiseased subjects is selected at random, the diseased subject will be ranked correctly by the test. Other summaries of the ROC curve include partial areas under the curve, values of sensitivity corresponding to selected values of specificity (and vice versa), and optimal operating points, defined according to specific criteria. ROC analysis and other statistical methods for diagnostic test evaluation are described in textbooks by Zhou et al. [19] and Pepe [20] and in chapters by Toledano et al. [21] and Toledano [22].

Digression to ROC analysis is necessary to highlight the role of the positivity threshold and its consequences. A direct implication of this issue in meta-analysis of sensitivity and specificity estimates is that the method has to account for the possibility of different thresholds across studies. The use of simple or weighted averages of sensitivity and specificity to draw statistical conclusions is not methodologically defensible. A simple example to illustrate this point is a meta-analysis of three studies with the sensitivity and specificity estimates described in Figure 1. The estimated sensitivity and specificity pairs are (0.1, 0.9), (0.8, 0.8), and (0.9, 0.1). The average pair is (0.6, 0.6). Clearly, the (0.6, 0.6) pair does not represent these data in any useful way; thus, a simple averaging of sensitivity and specificity is not an adequate approach.

"Average" values of sensitivity and specificity sometimes are used as descriptive summaries of the observed data. Typically, this approach would be the case when the observed variability in one or both of the two quantities is small.

Interpretation of Results

Interpretation of the findings from a meta-analysis of diagnostic performance must address the relevance of the results to the four general aims stated earlier. That is, this section of the report should highlight the specific ways in which the data provide information about the proper use of the particular test, preferably in comparison with alternative techniques; discuss how the findings can be used to make decisions about health care policy and financing; summarize the quality of the available studies, pointing to areas in which more research is needed; and provide information about possible areas of improvement in the performance of the techniques under review.

Statistical Methods for Meta-Analysis of Sensitivity and Specificity Data

Summary ROC (SROC) Curve for a Single Test

Our focus is on meta-analyses in which each study contributes a two-by-two table of data, on the basis of which a pair of estimates of sensitivity and specificity can be obtained. To introduce statistical notation, the *i*th study ($i = 1,\ldots I$) contributes data in the format shown in Table 2. With the notation of Table 2, the estimates of sensitivity and 1 – specificity from the *i*th study are $\text{TPR}_i = d_i \ / \ n_{i1}$ and $\text{FPR}_i = b_i \ / \ n_{i0}$, where

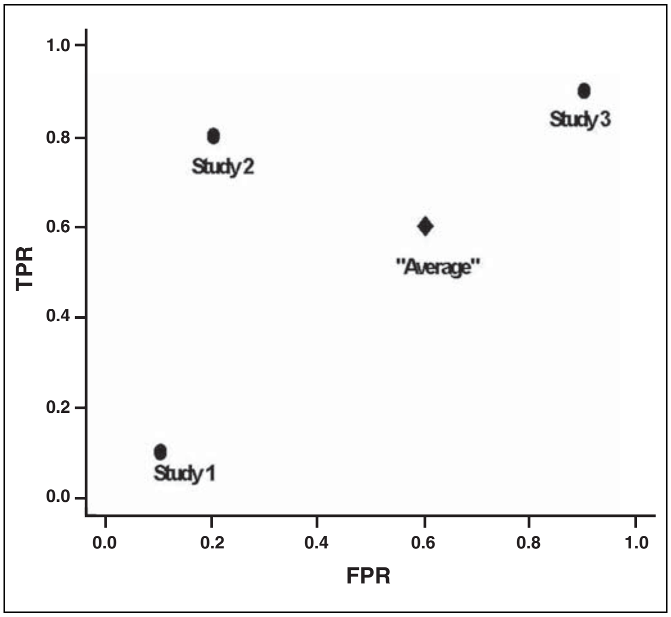


Fig. 1—Graph shows that averaging sensitivities and specificities can be misleading. TPR = true-positive rate, FPR = false-positive rate.

TPR is the true-positive rate and FPR is the false-positive rate.

The display of paired estimates of sensitivity and specificity in ROC coordinates (FPR, TPR) is a key step in the process of statistical analysis. Such plots ideally should include error bars for each of the two estimates. However, the bars often make the plot rather busy. An additional plot to consider is a forest plot, which shows the sensitivity and specificity estimates of each study side by side and may also include the numerators and denominators used to construct the estimates (Figure 2).

Simple derivation of an SROC curve—An easy way to construct a graphical summary of (FPR, TPR) estimates was introduced by Moses and colleagues in 1993 [23]. In this approach, the original data are first transformed into new variables *S* and *D*, defined as follows for the *i*th study:

$$S_i = \text{logit}(\text{TPR}_i) + \text{logit}(\text{FPR}_i)$$
$$D_i = \text{logit}(\text{TPR}_i) - \text{logit}(\text{FPR}_i)$$

TABLE 2: Format for *i*th Study

Test Result (T)	True Status (D)		
	Nondiseased	Diseased	Total
Negative	a_i	c_i	
Positive	b_i	d_i	
Total	n_{i0}	n_{i1}	n_i

where logit(a) = log [a / (1 – a)]. The next step is to fit a linear regression model of the form

$$D_i = a + bS_i + \text{error}.$$

The fitted model provides a value of *D* for each value of *S*. In the final step, the *D* and *S* pairs are transformed back into ROC coordinates to obtain an SROC curve.

