Statistical Methods in Diagnostic Medicine

XIAO-HUA ZHOU
Indiana University

NANCY A. OBUCHOWSKI
The Cleveland Clinic Foundation

DONNA K. McCLISH
Virginia Commonwealth University

A JOHN WILEY & SONS, INC., PUBLICATION

This book is printed on acid-free paper. ♾

Published simultaneously in Canada.

For ordering and customer service, call 1-800-CALL-WILEY.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 0-471-34772-8

Printed in the United States of America

10 9 8 7 6 5 4 3 2

CHAPTER 1

Introduction

1.1 WHY THIS BOOK?

Diagnostic tests play an important role in medical care and contribute significantly to health care costs (Epstein, Begg, and McNeil, 1986), yet the quality of diagnostic test studies has been poor (Begg, 1987). Reid, Lachs, and Feinstein (1995) reviewed articles on diagnostic tests that were published between 1978 and 1993 and reported many errors in design and analysis. These errors have fostered distrust in the conclusions of diagnostic test studies and have contributed to misunderstandings in the selection and interpretation of diagnostic tests.

Some examples of common errors in diagnostic test studies help illustrate the problem. One common error involves how the diagnostic tests are interpreted. Many investigators of new diagnostic tests attempt to develop criteria for interpreting such tests based only on the test results of healthy volunteers. For example, for a new test to detect pancreatitis, investigators might measure the amount of a certain enzyme in healthy volunteers. A typical decision criterion, or *cutpoint*, is three standard deviations (SDs) from the mean. Patients with an enzyme level of three SDs below the mean of healthy volunteers are labeled *positive* for pancreatitis; patients with an enzyme level above this cutpoint are labeled *negative*. In proposing such a criterion, investigators fail to recognize

1. the relevance of natural distributions (i.e., are they Gaussian [normal]?);
2. the amount of potential overlap with test results from patients with the condition;
3. the clinical significance of diagnostic errors, both attributed to falsely labeling a patient without the condition as positive and a patient with the condition as negative; and
4. the poor generalization of results based on healthy volunteers.

In Chapter 2, we discuss factors involved in determining optimal cutpoints for diagnostic tests; in Chapter 4, we discuss methods of finding optimal cutpoints and estimating diagnostic errors associated with them.

Another common error in diagnostic test studies is the notion that making a rigorous assessment of a patient's true condition—with the exclusion of patients for whom a less rigorous assessment was made—allows for a scientifically sound study. An example comes from literature on the use of ventilation-perfusion lung scans for diagnosing pulmonary emboli. The ventilation-perfusion lung scan is a noninvasive test used to screen high-risk patients for pulmonary emboli; its accuracy in various populations is unknown. Pulmonary angiography, on the other hand, is a highly accurate test for diagnosing pulmonary emboli, but it is invasive. In a study that assesses the accuracy of ventilation-perfusion lung scans, the study sample usually consists of patients who have undergone both a ventilation-perfusion lung scan and a pulmonary angiogram, with the angiogram serving as the reference for estimating accuracy. (See Chapter 2 for the definition and some examples of *gold standards*.) Patients who undergo a ventilation-perfusion lung scan but not an angiogram would be excluded from such a study. This study design can lead to serious errors in test accuracy estimates. These errors occur because the study sample is not truly representative of the patient population undergoing ventilation-perfusion lung scans—rather patients with positive scans are often recommended for angiograms and patients with negative scans are often not sent for angiograms because of the unnecessary risks. In Chapter 3, we discuss *workup bias* and its most common form, *verification bias*, as well as the strategies to avoid them. In Chapter 10, we present statistical methods developed specifically to correct for verification bias.

Another error involves problems with agreement studies, in which investigators often draw conclusions about a new test's diagnostic capabilities based on how often it agrees with a conventional test. For example, digital mammography, a new method of acquiring images of the breast for screening and diagnosis, has many advantages over conventional film mammography, including easy storage and transfer of images. In a study comparing these two tests on a sample of patients, if the results agree often, we will be encouraged by the new test. But what if the digital and film results do not agree often? It is incorrect for us to conclude that digital mammography has inferior accuracy. Clearly, if digital mammography has better accuracy than film mammography, then the two tests will not agree. Similarly, the two tests can have the same accuracy but make mistakes on different patients, resulting in poor agreement. A more valid approach to assessing a new test's diagnostic worth is to compare both tests against the true diagnoses of the patients to estimate and compare the accuracy of both tests. Assessment of diagnostic accuracy is usually more difficult than assessment of agreement, but it is a more relevant, valid approach (Zweig and Campbell, 1993). In Chapter 5, we present methods for comparing the accuracy of two tests when the true diagnoses of the patients are known; in Chapter 11, we present methods for comparing the accuracy of two tests when the true diagnoses are unknown.

There is no question that studies of diagnostic test accuracy are challenging to design and require specialized statistical methods for their analysis. There

are few good references and no comprehensive sources of information on how to design and analyze diagnostic test studies. This book fulfills this need. In it, we present and illustrate concepts and methods for designing, analyzing, interpreting, and reporting studies of the diagnostic test accuracy. In Part I (Chapters 2–7), we define various measures of diagnostic accuracy, describe strategies for designing diagnostic accuracy studies, and present basic statistical methods for estimating and comparing test accuracies, calculating sample sizes, and synthesizing literature for meta-analysis. In Part II (Chapters 8–12), we present more advanced statistical methods of describing a test's accuracy when patient characteristics affected it, of analyzing multireader studies and studies with verification bias or imperfect gold standards, and of performing meta-analyses.

1.2 WHAT IS DIAGNOSTIC ACCURACY?

A diagnostic test has two purposes (Sox, Jr. et al., 1989): (1) to provide reliable information about the patient's condition and (2) to influence the health care provider's plan for managing the patient. McNeil and Adelstein (1976) added a third possible purpose: to understand disease mechanisms and natural history through research (e.g., the repeated testing of patients with chronic conditions). A test can serve these purposes only if the health care provider knows how to interpret it. This information is acquired through an assessment of the test's diagnostic accuracy, which is simply the ability of a test to discriminate among alternative states of health (Zweig and Campbell, 1993). Although frequently there are more than two states of health, the clinical question can often be appropriately dichotomized (e.g., the presence or absence of Parkinson's disease or the presence or absence of an invasive carcinoma). In this book, we consider these types of situations (i.e., the binary health states).

In assessing the performance of the diagnostic test, we want to know if the test results differ for the two health states. If they do not differ, then the test has negligible accuracy; if they do not overlap for the two health states, then the test has perfect accuracy. Most test accuracies fall between these two extremes. The most important error to avoid is the assumption that a test result is a true representation of the patient's condition (Sox, Jr. et al., 1989). Most diagnostic information is imperfect; it may influence the health care provider's thinking, but uncertainty will remain about the patient's true condition. If the test is negative for the condition, should the health care provider assume that the patient is disease-free and thus send him or her home? If the test is positive for the condition, should the health care provider assume the patient has the condition and thus begin treatment? And if the test result requires interpretation by a trained reader (e.g., a radiologist), should the health care provider get a second opinion of the interpretation?

To answer these critical questions, the health care provider needs to have information on the test's absolute and relative capabilities and an understanding

of the complex interactions between the test and the trained readers (Beam et al., 1992). The health care provider must ask, How does the test perform among patients with the condition (i.e., the test's sensitivity)? How does the test perform among patients without the condition (i.e., the test's specificity)? Does the test serve to replace an older test, or should multiple tests be performed? If multiple tests are performed, how should they be executed (i.e., sequentially or in parallel)? How reproducible are interpretations by different readers?

Radiographic image quality is often confused with diagnostic accuracy. As noted by Lusted (1971), an image can reproduce the shape and texture of tissues most faithfully from a physical standpoint, but it may not contain useful diagnostic information. Fryback and Thornbury (1991) described a working model for assessing the efficacy of diagnostic tests in medicine. The model delineates image quality, diagnostic accuracy, treatment decisions, and patient outcome and describes how these conditions relate to the assessment of a diagnostic test. Expanding upon other works (Cochrane, 1972; Thornbury, Fryback, and Edwards, 1975; McNeil and Adelstein, 1976; Fineberg, 1978), Fryback and Thornbury (1991) proposed the following 6-level hierarchical model. Level 1, at the bottom, is *technical efficacy*, which is measured by such features as image resolution and sharpness for radiographic tests and optimal sampling times and doses for diagnostic marker tests; level 2 is *diagnostic accuracy efficacy*, that is, the sensitivity, specificity, and receiver-operating characteristic (ROC) curve; level 3 is *diagnostic thinking efficacy*, which can be measured, for example, by the difference in the clinician's estimated probability of a diagnostic before versus after the test results are known; level 4 is *therapeutic efficacy*, which can be measured by the percentage of time that therapy planned before the diagnostic test is altered by the results of the test; level 5 is *patient outcome efficacy*, which can be defined, for example, by the number of deaths prevented, or a change in the quality life because of, the test information; and level 6, at the top, is *societal efficacy*, which is often described by the cost-effectiveness of the test as measured from a societal perspective. A key feature of this model is that for a diagnostic test to be efficacious at a higher level, it must be efficacious at all lower levels. The reverse is not true; that is, the fact that a test can be efficacious at one level does not guarantee that it will be efficacious at higher levels. In this book, we deal exclusively with the assessment of diagnostic accuracy efficacy (level 2 of the hierarchical model), recognizing that it is only one step in the complete assessment of a diagnostic test's usefulness.

1.3 LANDMARKS IN STATISTICAL METHODS OF DIAGNOSTIC MEDICINE

In 1971, Lusted wrote a highly influential article in the journal *Science* in which he postulated that to measure the worth of a diagnostic test, one must measure the performance of the observers with the test. Lusted argued that ROC curves

provide an ideal means of studying observer performance. Lusted was writing about radiographic tests, but ROC curves are now used to assess diagnostic test accuracy in many disciplines of medicine.

An ROC curve is a plot of a diagnostic test's sensitivity (i.e., the test's ability to detect the condition of interest) versus its false-positive rate (i.e., the test's inability to recognize normal anatomy and physiology as normal). The curve illustrates how different criteria for interpreting a test produce different values for the test's false-positive rate and sensitivity.

ROC curves and their analyses are based on statistical decision theory; they were originally developed for electronic signal–detection theory (Peterson, Birdsall, and Fox, 1954; Swets and Pickett, 1982). They have been applied in many medical and nonmedical endeavors, including studies of human perception and decision making (Green and Swets, 1966), industrial quality control (Drury and Fox, 1975), and military monitoring (Swets, 1977).

Lusted (1971) indicated that in diagnostic medicine, as in electronic signal–detection theory, a distinction must be made between the criteria that an observer uses for deciding whether a condition is present or absent and the observer's abilities the sensory and cognitive attributes used for interpreting the test results) for detecting the condition. ROC curves can be used to make this distinction. Lusted gave the following example: Suppose that the six points in Fig. 1.1 represent the diagnoses of six different physicians. The physicians have identical sensory and cognitive abilities for detecting tuberculosis on a

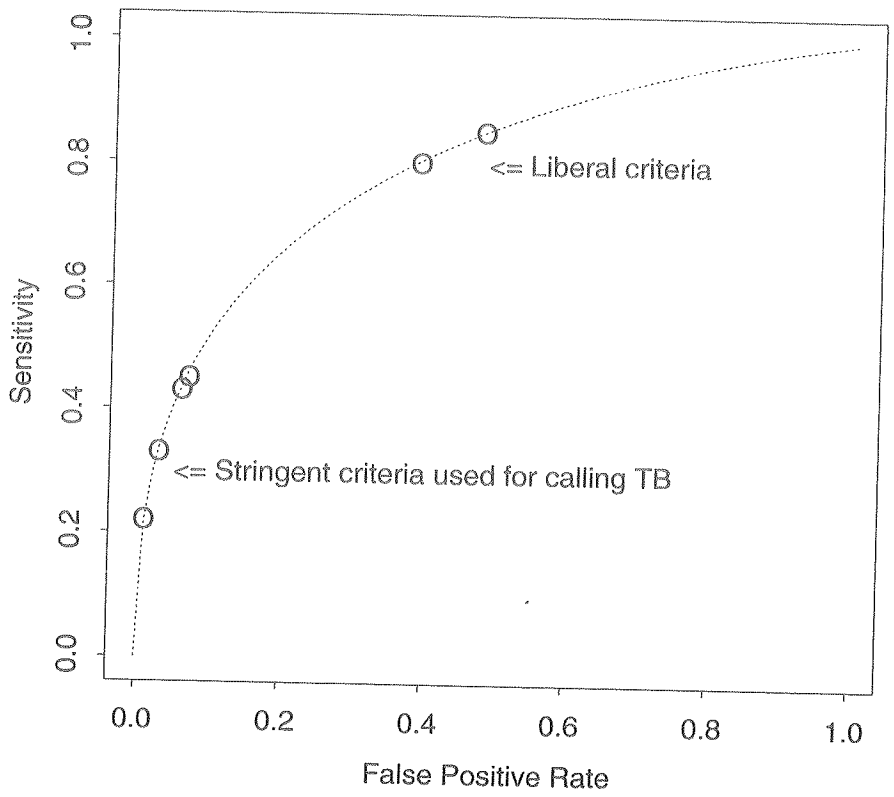

Figure 1.1 Lusted's (1971) example of tuberculosis (TB) detection.

chest radiograph, but they have different criteria for which densities actually should be called tuberculosis. The upper points on the curve represent individuals with more liberal decision criteria (i.e., the low-density nodules are called *positive*), whereas the lower points on the curve represent individuals with more stringent criteria (i.e., only high-density nodules are called *positive*). In diagnostic medicine, we are interested in measuring the observer's abilities for interpreting test results rather than his or her criteria for decisions.