The transformed variable *D* is actually the diagnostic odds ratio estimated from each individual primary study in the meta-analysis. The variable *S* has a less straightforward interpretation. A little algebra shows that *S* increases when the probability of a positive test result increases in both the diseased and nondiseased populations. Hence, *S* can be interpreted as a proxy for the test positivity threshold operating in the particular study. This way of constructing an SROC curve is roughly based on an implicit assumption that the variation in diagnostic odds ratio across studies is a function of the threshold for test positivity.

The foregoing model can be easily extended to incorporate covariates measuring study characteristics or group characteristics of the participants in the individual primary studies. The linear model would then have the following form:

$$D_i = a + b_0S_i + b_1X_1 + b_2X_2 + \dots \\ + b_kX_k + \text{error},$$

where the *X* variables can be suitably defined to represent characteristics of the study design, the test technology, and study participant characteristics as used in subgroup analyses. A model with appropriately defined indicator variables *X* can also be used to compare tests.

SROC summaries—In analogy with the usual ROC curve, a natural summary of the SROC is the AUC. However, the choice of the exact limits for defining the area is a matter of some debate. In particular, some authors prefer to compute the area only over the range of the observed FPR values to avoid the inherent uncertainties about extrapolating beyond the range of the observed data. Other authors support the use of a partial area over a range of FPR values of interest in the context of the particular test. In this primer we report the full AUC estimates because of their simplicity, intuitive interpretation, and avoidance of arbitrary choices of limits of FPR values.

Another global summary of the SROC curve is the so-called Q* ("Q-star") statistic, which measures the value of TPR at the point where the curve intersects the $x + y = 1$ diagonal line. This is the point on the curve where sensitivity equals specificity. For a symmetric curve, this value is also the point at which the curve is closest to the ideal point (FPR = 0, TPR = 1).

In addition to the global summary measures, the SROC curve can be used to estimate TPR for each fixed value of FPR and, conversely, standard errors of the estimates can be obtained using the delta method. We include such estimates in the analysis of the cervical cancer data (Fig. 5).

SROC summaries can be used to compare the performance of alternative diagnostic and screening tests for a particular diagnostic question. These comparisons are relatively straightforward when statistical independence can be assumed to hold, as when SROC curves of alternative tests are derived from separate sets of studies or from overlapping sets of studies in which test results were not correlated. However, the situation is technically more complex when test results within a study are correlated, as is the case when a paired design has been used to compare tests. We discuss this issue later, in Other Methods.

SROC properties and limitations—The shape of the SROC curve derived from the foregoing linear regression model depends on the values of the linear model parameters *a* and *b* [24]. The special case of $b = 0$ corresponds to the situation in which the true diag-

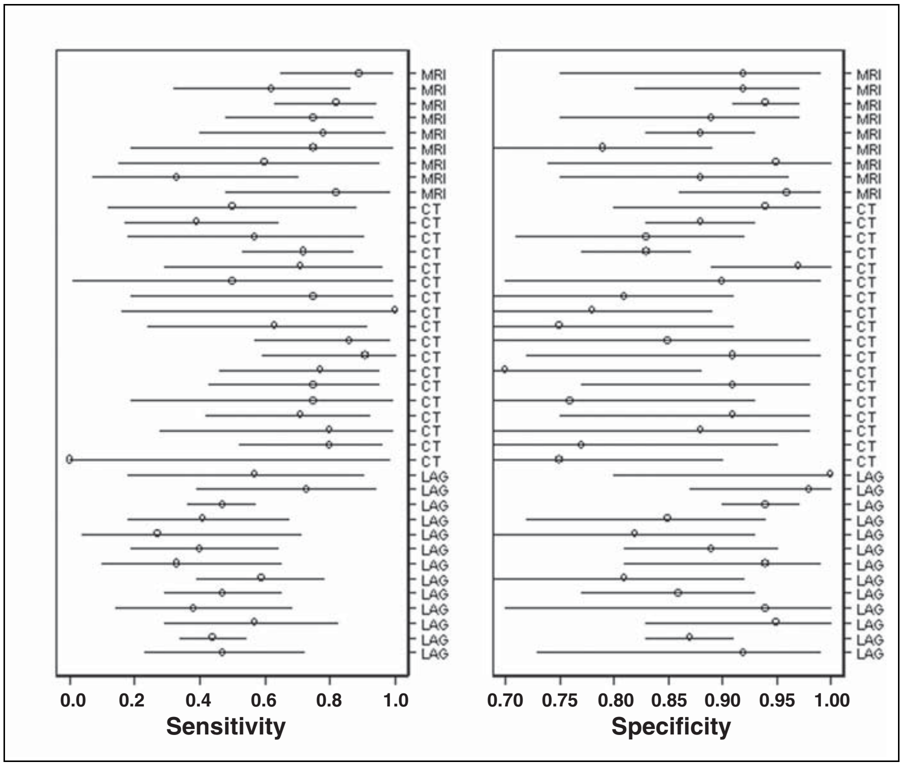


Fig. 2—Forest plot of CT sensitivity and specificity estimates and their confidence intervals.
LAG = lymphangiography.

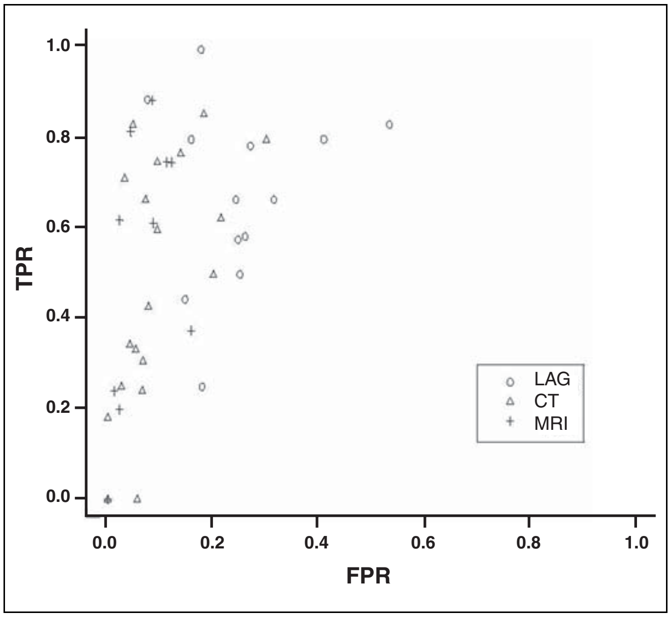


Fig. 3—Observed true-positive rates (TPR) and false-positive rates (FPR) for three imaging techniques.
LAG = lymphangiography.

nostic odds ratio is assumed to be constant across all studies. In this case, the SROC curve is symmetric along the $x + y = 1$ diagonal line. If $b \neq 0$, the curve is not symmetric. Indeed, it turns out that when $| b | > 1$, the SROC curve derived from the linear regression model has a counterintuitive property: According to the curve, the sensitivity of the test decreases as the FPR increases. Estimated values of b greater than 1 or less than –1 indicate that the simple linear regression model is not adequate for constructing an SROC curve.

SROC curve computations based on the linear regression model are a simple and useful method for developing such curves. There are, however, potentially important technical difficulties to overcome if the results of this approach are used to draw formal statistical inferences. First, the presence of sampling error in the variable S on the right hand side of the linear model may affect the magnitude of the estimates of b and its SE. The sampling error may increase the uncertainty in the estimate of b, leading erroneously to the conclusion that the SROC is symmetric. Second, the linear model uses summaries from the two-by-two tables of the individual studies and ignores the statistical precision of these summaries. Unfortunately, the precision of TPR and FPR estimates is somewhat complex because it depends not only on overall sample size but also on the sample sizes for diseased and nondiseased subjects in the study. Hence, simple weighting by sample size is not sufficient. In addition, the left-hand-side and right-hand-side variables in the linear model have their own estimates of statistical precision, making it difficult to decide on a single weight for the particular study. Third, the linear model does not account for the presence of correlations in the data, such as those resulting from the use of paired designs within individual primary studies.

Binary regression for SROC analysis—Because of the methodologic difficulties described, it is prudent for investigators to consider the use of alternative approaches to estimating SROC parameters for purposes of formal statistical inference. An early such approach predated the linear regression method and used the bivariate normal distribution of the estimates of sensitivity and specificity from each study, with a linear relation between the true values of sensitivity and specificity to account for the effect of threshold [25].

A streamlined alternative to the linear regression model is to use a variant of logistic regression, which models directly the data ineach two-by-two table [26–28]. If Y is the binary test result (yes = 1, no = 0) and D the binary disease status for an individual patient in a given study, the form of the model is as follows:

$$\text{logit}P[Y = 1] = (\theta - \alpha D)\exp(-\beta D).$$

The binary regression model is intuitively based on the usual conceptualization of the binary test outcome resulting from a positivity threshold (denoted here by θ). In other words, the binary test outcome is obtained by dichotomizing a continuous variable that has different distributions for diseased and nondiseased subjects. The parameter α measures the distance between the centers of the diseased and nondiseased populations, and the parameter β measures the ratio of the SDs in the two populations. The mathematic details of the model and its relation to the linear model approach are sketched in Appendix 1.