Swets and Pickett (1982) noted two other key features of ROC curves that make them ideal for studying diagnostic tests. First, the curves display all possible cutpoints and thus supply estimates of the frequency of various outcomes (i.e., true positives, true negatives, false positives, and false negatives) at each cutpoint. (See Chapter 2 for definitions.) Second, the curves allow the use of previous probabilities of the condition, as well as calculations of the costs and benefits of correct and incorrect decisions, to determine the best cutpoint for a given test in a given setting. (See Chapters 2 and 4.)

Green and Swets (1966) were first to use the Gaussian model for estimating the ROC curve. They assumed that the various sensory events (i.e., test results) could be mapped on a single line. The numerical value of an observed event (call it T) affects the observer's confidence about whether the condition is present or absent. They also assumed that a cutpoint c is present so that if $T < c$, the observer will choose the hypothesis that the condition is absent, whereas if $T > c$, the observer will choose the hypothesis that the condition is present. In addition, they assumed that the distribution of T is Gaussian under each hypothesis. Following these assumptions, Dorfman and Alf, Jr. (1968, 1969) developed maximum-likelihood estimates (MLEs) for the parameters of a binormal (i.e., two Gaussian distributions, usually overlapping) ROC curve, along with procedures for obtaining the variance–covariance matrix and confidence intervals. (See Chapter 4.) Also, they wrote a FORTRAN program called RSCORE to perform the MLE.

A decade later, Metz (1978) and Swets and Pickett (1982) described, in practical terms, how to design ROC curve studies and analyze the data, with particular emphasis on the area under the ROC curve as the measure of test accuracy. The MLE software RSCORE was modified and extended by Metz and colleagues. The FORTRAN programs—including ROCFIT, LABROC, CORROC, and CLABROC—written by Metz and colleagues are today commonly used to estimate and compare ROC curves based on the binormal model.

A pivotal paper was the article written by Hanley and McNeil (1982), which provided a computationally simple method of estimating the area under the ROC curve without any assumptions about the distribution of the test results. The paper also noted an interesting equivalence, first described by Bamber (1975)—that the quantity of area under the ROC curve is the same as that estimated by the Wilcoxon 2-sample test, a well-known nonparametric test statistic. This equivalence led to a simple interpretation for the area under the ROC curve, which is now used extensively. Another key development in the Hanley and McNeil (1982) article was a method of calculating sample size for

studies using the ROC area. Other nonparametric methods for estimating and comparing ROC curves have since been published (see Chapters 4 and 5), and several methods for sample size estimation now exist (see Chapter 6).

Swets and Pickett (1982) were the first to tackle the analysis of multireader studies, where typically several observers interpret the test results of the same sample of patients. They identified several sources of variability and correlations in multireader studies and proposed a method for estimating and comparing the test accuracy for multireader studies by estimating the different variance components and correlations. Several methods for analyzing multireader studies are now available. (See Chapter 9.)

Tosteson and Begg (1988) were the first to describe how general regression models for ordinal data can be used to estimate ROC curves. These regression models could be used to understand the effect of covariates (e.g., a patient's age and gender) on the test's accuracy. Since their 1988 article, new regression approaches and extensions of their basic model have been developed. (See Chapters 8 and 9.)

McClish (1989), recognizing that the ROC curve area is a global measure of a test's accuracy because it includes the entire range of false-positive rates from 0.0 to 1.0, developed parametric methods for estimating and comparing the partial areas under the ROC curve. These methods are based on a binormal model and parallel the commonly used MLEs of the area under the total ROC curve. (See Chapters 2, 4, and 5.)

Parallel to these landmarks in analyzing diagnostic accuracy data, Ransohoff and Feinstein (1978) were investigating issues of study design. They identified two common problems that can occur in sensitivity-and-specificity estimates of a diagnostic test: First, unless a broad spectrum of patients is chosen both with and without the condition, the study may yield falsely high sensitivity-and-specificity estimates, known as *spectrum bias*, and second, unless the interpretation of the test and the establishment of the true diagnosis are done independently, bias can falsely elevate the test's estimated accuracy, a problem known as *workup bias*. They illustrated these problems with several real examples of diagnostic tests that initially were found to be valuable in biased studies but later found to be useless. Since Ransohoff and Feinstein's investigations, many other problems have been identified in diagnostic test accuracy studies. (See Chapter 3.)

Many statistical methods that correct biased data were developed shortly after these investigations. For instance, in 1980 Hui and Walter proposed a method of estimating the sensitivity and specificity of a diagnostic test when the standard test against which it is compared has unknown error rates, a condition known as *imperfect gold standard bias*, and in 1983 Begg and Greenes developed a method to remove the effect of *verification bias* on estimates of sensitivity and specificity. From these articles evolved many other approaches to solving for imperfect gold standard bias (see Chapter 11) and verification bias (see Chapter 10).

Statistical methods for synthesizing diagnostic test accuracy studies (i.e.,

meta-analysis) have been developed more recently. Summary receiver-operating characteristic (SROC) curves were proposed by Littenberg, Moses, and Rabinowitz (1990) as a means of summarizing a test's sensitivity and specificity from multiple studies without the assumption (usually invalid; see Chapter 7) that all of the studies used the same cutpoint. New methods based on the SROC curve have since been developed. (See Chapter 12.)

1.4 SOFTWARE

Software to implement many of the statistical methods discussed in this book is available free of charge. Some of this software is in the format of FORTRAN programs; others are in the form of SAS macros (SAS Institute, Cary, North Carolina, USA). The authors have prepared a Web site that contains, links, or cites useful software relevant to statistical methods for diagnostic medicine; it is http://www.wiley.com/statistics and will be maintained and updated periodically for at least five years after this book's publication date.

1.5 TOPICS NOT COVERED IN THIS BOOK

Although this book covers the main themes in statistical methods of diagnostic medicine, it does not cover several related topics, as discussed in the following paragraphs.

In this book we discuss how ROC curves can be used to describe and compare the accuracies of diagnostic tests. An ROC curve and, in particular, an ROC area are also used to assess the predictive ability of a fitted model. For example, in SAS's PROC LOGISTIC, the *c*-statistic is reported; it is equivalent to the nonparametric estimate of the area under the ROC curve and used in PROC LOGISTIC to describe how well a fitted model discriminates between the two groups in the model. For more information on this particular use of ROC curves, see Harrell, Jr., Lee, and Mark (1996) and Hosmer and Lemeshow (2000).

Decision analysis, cost-effectiveness analysis, and cost-benefit analysis are methods commonly used to quantify the long-term, or downstream, effects of a test on the patient and society. In Chapters 2 and 4, we discuss how these methods can be applied to find the optimal cutpoint on the ROC curve. Description of how to perform these methods, however, is beyond the scope of this book. There are many excellent references on these topics, including Pauker and Kassirer (1975); Weinstein et al. (1980, 1996); Russell et al. (1996); and Gold et al. (1996).

We focus mainly on the assessment of diagnostic tests. However, many tests are used for screening asymptomatic people and for surveillance of patients with known disease. Many of the methods described in this book are applicable to these tests, but there are many issues specific to these applications

that are not covered here. For these issues, see, for example, Morrison (1992), Murtaugh (1995), and Black and Welch (1997).

Most of the methods we present for estimation and hypothesis testing are from a frequentist perspective. Bayesian methods can also be used, whereby one incorporates into the assessment of the diagnostic test some previously acquired information or expert opinion about a test's characteristics or information about the patient or population. Examples of Bayesian methods used in diagnostic testing are found in Hellmich et al. (1988); Gatsonis (1995); Joseph, Gyorkos, and Coupal (1995); Peng and Hall (1996); and O'Malley et al. (2001).

We present methods for a situation in which the condition status of a patient can be described by one of two states (e.g., Parkinson's disease—present or absent). In some situations, however, there are more than two truth states (e.g., chest radiograph findings of pneumothorax, interstitial disease, nodules, or normal). Some relevant references on the assessment of diagnostic test accuracy for multiple truth states are found in Steinbach and Richter (1987); Rockette (1994); Mossman (1999); and Obuchowski, Lieber, and Powell (2001).

We do not discuss regulatory requirements for the assessment of diagnostic tests. Such requirements can be found at Web sites maintained by the appropriate regulatory agency.

Finally, when multiple diagnostic tests are performed on a patient, one may want to combine the information from the tests to make the best possible diagnosis. See, for example, Pepe and Thompson (2000) for various methods for combining the results of tests to optimize diagnostic accuracy.

1.6 SUMMARY

Health care providers need to understand how to select and interpret diagnostic tests. However, much of the current literature on diagnostic test assessment is of poor quality, leading to misunderstanding and distrust. Considerable research has been done on methods for design, analysis, and interpretation of diagnostic test accuracy. This book provides a comprehensive, illustrative approach to these methods.

We note that statistical methods for the assessment of diagnostic tests are developed, modified, and extended constantly. Like the reader, the authors look forward to many advances in this field that extend beyond the coverage of this book.

REFERENCES

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating graph, *J. Math. Psych.* **12:** 387–415.

Beam, C. A., Baker, M. E., Paine, S. S., Sostman, H. D., and Sullivan, D. C. (1992). Answering unanswered questions: Proposal for a shared resource in clinical diagnostic radiology research, *Radiology* **183:** 619–620.

Begg, C. B. (1987). Biases in the assessment of diagnostic tests, *Stat. Med.* **6:** 411–423.

Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias, *Biometrics* **39:** 207–215.

Black, W. C. and Welch, H. G. (1997). Screening for disease, *AJR Am. J. Roentgenol.* **168:** 3–11.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services*, The Nuffield Provincial Hospital Trust, London.

Dorfman, D. D. and Alf, Jr., E. (1968). Maximum-likelihood estimation of parameters of signal-detection theory—a direct solution, *Psychometrika* **33:** 117–124.

Dorfman, D. D. and Alf, Jr., E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data, *J. Math. Psych.* **6:** 487–496.

Drury, C. G. and Fox, J. G. (1975). *Human reliability in quality control*, Halsted, New York.

Epstein, A. M., Begg, C. B., and McNeil, B. J. (1986). The use of ambulatory testing in prepaid and fee-for-service group practices, *N. Engl. J. Med.* **314:** 1089–1094.

Fineberg, H. V. (1978). Evaluation of computed tomography: Achievement and challenge, *AJR Am. J. Roentgenol.* **131:** 1–4.

Fryback, D. G. and Thornbury, J. R. (1991). The efficacy of diagnostic imaging, *Med. Decis. Making* **11:** 88–94.

Gatsonis, C. A. (1995). Random-effects models for diagnostic accuracy data, *Acad. Radiol.* **2:** S14–S21.

Gold, M. R., Siegel, J. E., Russell, L. B., and Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine*, Oxford University Press, New York.

Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psychophysics*, John Wiley and Sons, New York.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143:** 29–36.

Harrell, Jr., F. E., Lee, K. L., and Mark, D. B. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Stat. Med.* **15:** 361–387.

Hellmich, M., Abrams, K. R., Jones, D. R., and Lambert, P. C. (1988). A Bayesian approach to a general regression model for ROC curves, *Med. Decis. Making* **18:** 436–443.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*, John Wiley and Sons, New York.

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests, *Biometrics* **36:** 167–171.

Joseph, L., Gyorkos, T. W., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard, *Am. J. Epidemiol.* **3:** 263–272.

Littenberg, B., Moses, L. E., and Rabinowitz, D. (1990). Estimating diagnostic accuracy from multiple conflicting reports: A new meta-analytic method, *Clin. Res.* **138:** 415a.

Lusted, L. B. (1971). Signal detectability and medical decision making, *Science* **171:** 1217–1219.

McClish, D. K. (1989). Analyzing a portion of the ROC curve, *Med. Decis. Making* **9:** 190–195.

McNeil, B. J. and Adelstein, S. J. (1976). Determining the value of diagnostic and screening tests, *J. Nucl. Med.* **17:** 439–448.

Metz, C. E. (1978). Basic principles of ROC analysis, *Semin. Nucl. Med.* **8:** 283–298.

Morrison, A. S. (1992). *Screening in chronic disease*, Oxford University Press, New York.

Mossman, D. (1999). Three-way ROCs, *Med. Decis. Making* **19:** 78–89.

Murtaugh, P. A. (1995). ROC curves with multiple marker measurements, *Biometrics* **51:** 1514–1522.

Obuchowski, N. A., Goske, M. J., and Applegate, K. E. (2001). Assessing physicians' accuracy in diagnosing pediatric patients with acute abdominal pain: Measuring accuracy for multiple diseases, *Stat. Med.* **20:** 3261–3278.

O'Malley, A. J., Zou, K. H., Fielding, J. R., and Tempany, C. M. C. (2001). Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: Prostate biopsy and spiral CT of ureteral stones, *Acad. Radiol.* **8:** 713–725.

Pauker, S. G. and Kassirer, J. P. (1975). Therapeutic decision making: A cost-benefit analysis, *N. Engl. J. Med.* **293:** 229–234.

Peng, F. and Hall, W. J. (1996). Bayesian analysis of ROC curves using Markov-chain Monte Carlo methods, *Med. Decis. Making* **16:** 404–411.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy, *Biostatistics* **1:** 123–140.

Peterson, W. W., Birdsall, T. G., and Fox, W. C. (1954). The theory of signal detection theory, *Transactions of the IRE Professional Group on Information Theory*, 171–212.

Ransohoff, D. J. and Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests, *N. Engl. J. Med.* **299:** 926–930.

Reid, M. C., Lachs, M. S., and Feinstein, A. R. (1995). Use of methodologic standards in diagnostic test research: Getting better but still not good, *JAMA* **274:** 645–651.