The use of binary regression allows investigators to avoid key difficulties associated with the linear model approach, notably the errors-in-variables problem and the need to account for differences in sample size across studies. As shown in Appendix 1, it is possible to translate the findings of binary regression analysis into linear model parametrization. However, the SROC curves obtained from binary regression analysis always lead to values of the slope between –1 and 1 and hence avoid the counterintuitive properties of curves with $| b | > 1$ obtained from the linear model. Binary regression models can be fitted with standard software, such as Proc NLMixed in SAS [29]. The SAS code for fitting a binary regression model using Proc NLMixed is in Appendix 2.

Example: Meta-Analysis of Cervical Cancer Staging Data

To illustrate the SROC method we use data from a meta-analysis of diagnostic imaging tests in the detection of lymph node metastasis in patients with cervical cancer [30]. This systematic review was conducted to compare the performance of three imaging techniques: lymphangiography (LAG), CT, and MRI. The published report describes how the problem was formulated, how the relevant studies were identified and reviewed, and how the diagnostic performance data were extracted. Briefly, studies were located with a MEDLINE literature search combined with hand searching of bibliographies from retrieved articles. Included studies had histologic confirmation of cervical cancer, uniformly appropriate reference standard information, and evidence of blinding in study design. In addition, included studies had a minimum sample size of 20 patients, reported criteria for test positivity, and presented sufficient data to complete the necessary two-by-two table.

In our example we included data from 42 studies, 13 of which evaluated LAG, 19 evaluated CT, and 10 evaluated MRI. Nine studies evaluated more than one test, but this feature of the data is ignored for the purposes of this analysis. The pairs of observed values of sensitivity and specificity are presented in ROC coordinates in Figure 3. We are not using exactly the same set of studies presented in the published paper, and hence the results of this example may differ from those in the article, particularly in the case of the LAG evaluation.

SROC curves were derived separately for each test by both the binary and the linear regression methods. The results of the binary regression fit are presented in detail and are followed by summary tables from the linear regression fit. The latter are included for comparison purposes. For each test, the binary regression model assumed common location (α) and scale (β) parameters across the studies but a separate threshold value for each study. Table 3 summarizes the results from the binary regression fit.

TABLE 3: Estimates of Summary Receiver Operating Characteristic Curve Parameters, Area Under the Curve (AUC), and Q* Statistic for Each Technique (Binary Regression Model)

Technique	α (Location)	SE (α)	β (Scale)	SE (β)	AUC	SE (AUC)	Q*
LAG	–1.965	0.365	–0.500	0.639	0.719	0.054	0.677
CT	–3.380	0.737	–0.591	0.399	0.839	0.025	0.769
MRI	–3.349	0.501	0.1	0.376	0.933	0.021	0.862

Note—LAG = lymphangiography.

The scale parameter is not statistically different from zero for all three techniques. Instead of assuming it is zero and plotting the SROC curves as symmetric, we used the estimated value of β to derive the plots in Figures 3 and 4. The SROC curves are superimposed on the observed data in Figure 4. The SROC curves with superimposed 95% confidence intervals for TPR and FPR at three points are shown in Figure 5.

The SROC curve for LAG stays consistently below the curves of the other two techniques. The MRI curve dominates the CT curve, and its summary values of AUC and Q* estimates dominate the other two techniques. However, only one of the paired comparisons of the AUC estimates (LAG vs MRI) is statistically significant. A comparison of the confidence intervals for TPR and FPR also shows overlap at each of three points chosen in Figure 5. We conclude that although there is a trend for MRI to have better performance than CT and LAG, only the AUC of MRI is statistically different from that of LAG.

For comparative purposes, we present the numeric results from the linear regression fit of the SROC curve (Table 4). The actual curves and summary estimates of AUC and Q* are close but not identical to those derived from the binary regression analysis. For a more detailed view of the comparison, we converted the SROC equation from the binary regression to the form that would be obtained from a linear regression fit, using the formulas in Appendix 1. Table 5 shows the results for CT.

Other Methods

The SROC method is limited in two important respects. First, the statistical framework does not consider the presence of random variation between studies. This fixed-effects framework implicitly assumes that the universe of all studies to which inferences apply is only the specific studies used in the meta-analysis and that in addition to sampling variation within studies, the only other possible variation can be explained by study-level covariates. As a result of its assumptions, a fixed-effects approach to meta-analysis is generally expected to provide artificially more precise results than an approach that provides a fuller account of variability in the data [31]. The second important limitation of the specific fixed-effects approach is that it ignores correlations in the data within studies. In this section, we briefly discuss statistical methods based on hierarchical models de-