Rockette, H. E. (1994). An index of diagnostic accuracy in the multiple disease setting, *Acad. Radiol.* **1:** 283–286.

Russell, L. B., Gold, M. R., Siegel, J. E., Daniels, N., and Weinstein, M. C. (1996). The role of cost-effectiveness analysis in health and medicine, *JAMA* **276:** 1172–1177.

Sox, Jr., H. C., Blatt, M. A., Higgins, M. C., and Marton, K. I. (1989). *Medical decision making*, Butterworths-Heinemann, Boston.

Steinbach, W. R. and Richter, K. (1987). Multiple classification and receiver operating characteristic (ROC) analysis, *Med. Decis. Making* **7:** 234–237.

Swets, J. A. (1977). *Vigilance: Relationships among theory, physiological correlates and operational performance*, Plenum, New York.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*, Academic Press, New York.

Thornbury, J. R., Fryback, D. G., and Edwards, W. (1975). Likelihood ratios as a mea-

sure of diagnostic usefulness of excretory urogram information, *Radiology* **141:** 561–565.

Tosteson, A. A. N. and Begg, C. B. (1988). A general regression methodology for ROC curve estimation, *Med. Decis. Making* **8:** 204–215.

Weinstein, M. C., Fineberg, H. V., Elstein, A. S., Frazier, H. S., Neuhauser, D., Neutra, R. R., and McNeil, B. J. (1980). *Clinical decision analysis*, WB Saunders, Philadelphia.

Weinstein, M. C., Siegel, J. E., Gold, M. R., Kamlet, M. S., and Russell, L. B. (1996). Recommendations of the panel on cost-effectiveness in health and medicine, *JAMA* **276:** 1253–1258.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine, *Clin. Chem.* **39:** 561–577.

PART I

Basic Concepts and Methods

CHAPTER 2

Measures of Diagnostic Accuracy

In this chapter, we describe several measures of the accuracy of diagnostic tests. In Sections 2.1–2.9, we discuss measures of *intrinsic accuracy*, a test's abilities to correctly detect a condition when it is actually present and to correctly rule out a condition when it is truly absent. These attributes are fundamental and inherent to diagnostic tests themselves.

The intrinsic accuracy of a test is measured by comparing the test results to the true condition status of the patient. We assume that the true condition status is one of two mutually exclusive states: "*the condition is present*" or "*the condition is absent.*" Some examples are the presence versus the absence of Parkinson's disease, the presence of a malignant versus a benign tumor, and the presence of one versus more than one tumor. We determine the true condition status by means of a *gold standard*. A gold standard is a source of information completely different from the test or tests under evaluation and which tells us the true condition status of the patient. Different gold standards are used for different tests and applications; some common examples are autopsy reports, surgery findings, pathology results from biopsy specimens, and the results of other diagnostic tests that have perfect or nearly perfect accuracy. In Chapter 3, we discuss more about the selection of a gold standard; in Chapter 11, we present statistical methods for measuring diagnostic accuracy without a gold standard.

Once a test is shown to have some level of intrinsic accuracy, the role of that test in particular clinical situations must be evaluated. At this stage, we consider not only the intrinsic accuracy of the test but also the prevalence and nature of the disease, the patient characteristics, and the consequences of the test's misdiagnoses. In Sections 2.10 and 2.11, we discuss the application of diagnostic tests in clinical scenarios.

2.1 SENSITIVITY AND SPECIFICITY

Two basic measures of diagnostic accuracy are *sensitivity* and *specificity*. Their definitions are best illustrated by a contingency table with 2 rows and 2

columns, or *decision matrix*, where the rows summarize the data according to the true condition status of the patients and the columns summarize the test results. We denote the true condition status by the indicator variable D, where $D = 1$ if the condition is present and 0 if the condition is absent. Test results indicating the condition's presence are called *positive*; those indicating its absence, *negative*. We denote positive test results as $T = 1$, negative test results as $T = 0$. Table 2.1 has such characteristics; it is called a *count* table because it indicates the number of patients in various categories. The total number of patients with and without the condition is, respectively, n_1 and n_0; the total number of patients with the condition who test positive and negative is, respectively, s_1 and s_0; and the total number of patients without the condition who test positive and negative is, respectively, r_1 and r_0. The total number of patients in the study group, N, is expressed as $N = s_1 + s_0 + r_1 + r_0$.

The sensitivity (Se) of a test is its ability to detect the condition when it is present. We write sensitivity as $Se = P(T = 1|D = 1)$, which is read, "sensitivity (Se) is the probability (P) that the test result is positive ($T = 1$), given that the condition is present ($D = 1$)." Among the n_1 patients with the condition, s_1 test positive; thus $Se = s_1/n_1$.

The specificity (Sp) of a test is its ability to exclude the condition in patients without the condition. We write specificity as $Sp = P(T = 0|D = 0)$, which is read, "specificity (Sp) is the probability (P) that the test result is negative ($T = 0$), given that the condition is absent ($D = 0$)." Among n_0 patients without the condition, r_0 test negative; thus $Sp = r_0/n_0$.

Count data can be summarized by probabilities, as shown in Table 2.2. This table emphasizes that sensitivity and specificity are computed from different subsamples of patients, that is, the subsamples of patients with and without the condition. Note that the sum of the two probabilities in the top row ($D = 1$) is one and, similarly, the sum of the two probabilities in the bottom row ($D = 0$) is one. The probability that the test will be positive in a patient with the condition (i.e., the sensitivity) is given in the ($D = 1$, $T = 1$) cell of the table.

Another way that diagnostic accuracy is commonly described emphasizes the consequences associated with the test results. In this use, sensitivity is the true-positive fraction (TPF) or rate (TPR); s_1 is the number of true positives (TPs). Specificity is the true-negative fraction (TNF) or rate (TNR); r_0 is the

Table 2.1 A Basic 2 × 2 Count Table

	Test Result		
True Condition Status	Positive ($T = 1$)	Negative ($T = 0$)	Total
Present ($D = 1$)	s_1	s_0	n_1
Absent ($D = 0$)	r_1	r_0	n_0
Total	m_1	m_0	N

Table 2.2 A 2 × 2 Probability Table

True Condition Status	Test Result: Positive ($T = 1$)	Test Result: Negative ($T = 0$)	Total
Present ($D = 1$)	$Se = s_1/n_1$	FNR = s_0/n_1	1.0
Absent ($D = 0$)	FPR = r_1/n_0	$Sp = r_0/n_0$	1.0

number of true negatives (TNs). The "true" positives and negatives are, respectively, s_1 and r_0, because the diagnostic test indicates the correct diagnosis. In contrast, s_0 is the number of false negatives (FNs), and s_0/n_1 is the false-negative fraction (FNF) or rate (FNR). Here, the test falsely indicates the absence of the condition in a patient who truly has the condition. False-negative results cause harm by delaying treatment and providing false reassurance. Similarly, r_1 is the number of false positives (FPs), and r_1/n_0 is the false-positive fraction (FPF) or rate (FPR). False detection of the condition leads to unnecessary, perhaps risky confirmatory tests, as well as incorrect treatment and false labeling of patients. An exercise for the reader is to verify that TPR + FNR = 1 and TNR + FPR = 1.

To illustrate the foregoing calculations, consider as an example a mammographer's diagnoses of 60 patients presenting for breast cancer screening. (See Table 2.3.) These data were part of a 7-reader retrospective study to investigate the accuracy of screening mammography (Powell et al., 1999). The study sample consisted of 30 patients with pathology-proven cancer and 30 patients with normal mammograms for two consecutive years. The mammogram was considered positive if the mammographer recommended additional diagnostic workup for the patient. Of the 30 patients with breast cancer, 29 tested positive—that is, they were correctly asked to return for additional workups. Thus there were 29 TPs and 1 FN; the sensitivity was 29/30 = 0.967. Of the 30 patients without breast cancer, 11 tested negative (TNs). The specificity was 11/30, or 0.367. The FPR was 19/30 = 0.633, or 1 − the specificity.

The definition of positive and negative test results as well as the condition of interest must be clear, because a positive finding may correspond to the

Table 2.3 Mammogram Results of 30 Patients With and 30 Patients Without Breast Cancer

Cancer Status	Test Result: Positive	Test Result: Negative	Total
Present	29	1	30
Absent	19	11	30
Total	48	12	60

presence or absence of a condition, depending on the clinical application. For example, in a study of lung disease (Remer et al., 2000), patients with detected adrenal adenomas were labeled positive and patients with detected lung metastases were labeled negative. The fact that patients with adrenal adenomas are eligible for lung cancer surgery, whereas patients without this condition (i.e., the patients with lung metastases) are not, motivated the authors of the study to refer to the detection of an adenoma as a positive finding.

Many diagnostic tests yield a numeric measurement as a result rather than a binary result (i.e., positive or negative). Consider a digital-imaging algorithm to identify patients whose implanted artificial heart valves have fractured (Powell et al., 1996). One measure used to distinguish fractured valves from intact valves is the width of the gap between the valve strut legs. The larger the gap, the likelier the valve has fractured. Table 2.4 lists the gap measurements of 20 patients who have undergone elective surgery for valve replacement; Fig. 2.1 illustrates the data. At surgery, 10 patients were found to have fractured valves and 10 were found to have intact valves; the gap values ranged from 0.03 to 0.58 for patients with fractured valves, 0.0 to 0.13 for patients with intact valves. To describe the sensitivity and specificity of the imaging technique, we choose a value of, say, 0.05, in which case the patients with gap values greater than 0.05 are labeled positive and patients with gap values less than or equal to 0.05 are labeled negative. The corresponding sensitivity and specificity are, respectively, 0.80 and 0.70.

In this example, we arbitrarily chose a gap value of 0.05 to define the test results as either positive or negative. The test result of 0.05 is called a *decision threshold*, the test result used as a cutoff to define positive and negative test results and, subsequently, to define sensitivity and specificity. We could have used any gap value as a decision threshold. Sensitivity and specificity would have been, however, affected by our choice.

Table 2.4 Gap Measurements of 10 Patients With and 10 Patients Without Fractured Heart Valves

Fractured	Intact
0.58	0.13
0.41	0.13
0.18	0.07
0.15	0.05
0.15	0.03
0.10	0.03
0.07	0.03
0.07	0.00
0.05	0.00
0.03	0.00

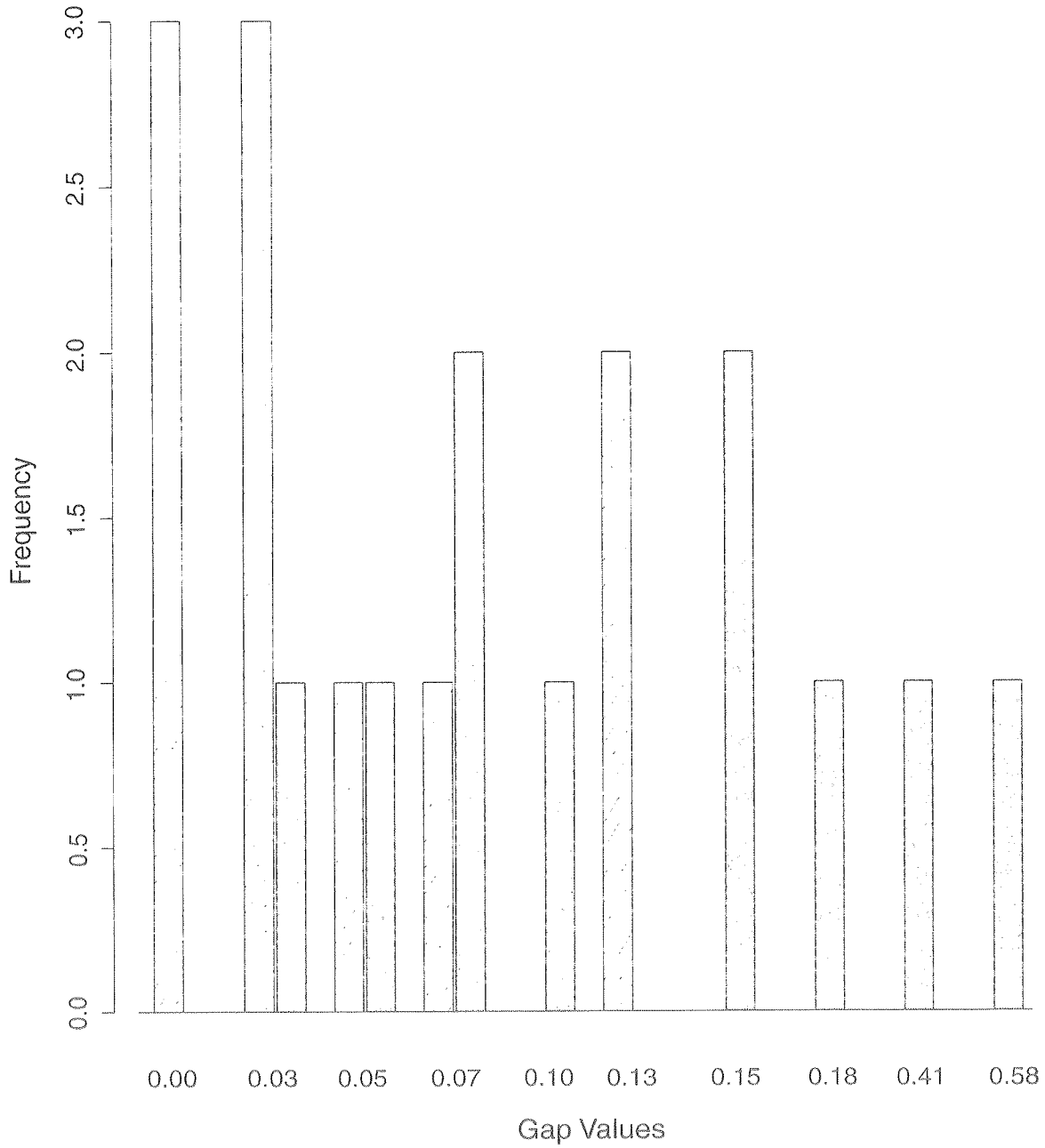

Figure 2.1 A histogram of gap measurements of patients with and without fractured heart valves.

Table 2.5 summarizes the sensitivity and specificity corresponding to several possible decision thresholds. If we choose a larger gap value of, say, 0.13, the sensitivity will decrease to 0.50 and the specificity will increase to 1.0. If, however, we choose 0.03, the sensitivity will increase to 0.90 and the specificity will decrease to 0.60. This example illustrates that the sensitivity and specificity of a test are inherently linked—as one increases, the other decreases.

Table 2.5 Estimates of *Se* and *Sp* From the Heart Valve–Imaging Study

Definition of + Test	*Se*	*Sp*	FNR	FPR
>0.58	0.0	1.0	1.0	0.0
>0.13	0.5	1.0	0.5	0.0
>0.07	0.6	0.8	0.4	0.2
>0.05	0.8	0.7	0.2	0.3
>0.03	0.9	0.6	0.1	0.4
>0.0	1.0	0.3	0.0	0.7
≥0.0	1.0	0.0	0.0	1.0

Thus in describing a diagnostic test, both sensitivity and specificity must be reported along with the corresponding decision threshold.

The gap measurement is an objective test result, calculated by a computer algorithm. Other tests yield results that must be interpreted subjectively, such as mammographic images for the detection of breast cancer or magnetic resonance (MR) images for the detection of multiple sclerosis. For these tests, the observer establishes a decision threshold in his or her mind and uses that threshold to label cases as positive or negative. The decision threshold that an observer adopts depends on many factors, including his or her "style," estimate of the condition's likelihood, and assessment of the consequences of misdiagnoses (Metz, 1978).

We might ask the mammographer, whose diagnoses were presented in Table 2.3, to use a stricter decision threshold to increase the specificity. The mammographer could reread the 60 cases, applying this new decision threshold, or, alternatively, he or she could assign a confidence score to each case to reflect his or her belief that the patient has the condition. In diagnostic radiology, two confidence scales are popular: an ordinal (rating) scale, which categorizes conditions as, for example, "definitely not present," "probably not present," "possibly present," "probably present," or "definitely present," and as a 0–100% scale, which describes the reader's confidence in the presence of the condition; 0% is no confidence, 100% is complete confidence in the presence of the condition. Certain tests have a specialized scale; for example, mammography uses the following rating scale: "normal," "benign," "probably benign," "suspicious," and "malignant." Table 2.6 summarizes the mammographer's results using this scale. If the mammographer uses a decision threshold at "suspicious" so that only cases assigned as "suspicious" or "malignant" are called positive, the corresponding sensitivity and specificity will be 0.767 and 0.733. (Note that the results in Table 2.3 were generated by using a decision threshold at "probably benign.") Here again, we see that an increase in specificity (from 0.367 in Table 2.3 to 0.733) was offset by a decrease in sensitivity (from 0.967 in Table 2.3 to 0.767).

Table 2.6 Mammogram Results Using a 5-Category Scale

	Test Result					
Cancer Status	Normal	Benign	Probably Benign	Suspicious	Malignant	Total
Present	1	0	6	11	12	30
Absent	9	2	11	8	0	30

Sensitivity and specificity are measures of intrinsic diagnostic accuracy because they are not affected by the prevalence of the condition. For example, in computing sensitivity in Table 2.3, it did not matter whether there were 30 or 30,000 patients without cancer; sensitivity is computed from only the subsample of patients with the condition, whereas specificity is computed from only the subsample of patients without the condition. Table 2.7 presents the test results of 3000 women—30 with cancer (as in Table 2.3) and 2970 without cancer—for a prevalence of 1%. The sensitivity is 0.967; the specificity, 0.367. These values are identical to the estimates from Table 2.3, where the prevalence was 50%. This property of sensitivity and specificity is important; in practical terms, it means that the sensitivity and specificity estimated from a study sample are applicable to other populations with different prevalence rates.

Although not affected by the prevalence of the condition, the sensitivity and specificity of some diagnostic tests are affected by the *spectrum of disease*. A disease's range of clinical severity or anatomic extent constitutes its spectrum. For example, large, palpable breast cancer tumors are easier to detect than sparse, dispersed malignant calcifications; thus mammography has greater sensitivity when it is applied to patients with advanced breast cancer. Similarly, patient characteristics affect the sensitivity and specificity of some diagnostic tests. Older women have fatty, less dense breasts than younger women, and mammography is better able to detect lesions in fatty breasts. In Chapter 3, we discuss more thoroughly the impact of the spectrum of disease and patient characteristics.

Table 2.7 Mammogram Results of 3000 Women

	Test Result		
Cancer Status	Positive	Negative	Total
Present	29	1	30
Absent	1881	1089	2970
Total	1910	1090	3000

Some interesting analogies are noted between Table 2.1 and the types I and II (or α and β) error rates used in statistical hypothesis testing. The type I (α) error rate is the probability of rejecting the null hypothesis when, in reality, the null hypothesis is true. The type II (β) error rate is the probability of failing to reject the null hypothesis when, in reality, the alternative hypothesis is true. In the diagnostic testing situation, let us define the null (H_0) and alternative (H_a) hypotheses as follows:

H_0: The condition is not present
H_a: The condition is present

Then, the type I error rate is analogous to the FPR and the type II error rate is analogous to the FNR. Statistical power, that is, 1 − type II error rate, is analogous to sensitivity. In statistical hypothesis testing, it is standard to set the type I error rate at 0.05 (5%). With diagnostic tests, however, the particular clinical application dictates the allowable error rates. (See Section 2.11.)

2.2 COMBINED MEASURES OF SENSITIVITY AND SPECIFICITY

It is often useful to summarize the accuracy of a test by a single number. For example, when comparing two tests, it is easier to compare a single number than to compare both the sensitivities and specificities of the two tests. There are several measures that incorporate sensitivity and specificity into a single index (accuracy, odds ratio, Youden's index). We start with a popular measure often referred to simply as *accuracy*; however, we refer to it more precisely as the *probability of a correct test result*. From Table 2.1, the probability of a correct test result is equal to $(s_1 + r_0)/N$ and constitutes the proportion of TPs and TNs in the entire sample. This measure is easily verified as a weighted average of sensitivity and specificity, with weights equal to the prevalence [that is, $P(D = 1)$] and to the complement of prevalence [that is, $P(D = 0)$] as follows:

$$P(\text{TP or TN}) = (\text{no. of correct decisions})/N = Se \times P(D = 1) + Sp \times P(D = 0)$$

The strength of this measure of accuracy is in its simple computation. However, this measure has many limitations, as illustrated by several examples. First, consider an 1885 editorial by Gilbert in which he writes about the extremely high "accuracy" of a fellow meteorologist in predicting tornadoes. Gilbert pointed out that because of the rarity of this meteorological event, high accuracy could be achieved simply by "calling" for "no tornado" every day.

As a second example, consider the mammography data in Tables 2.3 and 2.7. The sensitivity (0.907) and specificity (0.367) calculated from these two tables are the same, but the prevalence is different. From Table 2.3, the probability of a correct test result is (29 + 11)/60, or 0.667; from Table 2.7, the

probability of a correct test result is only 0.373. In Table 2.3, sensitivity and specificity are given equal weight, because the prevalence is 50%; in Table 2.7, specificity is given much more weight, because the prevalence is very low. This example illustrates that although sensitivity and specificity *are* measures of the intrinsic accuracy of a test, the probability of a correct test result is *not* a measure of intrinsic accuracy.

Another limitation of the probability of a correct result is that it is calculated based on only one decision threshold. However, there are many potential decision thresholds, and the clinical application should determine which of these is relevant. This also represents a limitation of single pairs of sensitivity and specificity.

Still another limitation of the probability of a correct result is that it treats FP and FN results as if they were equally undesirable, but often this is not the case (Zweig and Campbell, 1993). One might be tempted to use this measure to compare two tests applied to the same population. Metz (1978) indicated the problem with this use—that the two tests can have the same probabilities of a correct result but different sensitivities and specificities. For example, test *A* might have a sensitivity of 100% but a specificity of 0%; test *B* might have a specificity of 100% but a sensitivity of 0%. If the prevalence of the condition is 50%, both tests will yield the same probability of a correct result yet perform differently, and patient management will differ radically.

We mention two other measures here because they are sometimes used in meta-analyses of the accuracy of diagnostic tests. (See Chapter 12.) One is the *odds ratio*, defined as the odds of a positive test result relative to a negative test result among patients with the condition divided by the odds of a positive test result relative to a negative test result among patients without the condition. The odds ratio can be written as follows in terms of sensitivity and specificity:

$$\text{Odds ratio} = \frac{Se/(1 - Se)}{(1 - Sp)/Sp} = \frac{Se \times Sp}{\text{FNR} \times \text{FPR}}$$

For the data in Tables 2.3 and 2.7, the odds ratio is the same: 16.99. An odds ratio of 1.0 indicates that the likelihood of a positive test result is the same for patients with and without the condition (i.e., Se = FPR). Odds ratios greater than 1.0 indicate that the odds of a positive test result is greater for patients with the condition; odds ratios less than 1.0 indicate that the odds of a positive test result is greater for patients without the condition.

The other measure sometimes used in meta-analyses is *Youden's index:* $Se + Sp - 1$, or written another way, Se – FPR. It has a maximum value of 1.0 and a minimum value of 0.0, and it reflects the likelihood of a positive result among patients with versus without the condition.

Unlike the probability of a correct test result, the odds ratio and Youden's index are not dependent on the prevalence of the condition in the sample, for which reason they are superior summary measures of accuracy. However, both the odds ratio and the Youden's index share two limitations with the probability

of a correct result: First, they are based on only one decision threshold when, in reality, many potential decision thresholds exist, and second, they treat FP and FN results as equally undesirable. For example, suppose that test A has a sensitivity of 0.90 and a specificity of 0.40 and test B has a sensitivity of 0.40 and a specificity of 0.90. The odds ratio and Youden's index of both tests are equivalent at 6.0 and 0.3, respectively, yet the two tests have very different properties.

In later sections of this chapter, we discuss several other summary measures of accuracy that are superior to the probability of a correct test result, the odds ratio, and Youden's index. These measures are associated with the *receiver operating characteristic (ROC) curve*.

2.3 THE ROC CURVE

In 1971, Lusted described how a method used often in psychophysics could be adopted for medical decision making. This method overcomes the limitations of a single sensitivity and specificity pair and the summary measures associated with single sensitivity and specificity pairs by including all of the decision thresholds. A Receiver Operating Characteristic, or ROC curve is a method of describing the intrinsic accuracy of a test apart from the decision thresholds. Since the 1970s, it has been the most valuable tool for describing and comparing diagnostic tests.

An ROC curve is a plot of a test's sensitivity (plotted on the y axis) versus its FPR, or (1 − specificity)(plotted on the x axis). Each point on the graph is generated by a different decision threshold. We use line segments to connect the points from all the possible decision thresholds, forming an *empirical ROC curve*. We know that as the sensitivity increases, the FPR increases, and the ROC curve shows precisely the magnitudes of these increases.

Figures 2.2 and 2.3 illustrate the ROC curves for the heart valve–imaging data (Table 2.4) and mammography data (Table 2.6), respectively. In Figure 2.2, each circle on the empirical ROC curve represents a (FPR, Se) point corresponding to a different decision threshold. For example, the point at the far left (FPR = 0.0, Se = 0.5) corresponds to the decision threshold of >0.13. (See Table 2.5.) The point at the far right at (FPR = 0.7, Se = 1.0) corresponds to the decision threshold at >0.0. Line segments connect the points generated from all possible decision thresholds. In this example data, there are nine decision thresholds that provide unique (FPR, Se) points in addition to the two trivial points of (0, 0) and (1, 1).