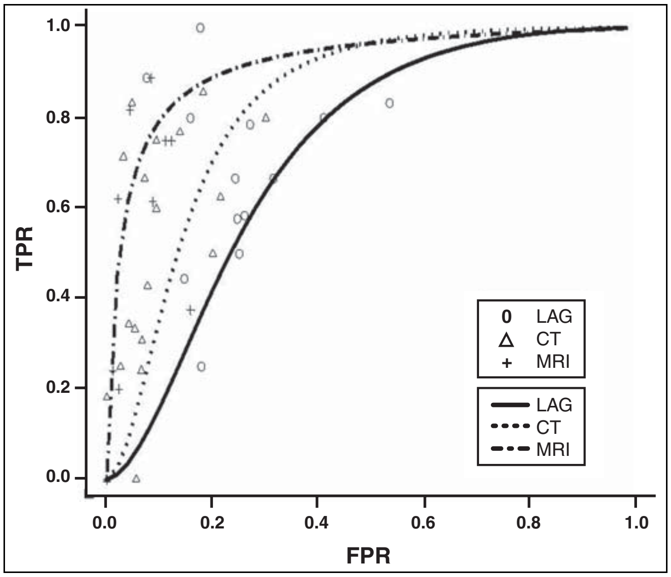


Fig. 4—Estimated SROC curves and original data points for three imaging techniques. TPR = true-positive rate, FPR = false-positive rate, LAG = lymphangiography.

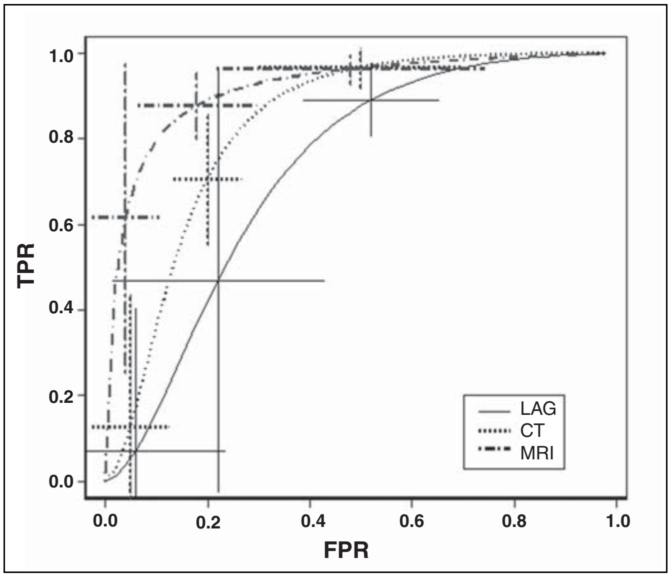


Fig. 5—Summary Receiver Operating Characteristic curves with confidence intervals for selected (TPR, FPR) points. TPR = true-positive rate, FPR = false-positive rate, LAG = lymphangiography.

signed to address these limitations. We also provide references to the literature on meta-analysis of ROC studies.

Hierarchical Summary ROC Analysis

The binary regression model is the building block for a hierarchical model describing the full range of variation in the data. In particular, the hierarchical model differentiates within-study from between-studies variability and systematic from random variability. For example, a model for the cervical cancer data accounts for two levels of variability. In level 1, within-study variation is modeled by binary regression. In level 2, between-studies variation is modeled by distributions of the threshold and location parameters. The mean of the distribution of the parameters may depend on study-level covariates (e.g., test type).

A hierarchical model can be fitted with fully bayesian methods [27] or likelihood-based approximations as implemented in the Proc NLMixed procedure of SAS [28]. A Hierarchical SROC (HSROC) curve can be derived by use of the population means of the parameters. In addition to providing a full account of the variability in the data, the hierarchical model accounts implicitly for correlations within studies. If information exists for such correlations, it can be included explicitly by suitable extensions of the model. In particular, such formulations are useful for modeling data from studies conducted with paired designs.

An alternative way to build hierarchical models for diagnostic accuracy data is to consider a variant of the Kardaun approach and use the bivariate asymptotic normal distribution of the estimates of sensitivity and specificity from each study [32]. Although this approach has been used to derive "average" estimates of sensitivity and specificity, a practice criticized earlier in this article, it is easy to modify the model to derive SROC curves.

Meta-Analysis with Multiple Thresholds from Individual Studies

In the SROC methods discussed earlier, it is assumed that a single two-by-two table is obtained from each study. If multiple thresholds for test positivity are used in the primary studies, ordinal regression methods and their hierarchical formulations can be used to perform the statistical analysis [33].

Meta-Analysis of ROC Data

The choice of suitable statistical methods for combining data from ROC studies depends on the type of data considered. If the full ROC data are available—for example, the complete two-by-five table of disease status by test results when a five-point ordinal categoric scale is used—then ordinal regression methods can be used. It is not necessary for all studies to use the same number of categories in reporting of test results [33, 34].

If the emphasis is on meta-analysis of summaries of the ROC curve, the appropriate methods have to be tailored to the specific summary. For meta-analysis of estimates of the AUC from independent studies, McClish [35] describes weighted average estimators, Zhou [36] describes a generalized estimating equation approach, and Hellmich et al. [37] describe a bayesian method. A hierarchical model for such data can be constructed in a straightforward manner with the asymptotic distribution of the estimate of the AUC for the first level of the model and proceeding as in the HSROC model for the other levels. Because the distributions involved are all normal, the process of fitting and checking such models is fairly routine [31, 38].