In Table 2.6, there are $k = 5$ categories for the test results, that is, normal, benign, probably benign, suspicious, and malignant. In the corresponding empirical ROC curve (Fig. 2.3), there are $k - 1$ (or 4) nontrivial points connected with line segments. Point A on the curve, corresponding to the cutoff at the malignant category, is a strict threshold in that only cases judged malignant are considered positive. Point B corresponds to the cutoff at the suspicious

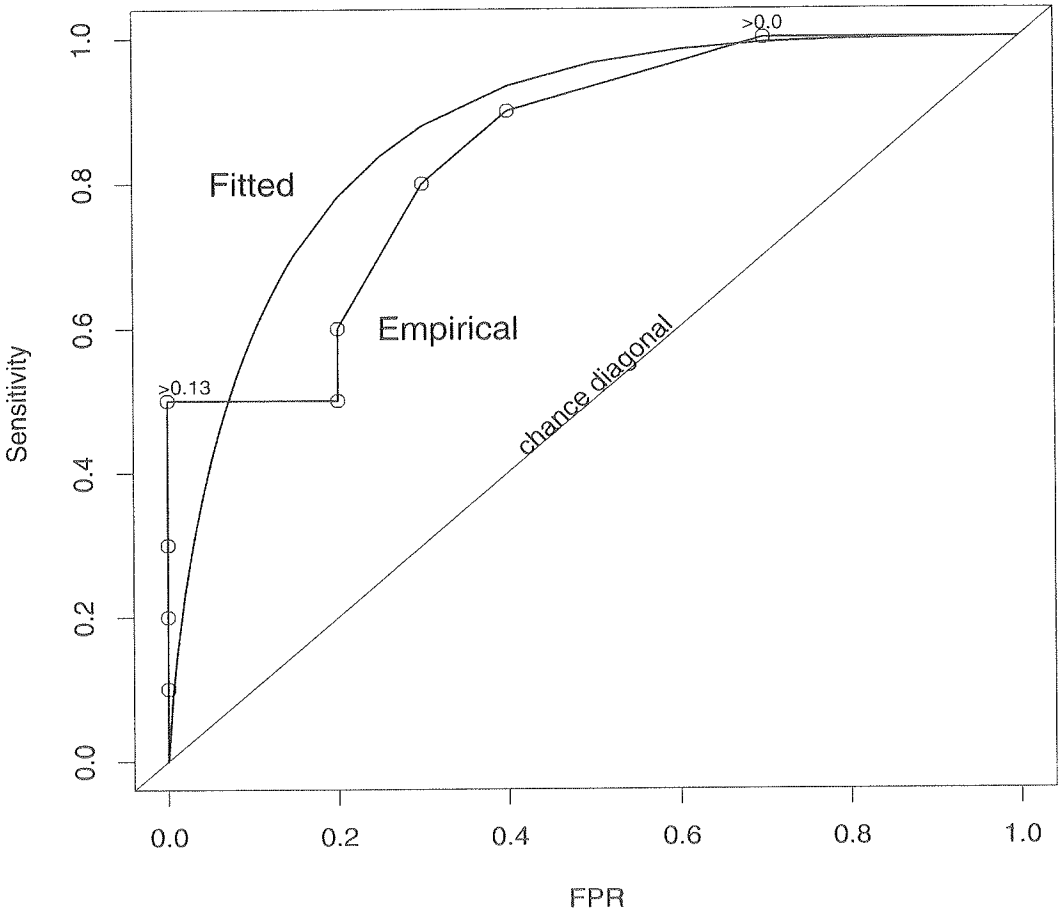

Figure 2.2 Empirical and fitted ROC curves for the heart valve–imaging data.

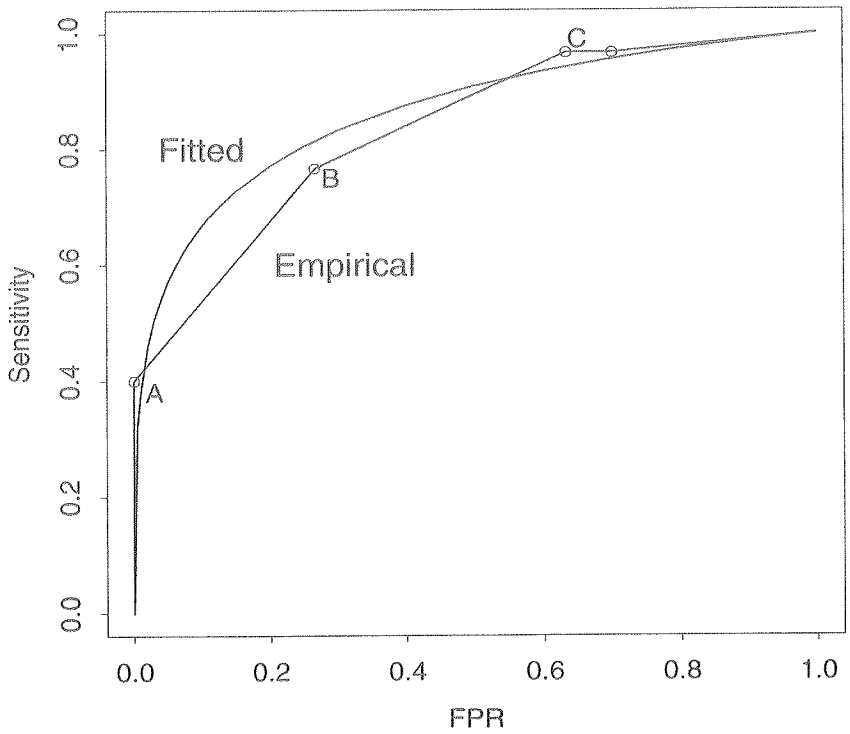

Figure 2.3 Empirical and fitted ROC curves for the mammography data.

category; it is a moderate threshold. Point *C* corresponds to the cutoff at the probably benign category; it is a lax threshold.

It is often convenient to fit a statistical model to the test results of a sample of patients. The *fitted ROC curves* (sometimes called *smooth curves*) for the heart valve–imaging test and mammography are also plotted in Figs. 2.2 and 2.3. The statistical model used is a binormal distribution (i.e., two Gaussian distributions: one for the test results of patients without fractured heart valves, the other for the test results of patients with fractured heart valves); it is the most commonly used model for fitting ROC curves in diagnostic medicine. When the binormal model is used, the curve is completely specified by two parameters. The first parameter, denoted as a, is the standardized difference in means of the distributions of test results for patients with and without the condition. The second parameter, denoted as b, is the ratio of the standard deviations (SDs) of the distributions of test results for patients without versus with the condition. In Chapter 4, we discuss the binormal model of ROC curves in detail; in this chapter, it is important to note that the intrinsic accuracy of a test is completely defined by its ROC curve, which in many cases can be defined by the two parameters a and b.

An ROC curve can be constructed from objective measurements of a test (e.g., the gap value from the digitized image of a heart valve), objective evaluation of image features (e.g., the attenuation coefficient from computed tomography), or subjective diagnostic interpretations (e.g., the 5-category scale used for mammographic interpretation) (Dwyer, 1997). The only requirement is that the measurements or interpretations can be meaningfully ranked in magnitude. With objective measurements, the decision variable is explicit, so one can choose from an infinite number of decision thresholds along the continuum of test results. For diagnostic tests interpreted subjectively, the decision thresholds are implicit or latent, for they exist only in the mind of the observer (Hanley, 1989). An essential assumption for the ROC curve is that these decision thresholds are the same for the subsamples of patients with and without the condition. When the decision thresholds are implicit, this assumption may need to be tested (Zhou, 1995). (See Chapter 4.) The concept of the ROC curve is the same whether the decision thresholds are explicit or implicit; the curve illustrates the trade-off between the sensitivity and the FPR as the decision threshold changes.

The name "receiver operating characteristic" curve comes from the notion that given the curve, we—the receivers of the information—can use (or operate at) any point on the curve by using the appropriate decision threshold. The clinical application determines which characteristics of the test are needed. Consider the heart valve–imaging data in Fig. 2.2. If the imaging technique is used to screen asymptomatic patients, we will want good specificity to minimize the number of FPs because surgery to replace the valve is risky. We might choose a cutoff at 0.07, where the FPR is 0.20 and the sensitivity is 0.50. On the other hand, if the imaging technique is used to diagnose patients with chest pain, a higher sensitivity will be needed. In this setting, a cutoff

at 0.03 is more appropriate, with a sensitivity of 0.90 and an FPR of 0.40. We discuss the optimal choice of operating points for particular applications in Section 2.11.

Most ROC curves are concave as in Figs. 2.2 and 2.3. Occasionally, however, a diagnostic test has an ROC curve with a "hook," defined as a portion of the ROC curve that lies below the chance diagonal (Pan and Metz, 1997). These curves are called *improper ROC curves* (Metz and Kronman, 1980); in Section 2.7, we discuss them in more detail when we introduce likelihood ratios.

In evaluating the accuracy of a test, a sensible person might ask if it is really necessary to generate a test's ROC curve. The ROC plot has many advantages over isolated measurements of sensitivity and specificity (Zweig and Campbell, 1993). In contrast with a figure such as Fig. 2.1, an ROC curve is a visual representation of accuracy data. The scales of the curve—sensitivity and FPR—are the basic measures of accuracy and are easily read from the plot; often, the values of the decision variable that generate the points are labeled on the curve. The ROC curve does not require selection of a particular decision threshold since all possible decision thresholds are included. Because sensitivity and specificity are independent of prevalence, so, too, is the ROC curve. Like sensitivity and specificity, however, the ROC curve and associated indices may be affected by the spectrum of disease as well as by patient characteristics. A good example is a test for fetal pulmonary maturity; for this test, the ROC curve is strongly affected by gestational age (Hunink et al., 1990).

Another advantage of the ROC curve is that it does not depend on the scale of the test results; that is, it is invariant to monotonic transformations of the test results, such as linear, logarithm, and square root (Campbell, 1994). In fact, the empirical curve depends only on the ranks of the observations, not on the actual magnitude of the test results.

Finally, the ROC curve provides a direct visual comparison of two or more tests on a common set of scales. It is difficult to compare two tests when there is only one sensitivity and specificity pair. The performance of one test is superior to another only if that test is more specific and more sensitive, equally specific and more sensitive, or equally sensitive and more specific. Even if one of these cases holds, however, it is difficult to determine how much better the test is when a change in the decision threshold occurs, because such a change may affect the two tests differently (Turner, 1978). By constructing the ROC curve, a comparison of tests at all decision thresholds is possible.

2.4 THE AREA UNDER THE ROC CURVE

As noted in Section 2.2, it is often useful to summarize the accuracy of a test by a single number. Several such summary indices are associated with the ROC curve, one of which is the *area under the ROC curve*, or just A.

The ROC curve area can take values between 0.0 and 1.0. An ROC curve

with an area of 1.0 consists of two line segments: (0, 0)–(0, 1) and (0, 1)–(1, 1). Such a test is perfectly accurate because the sensitivity is 1.0 when the FPR is 0.0. Unfortunately, such diagnostic tests are rare. In contrast, a test with an area of 0.0 is perfectly inaccurate; that is, patients with the condition are labeled incorrectly as negative and patients without the condition are labeled incorrectly as positive. If such a test existed, it would be trivial to convert it into one with perfect accuracy only by reversing the test results. The practical lower bound for the ROC curve area is 0.5. The (0, 0)–(1, 1) line segment has an area of 0.5; it is called the *chance diagonal*. If we relied on pure chance to distinguish patients with versus without the condition, the resulting ROC curve will fall along this diagonal line. (See Exercise 2.2 at the end of this chapter.)

Diagnostic tests with ROC curves above the chance diagonal have at least some ability to discriminate between patients with and without the condition. The closer the curve to the (0, 1) point (left upper corner), the better the test. As we discuss in Chapter 4, to statistically evaluate whether the ROC curve area differs from 0.5 is often appropriate. Rejection of this hypothesis implies that the test has some ability to discriminate between patients with versus without the condition.

The ROC curve area has several interpretations:

1. the average value of sensitivity for all possible values of specificity;
2. the average value of specificity for all possible values of sensitivity (Metz, 1986, 1989); and
3. the probability that a randomly selected patient with the condition has a test result indicating greater suspicion than that of a randomly chosen patient without the condition (Hanley and McNeil, 1982).

The third interpretation comes from work of Green and Swets (1966) and Hanley and McNeil (1982). Green and Swets showed that the area under the true ROC curve is linked to the *2-alternative forced-choice (2-AFC) experiment* used in psychophysics. (By "true ROC curve," we mean the empirical curve if it is constructed from an infinitely large sample of patients and an infinite number of decision thresholds. Note that the fitted curve is an estimate of this true curve.) In a 2-AFC experiment, two stimuli are presented to an observer: one is *noise*, the other is *signal*. The observer identifies the signal stimulus; the area under the ROC curve is the frequency with which the observer correctly identifies the signal. The area under the ROC curve constructed from ordinal or continuous data retains this same meaning, even though the 2-AFC experiment is not performed (Hanley and McNeil, 1982). The area under the empirical ROC curve is actually computed from the mathematical reconstruction of random pairs of patients with and without the condition. Out of the pair, the patient with the more suspicious test result is considered the signal stimulus. (See Chapter 4.)

Bamber (1975) noted that the area under the empirical ROC curve is equivalent to the quantity obtained when one performs the Mann–Whitney version of the Wilcoxon 2-sample rank-sum statistic. This link is important because the properties of the Wilcoxon statistic are used to predict the statistical properties of the ROC curve area (Hanley and McNeil, 1982). (See Chapter 4.)

Table 2.8 describes the ROC curve areas of some common diagnostic tests. One can see that a large range in ROC curve areas exists for these tests. We cannot say which ROC curve area is a good one, because what is considered

Table 2.8 ROC Curve Areas for Some Common Diagnostic Tests

Target Disorder	Patient Population	Diagnostic Test (and Gold Standard)	ROC Curve Area
Breast cancer	Women presenting for screening	Film-screen mammography (biopsy or two year followup	Range: 0.74–0.95 Mean: 0.85 (Beam et al., 1996)
Multiple Sclerosis (MS)	Patients with signs and symptoms of MS	MRI CT (expert panel)	0.82 0.52 (Mushlin et al., 1993)
Herniated nucleus pulposus–caused nerve compression	Patients with acute low-back and radicular pain	MRI CT CT myelography (expert panel)	0.81–0.84 0.86 0.83 (Thornbury et al., 1993)
Fetal pulmonary maturity	Infants who were delivered within 72 hours of amniotic fluid testing	Lecithin/sphingomyelin ratio Saturated phosphatidylcholine (evaluation of newborn)	0.70–0.88 0.65–0.85 (Hunink et al., 1990)
Tumor staging in non–small cell bronchogenic carcinoma	Patients with known or suspected non–small cell bronchogenic carcinoma	CT/MRI (surgery or biopsy)	Chest wall invasion: 0.86/0.87 Bronchial involvement: 0.83/0.78 Mediastinal invasion: 0.83/0.92 Mediastinal node metastasis: 0.60/0.60 (Webb et al., 1991)
Obstructive airways disease	Subjects presenting to the pulmonary function test lab	Forced expiratory time (spirometry)	0.63 (Schapira et al., 1993)

good depends on the disorder and clinical application. However, the table does allow us to put the ROC curve areas of new tests in context with some commonly used and accepted diagnostic tests.