Discussion

As interest in evidence-based diagnosis increases, so does the demand for information from systematic reviews of studies of diagnostic accuracy. The information from such reviews is a key ingredient for all subsequent evaluation of diagnostic techniques. Because empiric studies of test outcomes can be prohibitively difficult to conduct in practice, research synthesis and modeling of health outcomes and costs often remain the only viable options. For such undertakings, the informa-

TABLE 4: Estimates of Summary Receiver Operating Characteristic Curve Parameters, Area Under the Curve (AUC), and Q* Statistic for Each of Three Techniques (Linear Regression Model, Unweighted)

Technique	α	SE (α)	β	SE (β)	AUC	SE (AUC)	Q*
LAG	2.006	0.308	0.299	0.276	0.779	0.031	0.731
CT	2.788	0.360	0.219	0.118	0.861	0.024	0.801
MRI	3.508	0.609	0.255	0.187	0.916	0.028	0.852

Note—LAG = lymphangiography.

TABLE 5: Comparison of Binary and Linear Regression Summary Receiver Operating Characteristic Analyses for CT Data

Analysis	α	SE (α)	β	SE (β)	AUC	SE (AUC)	Q*
Linear (unweighted)	2.788	0.360	0.219	0.118	0.861	0.024	0.8012
Linear (weighted)	2.606	0.329	0.141	0.114	0.854	0.028	0.7863
Binary regression	2.409	0.410	0.288	0.183	0.839	0.025	0.7694

Note—AUC = area under the curve.

tion from meta-analysis of diagnostic performance is crucial.

Meta-analysis of accuracy evaluations is not as streamlined or easy to perform and summarize as meta-analysis of therapy evaluations. A key difference is the nature of the summary measure. In therapy studies, the summary can be as simple as an overall success rate with appropriately quantified variability and uncertainty. In diagnostic accuracy studies, however, the summary is a curve (or several curves if patient subsets are considered). Comparisons of curves are inherently more complex and nuanced than comparisons of means or proportions. Thus, systematic reviews of diagnostic accuracy present the research community with a challenging set of questions about how best to summarize the information and how to use it in analysis and decision making. For example, the methodology for incorporation of SROC curves in modeling outcomes and costs is not fully developed, and practical experience in this type of analysis is relatively scarce. In most published modeling exercises, the sensitivity and specificity of tests are assumed to be a single pair of numbers.

Two major determinants of the success of systematic reviews of diagnostic accuracy are the availability of relevant studies of adequate quality and the development of a consensus around the methods for such reviews. In recent years, the quality of diagnostic and screening test evaluations has improved, but the hill still seems steep [14, 39–41]. In the same period, the methods for systematic review of diagnostic accuracy have progressed and matured. Evidence of methodologic progress is the growing list of published work and the formation of the Cochrane Diagnostic Reviews initiative [42] late in 2003. The researchers involved in this initiative are at work preparing the methodologic infrastructure for performing diagnostic accuracy reviews and including them in a new division of the Cochrane Library.

Despite progress in study quality and reporting and in methodologic development, major challenges confront investigators venturing into the world of systematic reviews of diagnostic and screening tests. The following is a partial list of challenges:

- The literature contains many small studies, which are usually retrospective and of uncertain quality.
- The detail and accuracy of reporting on study methods and results vary greatly. It is often impossible to determine key study characteristics, such as study cohort, technical aspects of the techniques involved, and definition of gold standard information.
- Even for relatively tightly defined clinical questions, multiple sources of heterogeneity among studies are operating, threshold differences being only one. It is therefore important for the review to explore such sources of variation and to use appropriate statistical techniques.
- An important source of heterogeneity not addressed in this article is heterogeneity due to observer. Empiric data suggest that within-study observer variability can be of the same order of magnitude as variability across studies. Hierarchical modeling can be a powerful framework for incorporating observer variability in the analysis of individual studies [43, 44]. However, detailed data on observer variability are not usually reported, making it necessary for investigators to contact the authors of studies if such an analysis is to be undertaken.
- As is the case with most technology, diagnostic and screening techniques evolve rapidly. In the absence of a consensus on a framework for diagnostic technology assessment, there is risk of increasing the heterogeneity in a systematic review by inclusion of studies that clearly do not reflect the current state of a technique. By contrast, such a framework is in place for the evaluation of therapy. In that context, a systematic review, for example, would not combine estimates of effects reported in phase 1 and 2 studies with those reported in phase 3 studies.
- Particular forms of bias exist within many primary studies [45]. The effect of such within-study bias on systematic reviews has to be considered. Methods for handling bias within the primary studies need to be developed.

In confronting methodologic and practical challenges, investigators conducting systematic reviews of diagnostic accuracy are likely to find colleagues and collaborators. The era of evidence-based diagnosis is here to stay.

Acknowledgments

We thank the editors for inviting us to prepare this review and the referees for their comments and suggestions.

References

1. Knottnerus JA, ed. *The evidence base of clinical diagnosis.* London, UK: BMJ Books, 2002
2. Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. *AJR* 1994; 162:1–8
3. Gatsonis C. Design of evaluations of imaging technologies: development of a paradigm. *Acad Radiol* 2000; 7:681–683
4. Gatsonis C. Do we need a checklist for reporting the results of diagnostic test evaluations? *Acad Radiol* 2003; 10:599–600
5. Irwig L, Tosteson AN, Gatsonis CA, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994; 120:667–676
6. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995; 48:119–130
7. Pai M. Systematic reviews of diagnostic test evaluations: what's behind the scenes? *ACP J Club*

2004; 141:11–13

8. Gatsonis C, McNeil B. Collaborative evaluation of diagnostic tests: experience of the Radiologic Diagnostic Oncology Group. *Radiology* 1990; 175:571–575
9. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol* 2000; 53:65–69
10. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc* 2002; 9:653–658
11. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Technol Assess Health Care* 1996; 12:195–208
12. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003; 3:25
13. Bossuyt P, Reitsma J, Bruns D, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003; 49:7–18
14. Bossuyt P, Reitsma J, Bruns D, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41–44
15. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004; 140:189–202
16. Weinstein S, Obuchowski N, Lieber M. Clinical evaluation of diagnostic tests. *AJR* 2005; 184:14–19
17. Hanley J. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging* 1989; 29:307–335
18. Obuchowski N. ROC analysis. *AJR* 2005; 184:364–372
19. Zhou XH, Obuchowski N, McClish D. *Statistical methods in diagnostic medicine.* New York, NY: Wiley, 2002
20. Pepe M. *The statistical evaluation of medical tests for misclassification and prediction.* New York, NY: Oxford University Press, 2003
21. Toledano AY, Herman BA. Case study: evaluating accuracy of cancer diagnostic tests. In: Beam C, ed. *Biostatistical applications in cancer research.* Boston, MA: Kluwer, 2002:219–232
22. Toledano AY. Cancer diagnostics: statistical methods. In: Beam C, ed. *Biostatistical applications in cancer research.* Boston, MA: Kluwer, 2002:183–218
23. Moses LE, Littenberg B, Shapiro D. Combining independent studies of a diagnostic test into a summary ROC curve: data–analytic approaches and some additional considerations. *Stat Med* 1993; 12:1293–1316
24. Walter S. Properties of the SROC for diagnostic test data. *Stat Med* 2002; 21:1237–1256
25. Kardaun JWPF, Kardaun OJWF. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Methods Inf Med* 1990; 29:12–22
26. Rutter C, Gatsonis C. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995; 2:S48–S56
27. Rutter C, Gatsonis C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001; 20:2865–2884
28. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol* 2004; 57:925–932
29. *SAS / STAT 9.1 user's guide.* Cary, NC: SAS Institute, 2004
30. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. *JAMA* 1997; 278:1096–1101
31. Normand SL. Tutorial in biostatistics: meta-analysis—formulating, evaluating, combining, and reporting. *Stat Med* 1999; 18:321–359
32. Reitsma J, Glas A, Rutjes A, Scholten R, Bossuyt P, Zwinderman A. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005, 58:982–990
33. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics* 2003; 59:936–946
34. Kester ADM, Buntinx F. Meta-analysis of ROC curves. *Med Decis Making* 2000; 20:430–439
35. McClish DK. Combining and comparing area estimates across studies or strata, *Med Decis Making* 1992; 12:274–279
36. Zhou X. Empirical Bayes combination of estimated areas under ROC curves using estimating equations. *Med Decis Making* 1996; 16:24–28
37. Hellmich M, Abrams KR, Sutton AJ. Bayesian approaches to meta-analysis of ROC curves. *Med Decis Making* 1999; 19:252–264
38. DuMouchel W, Normand SL. Computer modeling and graphical strategies for meta-analysis. In: Stangle D, Berry D, eds. *Meta-analysis in medicine and health policy.* New York, NY: Dekker, 2000
39. Beam C, Sostman HD, Zheng J-Y. Status of clinical MR evaluations 1985–1988: baseline and design for further assessments. *Radiology* 1991; 180:265–270
40. Black WC. How to evaluate the radiology literature. *AJR* 1990; 154:17–22
41. Cooper LS, Chalmers TC, McCally M, Berrier J, Sacks HS. The poor quality of early evaluations of magnetic resonance imaging. *JAMA* 1988; 259:3277–3280
42. Cochrane reviews of diagnostic test accuracy. The Cochrane Collaboration Web site. Available at: www.cochrane.org/newslett/ccnews31-lowres.pdf. Accessed May 31, 2006
43. Gatsonis CA. Random effects models for diagnostic accuracy data. *Acad Radiol* 1995; 2:S14–S21
44. Ishwaran H, Gatsonis C. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Can J Stat* 2000; 28:731–750
45. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999; 282:1061–1066

Appendices continue on next page

APPENDIX 1: Binary Regression Model

For a single study, the model can be described as follows.