In Fig. 2.3, the area under the empirical ROC curve for mammography is 0.83; that is, if we select, at random, two patients—one with and one without breast cancer—the probability is 0.83 that the patient with breast cancer will have a more suspicious test result. The area under the binormal-fitted curve is slightly larger at 0.86. Unless the number of decision thresholds is large, the area under the empirical ROC curve is usually less than the area under the fitted curve. (See Chapter 4.)

In Fig. 2.4, the fitted ROC curve for gap is illustrated along with a possible alternative diagnostic measure, offset. Although gap describes the distance between the legs of the artificial heart valve, offset describes the deviation of the strut leg from a straight line. The areas under these fitted curves are, respectively, 0.87 and 0.65. Based on the ROC curve areas, it is gap, not offset, that has superior performance.

On rare occasions, the ROC curve area, when used as a measure of diagnostic accuracy, can be misleading. Hilden (1991) offers a hypothetical example

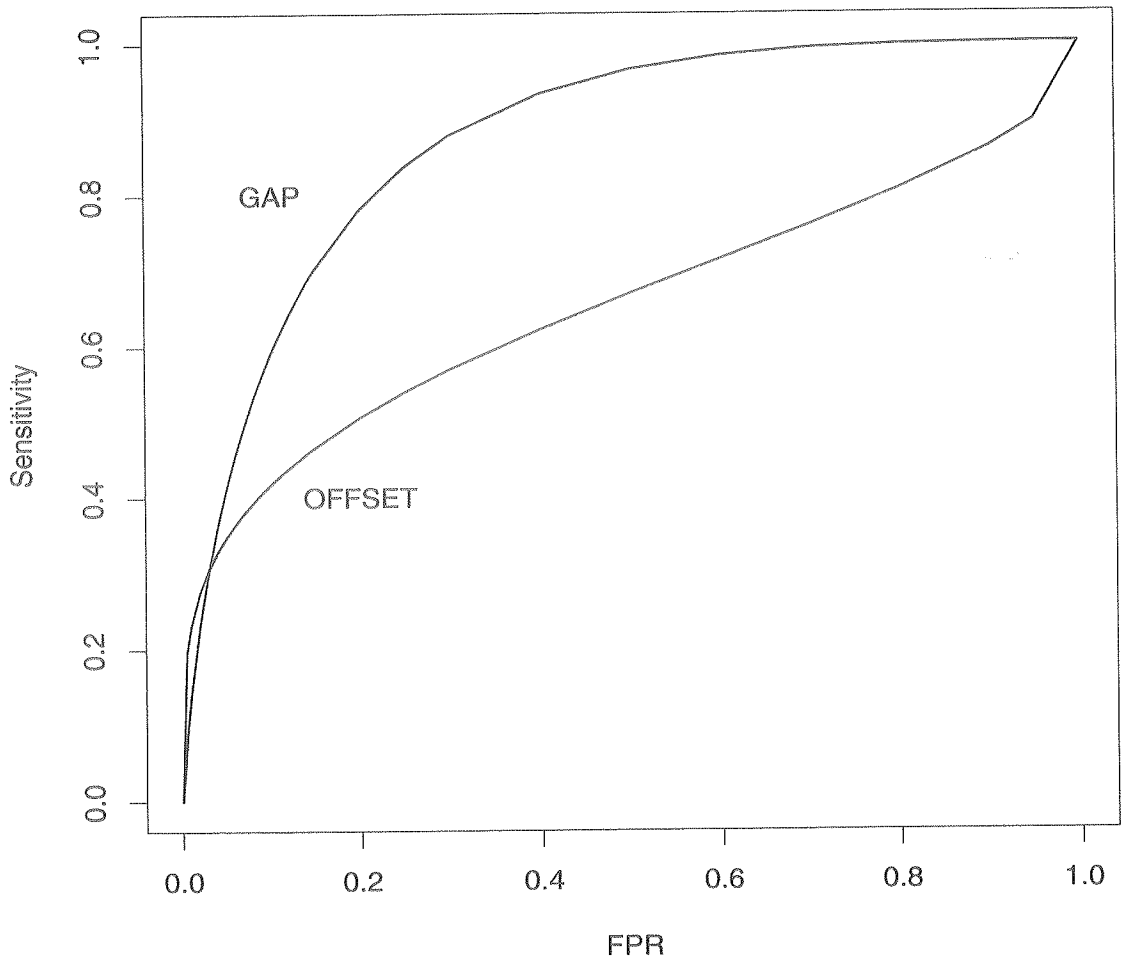

Figure 2.4 Fitted ROC curves for gap and offset.

of a perfectly discriminating test with an ROC curve area of only 0.5. Suppose that patients without the condition have test values between 80 and 120, while one half of the patients with the condition have values less than 80 and the other half have values greater than 120. The ROC curve, shown in Fig. 2.5, consists of the following line segments: (0.0, 0.0)–(0.0, 0.5); (0.0, 0.5)–(1.0, 0.5); and (1.0, 0.5)–(1.0, 1.0). The ROC curve area is 0.5, yet the test discriminates perfectly between patients with and without the condition. The transformation $T' = |T - 100|$ leads to an ROC curve with area of 1.0. We now assume that, when appropriate, a test's results have been transformed so that as the value of the test result increases, the likelihood of the condition increases. A real example of when such a transformation is necessary is a test for atherosclerosis of the carotid arteries. Ultrasound is used to measure the velocity of blood as it passes through the vessels. The velocity increases as the extent of disease increases; however, when a vessel is completely occluded, the velocity is zero. To estimate the ROC curve area of the velocity measurements, Hunink et al. (1993) assigned ranks to the velocity measurements, but instead of assigning a rank of one to the zero velocities, they assigned the highest rank.

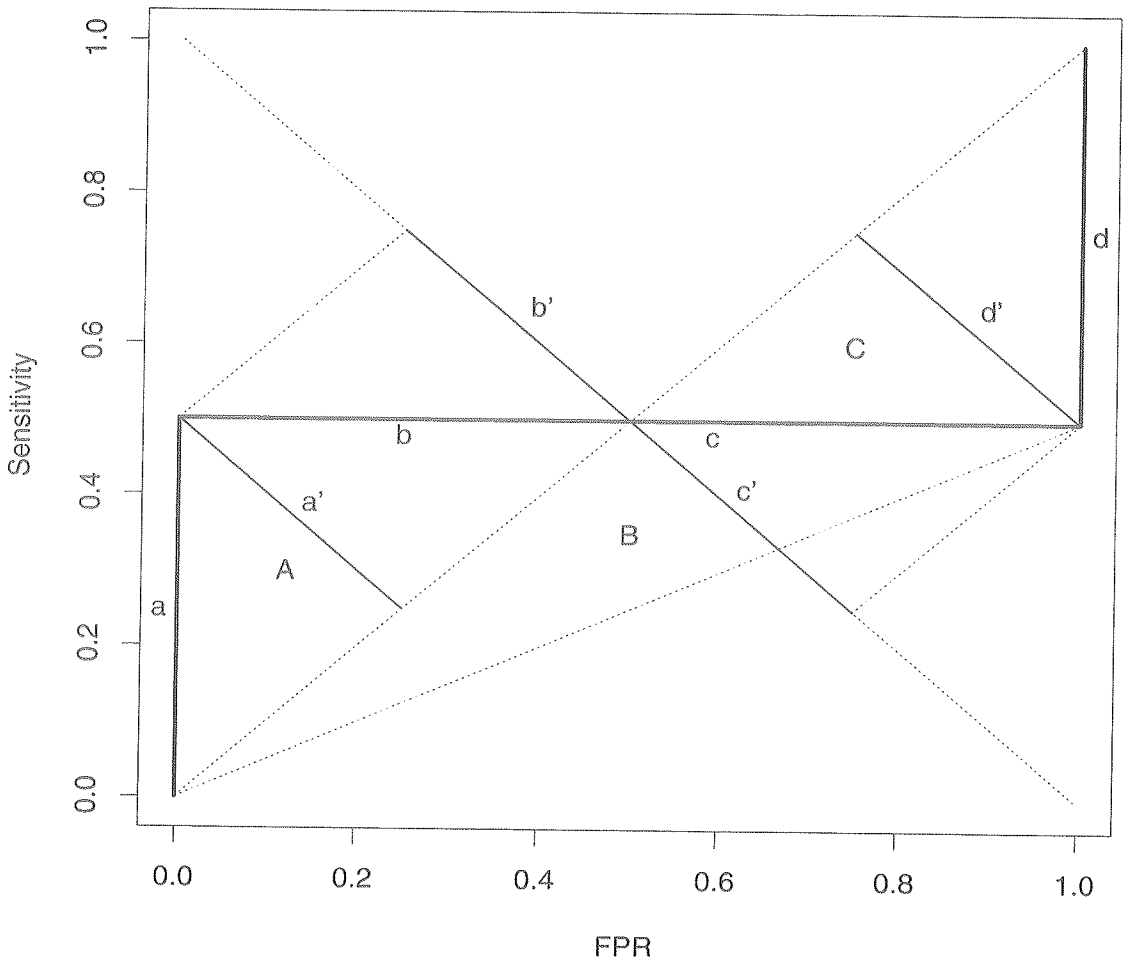

Figure 2.5 A perfectly discriminating test with an ROC area of 0.5.

The ROC curve area describes a test's inherent ability to discriminate between patients with versus without the condition, for the ROC curve area is invariant to the prevalence of the condition and the cutoffs used to form the curve. Such a measure of diagnostic accuracy is useful in the early stages of a diagnostic test's evaluation, but once a test's ability to distinguish well is shown, its role for particular applications must be evaluated. At this stage, we may be interested only in a small portion of the ROC curve. For example, if we use the heart valve–imaging technique to screen asymptomatic patients, we are interested only in the part of the ROC curve where the specificity is high; we will adjust our decision threshold to ensure that the specificity is high. We are not interested in the average sensitivity over all specificities or the average specificity over all sensitivities. As a global measure of intrinsic accuracy, the ROC curve area is not always relevant.

Similarly, the ROC curve area may be misleading when comparing the accuracy of two tests. The ROC curve areas of two tests may be equal, but the tests may differ in clinically important regions of the curve. Likewise, the ROC curve areas may differ, but the tests may have the same area in the clinically relevant region of the curve. Figure 2.6 illustrates two ROC curves that cross

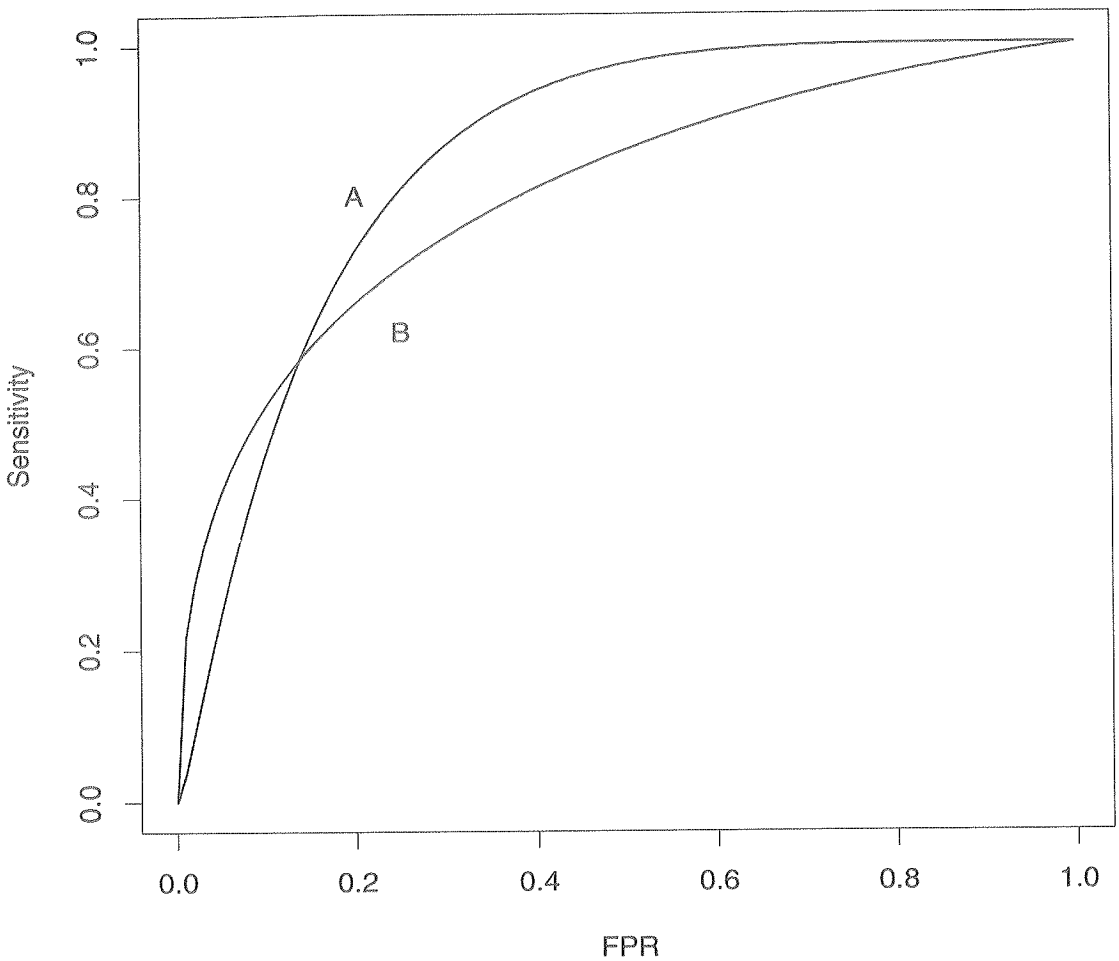

Figure 2.6 Two tests with crossing ROC curves.

at an FPR of 0.14. The area under curve A is greater than the area under curve B (i.e., 0.85 versus 0.80). If the clinically relevant region of the curve is at low FPRs, test B is preferrable to test A, despite the greater ROC curve area for A.

In Sections 2.5 and 2.6, we present two alternative summary measures of intrinsic accuracy that focus on only a portion of the ROC curve, thus overcoming the main limitation of the area under the whole curve.

2.5 THE SENSITIVITY AT A FIXED FPR

An alternative summary measure of intrinsic accuracy is the *sensitivity at a fixed FPR* or, similarly, the FPR at a fixed sensitivity, which we write as $Se_{(\mathrm{FPR}=e)}$ or $\mathrm{FPR}_{(Se=e)}$, respectively. For a predetermined FPR of e (or predetermined sensitivity of e), the sensitivity (or FPR) is estimated from the ROC curve.

This measure of accuracy allows us to focus on the particular portion of the ROC curve of clinical relevance. The characteristics of the clinical application, such as the prevalence of the condition and the consequences of misdiagnoses (see Section 2.10), determine at which FPR or sensitivity we need to operate. The ROC curves for gap and offset were illustrated in Fig. 2.4. Suppose that the clinical situation requires a low FPR. At FPR = 0.05, the sensitivity is 0.41 and 0.35 for gap and offset, respectively; at FPR = 0.20, the respective sensitivities are 0.78 and 0.51. (See Chapter 4 for a description of these MLEs.) Thus at both of these FPRs, the observed sensitivity is greater for gap.

The sensitivity at a fixed FPR is often preferable to the ROC curve area when evaluating a test for a particular application. This measure also has a simple and clinically useful interpretation. One disadvantage of this measure is that reported sensitivities from other studies are often at different FPRs; thus comparisons with published literature can be problematic. A second disadvantage is that published reports are not always clear regarding whether the FPR was selected before the start of the study (as it should be) or after the data were examined (a practice that can introduce bias) (Hanley, 1989). A third disadvantage is that the statistical reliability of this measure is lower (i.e., the variance is larger) than that of the ROC curve area (Hanley, 1989; Obuchowski and McClish, 1997). (See Chapter 6.)

2.6 THE PARTIAL AREA UNDER THE ROC CURVE

Another summary measure of intrinsic accuracy is the *partial area under the ROC curve*. As its name implies, it is the area under a portion of the ROC curve, often defined as the area between two FPRs, e_1 and e_2, for which we write $A_{(e_1 \le \mathrm{FPR} \le e_2)}$. Similarly, we can define the area between two sensitivities, for which we write $A_{(e_1 \le Se \le e_2)}$. If $e_1 = 0$ and $e_2 = 1$, the area under the entire ROC curve will be specified; if $e_1 = e_2$, the sensitivity at a fixed FPR of e (or

FPR at a fixed sensitivity of e) will be given. The partial area measure is thus a compromise between the ROC curve area and the sensitivity at a fixed FPR.

Like the sensitivity at a fixed FPR index, the partial area allows one to focus on the portion of the ROC curve relevant to a particular clinical application. In Fig. 2.4, if an FPR range is restricted to 0.0–0.05, the partial area for offset will be slightly larger than it will be for gap (though not statistically significant) at 0.0139 versus 0.0126. If we include larger FPRs, such as 0.0–0.20, the partial area for gap (0.108) will be larger than it will be for offset (0.080). (See Chapter 4 for descriptions of these MLEs.)

To interpret the partial area, we must consider its maximum possible value. The maximum area is equal to the width of the interval, that is, $(e_2 - e_1)$ (McClish, 1989). McClish (1989) and Jiang, Metz, and Nishidawa (1996) recommend standardizing the partial area by dividing by its maximum value; Jiang et al. refer to this standardized partial area as the *partial area index*. This index is interpreted as the average sensitivity for the range of specificities examined (or average specificity for the range of sensitivities examined), an interpretation that is highly useful clinically. For the heart valve–imaging example, the average sensitivities in the 0.0–0.20 FPR range are 0.54 and 0.41, respectively, for gap and offset.

Dwyer (1997) offers a probabilistic interpretation of the partial area index when the partial area is defined for sensitivities greater than e_1, that is, $A_{(e_1 \le \mathrm{TPR} \le 1.0)}$. The partial area index equals the probability that a randomly chosen patient without the condition will be distinguished correctly from a randomly chosen patient with the condition who tested negative for the criterion that corresponds to TPR = e_1. For example, suppose we want to estimate $A_{(0.8 \le \mathrm{TPR} \le 1.0)}$ from the gap values in Table 2.4. From Table 2.5, we know that a cutoff of >0.05 corresponds to an observed sensitivity of 0.80. Among the 10 patients with a fractured valve, 2 tested negative using this criterion (i.e., 2 patients had gap values of ≤0.05). The partial area index is the probability that a randomly chosen patient with an intact valve will be correctly distinguished from a patient like one of the foregoing. An analogous interpretation for $A_{(0.0 \le \mathrm{FPR} \le e_2)}$ is the probability that a randomly chosen patient with the condition will be correctly distinguished from a randomly chosen patient without the condition who tested positive for the criterion that corresponds to FPR = e_2. Note the similarities between this interpretation and the probabilistic interpretation of the ROC curve area.

A potential problem with the partial area measure is that the minimum possible value depends on the location along the ROC curve. The minimum partial area is equal to $(1/2)(e_2 - e_1)(e_2 + e_1)$ (McClish, 1989). For example, the minimum value for $A_{(0 \le \mathrm{FPR} \le 0.2)}$ is 0.02 (maximum value is 0.20) and the minimum value for $A_{(0.8 \le \mathrm{FPR} \le 1.0)}$ is 0.18 (maximum value is 0.20). Suppose that we estimated a partial area of 0.19 for both of these FPR ranges; the partial-area index, 0.95, is the same for both ranges. However, we would probably not value these two areas the same. To remedy this problem, McClish (1989) offers a transformation of the partial area to values between 0.5 and 1.0. The

formula is as follows:

$$\frac{1}{2}\left[1 + \frac{A_{e_1 \le \text{FPR} \le e_2)} - \min}{\max - \min}\right] \tag{2.1}$$

where min and max are the minimum and maximum possible values for the partial area. Continuing with this example, the partial area of 0.19 is transformed to 0.972 for the 0.0–0.2 FPR range and 0.75 for the 0.8–1.0 FPR range. For the heart valve–imaging example, the transformed partial area for gap and offset in the 0.0–0.20 FPR range is 0.744 and 0.672, respectively.

The partial area measure has similar limitations to the sensitivity at a fixed FPR. First, it is difficult to compare this measure with the published literature if different ranges are used. Second, the relevant range should be specified a priori; it is not always clear from published reports whether this specification had occurred. Third, the statistical reliability of this measure is lower than that of the ROC area but greater than that of the sensitivity at a fixed FPR (Hanley, 1989; Obuchowski and McClish, 1997). (See Chapter 6.)

2.7 LIKELIHOOD RATIOS

Still another single index of diagnostic accuracy is the *likelihood ratio* (LR), the ratio of two probabilities: the probability of a particular test result among patients with the condition to the probability of that test result among patients without the condition. The LR can be defined for a single test result value, for an interval of test values, and for the results of one side of a decision threshold. In symbols,

$$\text{LR}(t) = \frac{P(T = t|D = 1)}{P(T = t|D = 0)} \tag{2.2}$$

where t can be a single test value, an interval of test values, or one side of a decision threshold. When the test result refers to one side of a decision threshold, we have *positive and negative LRs*, where

$$\text{LR}(+) = \frac{P(T = 1|D = 1)}{P(T = 1|D = 0)}$$

and

$$\text{LR}(-) = \frac{P(T = 0|D = 1)}{P(T = 0|D = 0)}$$

Note that the LR(+) is the ratio of sensitivity to the FPR; likewise, LR(−) is the ratio of the FNR to specificity. For example, from Table 2.6, let us choose

a decision threshold at probably benign so that patients classified as probably benign, suspicious, or malignant will be called positive. The LR(+) = Se/FPR = 0.967/0.633 = 1.53; the LR(−) = FNR/Sp = 0.033/0.367 = 0.09.

The LR reflects the magnitude of evidence that a particular test result provides in favor of the presence of the condition relative to the absence of the condition. An LR of 1.0 indicates that the test result is equally likely in patients with and without the condition; an LR > 1.0 indicates that the test result is more likely among patients with the condition than without the condition; and an LR < 1.0 indicates that the test result is more likely among patients without the condition. The higher the LR, the likelier the test result among patients with the condition relative to patients without the condition. With the mammography example, a positive test result is 53% more likely in patients with breast cancer than patients without breast cancer; a negative test result is 11 (i.e., 1/0.09) times more likely in patients without breast cancer.

Table 2.9 summarizes the probability of various gap values for patients with and without valve fractures. The last column gives the LR(t). Gap values between 0.031 and 0.050 are equally likely in patients with and without a fractured valve (i.e., the probability of 0.10), whereas gap values between 0.051 and 0.070 are twice as likely in patients with a fractured valve.

Using LRs is a convenient means of describing the degree of abnormality. Radack, Rouan, and Hedges (1986) describe a 773-patient prospective study of the usefulness of creatine kinase concentration in the diagnosis of acute myocardial infarction (AMI). They calculated five LRs corresponding to five serum creatine kinase value ranges: LR = 9.26, for a value > 480 IU/L; 7.31, for 361–480; 4.15, for 241–360; 0.42, for 121–240; and 0.69, for 1–120. A test result of 150 is four-tenths as likely in a patient with AMI, whereas a value greater than 480 is nine times more likely in a patient with AMI than a patient without AMI. Differences in LR magnitude provide important clinical information not available when a single decision threshold is chosen.

Table 2.9 Estimating LT(t) From the Heart Valve–Imaging Study

Test Result, t	$P(T = t \mid D = 0)$	$P(T = t \mid D = 1)$	LR(t)
0.0	0.3	0.0	0.0
0.001 – 0.030	0.3	0.1	0.33
0.031 – 0.050	0.1	0.1	1.0
0.051 – 0.070	0.1	0.2	2.0
0.071 – 0.100	0.0	0.1	Undefined
0.101 – 0.130	0.2	0.0	0.0
>0.0	0.7	1.0	1.43
>0.03	0.4	0.9	2.25
>0.05	0.3	0.8	2.67
>0.07	0.2	0.6	3.0
>0.15	0.0	0.3	Undefined

Another example comes from a study by Mushlin et al. (1993) of the accuracy of MRIs in identifying multiple sclerosis (MS). Two observers assigned one of the following rating categories to each of 303 patients: "definitely not MS," "probably not MS," "possible MS," "probable MS," and "definite MS." The corresponding LRs were 0.3, 0.3, 1.3, 2.9, and 24.9. Although the accuracy of the MRI was less than definitive (ROC curve area = 0.82), the authors concluded that a "definite MS" reading essentially established the diagnosis of MS. However, 25% of patients with MS were classified as "probably not MS" or "definitely not MS"; thus these diagnoses were not sufficient to rule out MS.

Zweig and Campbell (1993) note that LRs can be easily misinterpreted. Consider the mammography data in Tables 2.3 and 2.7. The LR(+), 1.53, is the same in both tables; thus it *is* correct to say that a positive result is 1.53 times more likely in patients with cancer as compared with patients without cancer. It is *not* necessarily correct to say that given a positive test result, a patient is 1.53 times more likely to have cancer than to not have cancer. The latter statement is a reflection of the prevalence in the population. For example, in Table 2.3 (showing a prevalence of breast cancer of 50%), given a positive test result, the ratio of patients with cancer to without cancer *is* 1.53 (i.e., 29 : 19 equals 1.53 : 1), but in Table 2.7 (showing a prevalence of 1%), the ratio is 0.015 (i.e., 29 : 1881), indicating that it is much more likely that a patient with a positive test result does *not* have cancer.

The LR is linked to the empirical ROC curve. The numerator of the LR(+) is the y coordinate of the curve; the denominator, the x coordinate of the curve. The LR for an interval of test values, $t_1 - t_2$, corresponds to the slope of the line segment between t_1 and t_2 on the ROC curve (Choi, 1998). The ROC curve labeled *A* in Fig. 2.7 corresponds to the gap measurement of the heart valve–imaging study. (See Table 2.9.) The line connecting the (FPR, *Se*) coordinate for the decision threshold at 0.0 and the (FPR, *Se*) coordinate for the decision threshold at 0.03 has a slope of 0.33, which corresponds to the LR(0.001 – 0.030) from Table 2.9. One can verify this equivalence by computing the change in sensitivity divided by the change in FPR for these two points—that is, from the bottom of Table 2.9: $(1.0 - 0.9)/(0.7 - 0.4) = 0.33$. Similarly, the slope of the line between the (FPR, *Se*) coordinates corresponding to decision thresholds at 0.03 and 0.05 is 1.0, which is the LR(0.031 – 0.050). The ROC curve labeled *B* has the single point ($Se = 0.9$, FPR = 0.4) from the decision threshold > 0.03. The slope of the line from the origin to this point is 2.25—that is, the LR(+) for the >0.03 cutoff. For ROC curve *B*, the slope is the ratio of *Se*/FPR, or LR(+). Generally, though, the slope is the change in sensitivity divided by the change in FPR over the defined interval of test results as in ROC curve *A* (Zweig and Campbell, 1993).