Let Y_{ij} represent the test result (1 = positive, 0 = negative) and D_{ij} the true disease status on the jth individual in the ith study. In our notation, we code $D = 1 / 2$, if diseased, and $-1 / 2$ if nondiseased. The binary regression model is based on the assumption that the response arises from the discretization of an underlying continuous latent variable with threshold θ_i. The latent variable follows logistic distributions for diseased and nondiseased subjects, and the two distributions can be distinguished by a location parameter (α_i) and a scale parameter (β_i). The diagnostic performance of the test in the ith study is a function of the location and the scale parameters. Formally,

$$\text{logit}P[Y_{ij} = 1 \mid D_{ij}] = (\theta_i - \alpha_i D_{ij})\exp(-\beta_i D_{ij})$$

The binary regression model is closely related to the usual ROC model and implies that for the ith study:

$$\text{logit}(\text{FPR}_i) = (\theta_i + \alpha_i / 2)\exp(\beta_i / 2)$$

$$\text{logit}(\text{TPR}_i) = (\theta_i - \alpha_i / 2)\exp(-\beta_i / 2)$$

If the location and scale parameters are assumed to be constant across all studies, the model reduces to a relation between the true-positive rate (TPR) and false-positive rate (FPR) that is similar to the relation postulated in the model described by Moses et al. [23]. In particular:

$$\text{logit (TPR}_i) = \text{c}_0 + \text{c}_1 \text{ logit (FPR}_i) \text{ or}$$
$$D_i = \frac{2c_0}{c_1 + 1} + \frac{c_1 - 1}{c_1 + 1}S_i$$

where $\text{c}_0 = -\alpha e^{-\beta / 2}$, $c_1 = e^{-\beta}$. It is clear that c_1 is greater than 0; and that b, which is equal to $(\text{c}_1 - 1) / (\text{c}_1 + 1)$, takes values between –1 and 1.

An SROC curve and its summary measures can be estimated from the binary regression model. In addition, study-level and subject-level covariates can be easily incorporated, resulting in models of the form:

$$\text{logit}P[Y_{ij} = 1 \mid D_{ij}, X_\text{i}] = (\theta_i - \alpha_i D_{ij} - \gamma X_\text{i})\exp(-\beta_i D_{ij}),$$

which corresponds to simultaneously fitting several SROC curves to subsets of the data.

The large number of parameters in the binary regression model creates identifiability problems without additional assumptions. For example, the model without covariates has three parameters for each table; hence, it is not identifiable for a single table. However, with suitable assumptions, such as the one leading to the analogue of the Moses model, the binary regression model can be made identifiable. Other assumptions about the parameters allow the exploration of heterogeneity across studies. For example, studies may have different location parameters (thus different overall accuracies) but the same scale parameter and the same threshold. Such exploration of heterogeneity is rather limited within the fixed-effects type of approach we present in this article. More elaborate exploration of heterogeneity requires the use of hierarchical models.

APPENDIX 2: Software for Fitting a Binary Regression Model

```
   SAS Code for Binary Regression Model (for CT);
data binreg1;
input study test n_pos n_tp dis dis1;
cards;
1 1 10 8 1 0.5
………………………………………………
42 3 24 2 zero −0.5

   data final; set binreg1; if test=2;
/* create indicator variable for each study */
if study=1 then s1=1; else s1=0;
...............
if study=42 then s42=42; else s42=0;
   run;
   proc nlmixed data=final1 maxiter=5000 cov ;
   parms t18=0 t19=0 t20=0 t21=0 t22=0 t23=0 t24=0 t25=0 t26=0 t27=0 t28=0 t29=0 t30=0 t31=0 t32=0 t33=0 t34=0 t35=0 t36=0 ;
logitp=(t18*s18+t19*s19+t20*s20+t21*s21+t22*s22+t23*s23+t24*s24+t25*s25+t26*s26+t27*s27+t28*s28+t29*s29+t30*s30+t31*s31+
t32*s32+t33*s33+t34*s34+t35*s35+t36*s36–a*dis1) / exp (b*dis1) ;
p=exp(logitp) / (1+exp (logitp) ) ;
model n_tp~binomial (n_pos , p) ;
run;
```

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005
16. Statistical Inference for Proportions, April 2005
17. Reader Agreement Studies, May 2005
18. Correlation and Regression, July 2005
19. Survival Analysis, July 2005
20. Multivariate Statistical Methods, August 2005
21. Decision Analysis and Simulation Modeling for Evaluating Diagnostic Tests on the Basis of Patient Outcomes, September 2005
22. Radiology Cost and Outcomes Studies: Standard Practice and Emerging Methods, October 2005