In Section 2.3, we described improper ROC curves. The distinction between *proper* and *improper* ROC curves is based on the LR. Figures 2.8 and 2.9 illustrate a proper and improper ROC curve, respectively. The insets in both figures depict the corresponding distributions of test results for hypothetical

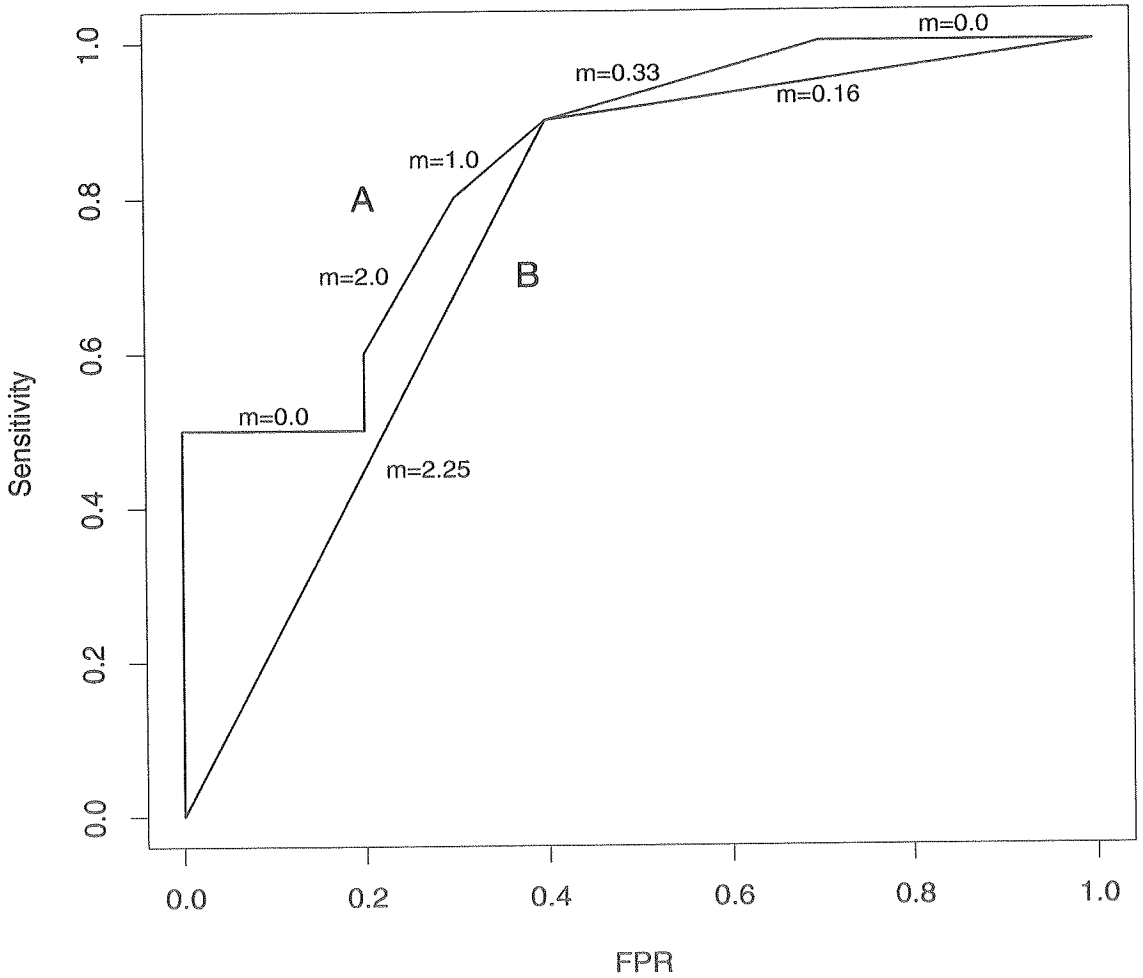

Figure 2.7 ROC curves and their LRs.

patients: the light shading denotes the distribution of test results for patients without the condition, whereas the dark shading denotes the test results of patients with the condition. In Fig. 2.8, the distributions of the test results of patients with and without the condition are identical but shifted apart. The corresponding ROC curve is a decreasing function of the LR. At the bottom left corner of the curve, corresponding to large test values, the LR is >1.0. The LR decreases along the curve's path. At $T = 17$, the LR = 1.0. At the top right corner of the curve, corresponding to small test values, the LR is <1.0. Proper ROC curves such as the one depicted in Fig. 2.8 are monotonic functions of the LR (Pan and Metz, 1997).

In contrast, Fig. 2.9 shows more variability in the test results of patients without the condition. At the far bottom left corner of the ROC curve and at the far top right corner, the LR is <1.0. The probability that $T = 16$ is the same for patients with and without the condition; thus the LR = 1.0. Similarly, at $T = 21$, the LR = 1.0, and when T is between 16 and 21, the LR is >1.0. This ROC curve is an improper one because it is not a monotonic function of the LR. The curve has the characteristic "hook" (Pan and Metz, 1997) at the bottom

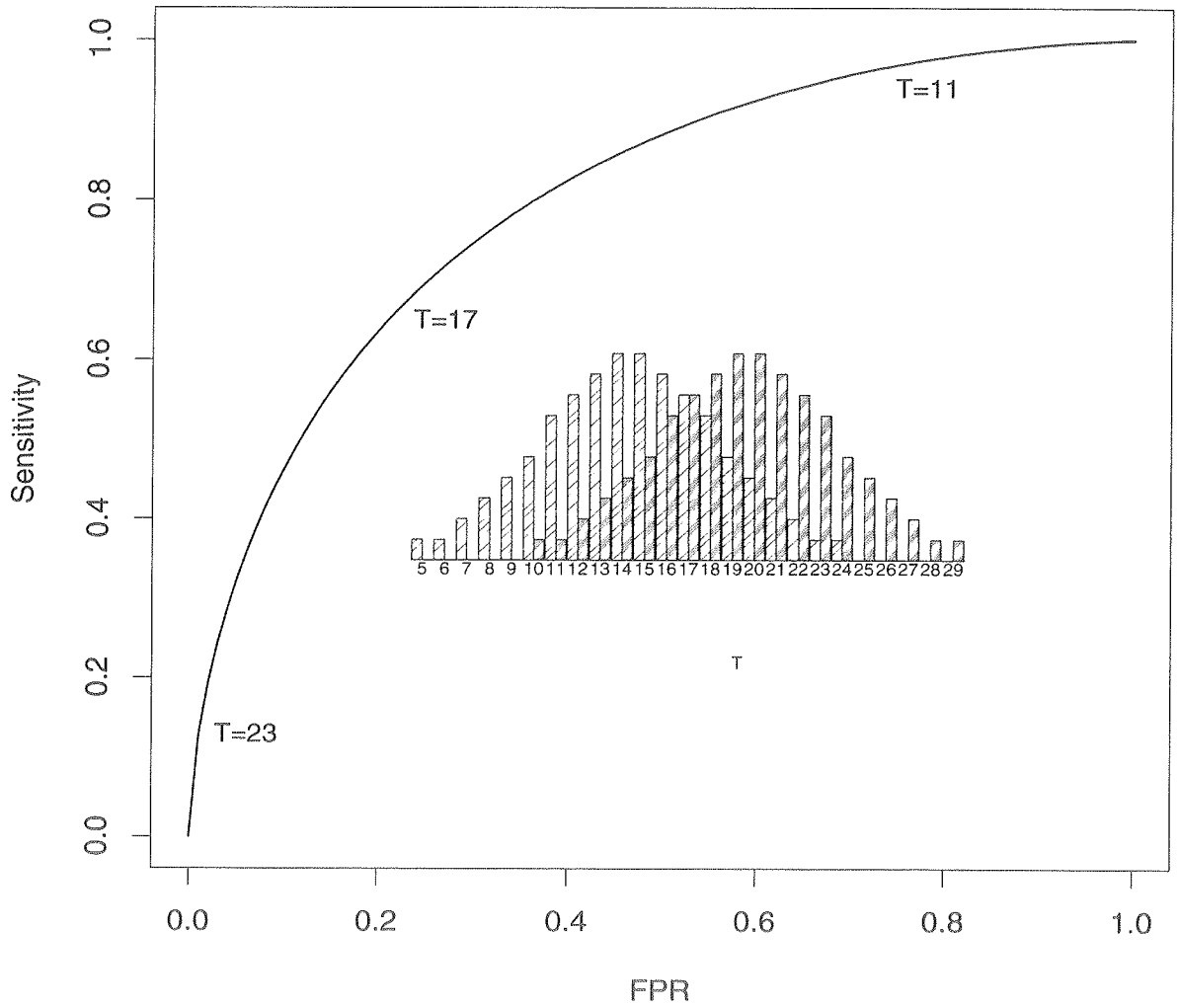

Figure 2.8 A proper ROC curve.

left. Problems in estimating improper ROC curves can occur; see Chapter 4 for a discussion.

The LR is an intrinsic measure of diagnostic accuracy because it is unaffected by the prevalence of disease. For example, the LR(+) and LR(−) from the mammography data in Tables 2.3 and 2.7 are identical at 1.53 and 0.09, despite the differing prevalence of breast cancer. The LR, however, has some limitations when it is used as a single measure of accuracy. Like all ratios of two random variables, it is difficult to estimate its standard error (SE) and statistical distribution. (See Chapters 4.) Zweig and Campbell (1993) illustrate that an LR without an accompanying ROC curve can be misleading. They present two ROC curves with identical LRs for the line segments forming the curves but with vastly different ROC curve areas. The two identical curves are parallel, but one is located near the upper left corner, the other near the chance diagonal. The primary role of the LR lies in using Bayes' theorem (see Section 2.10) and in defining the optimal decision threshold for particular clinical applications (see Section 2.11).

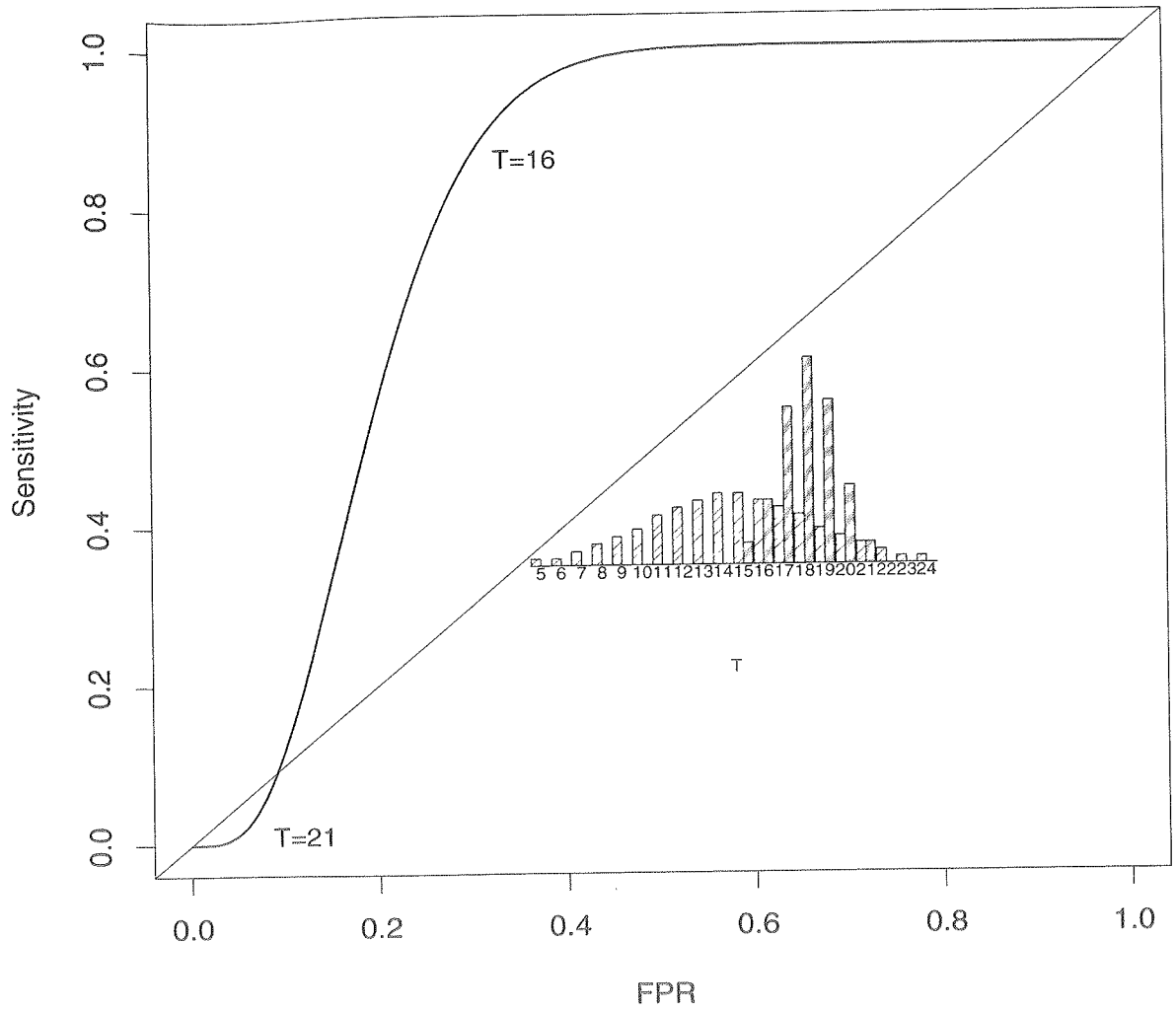

Figure 2.9 An improper ROC curve.

2.8 OTHER ROC CURVE INDICES

Several ROC curve summary indices are used in psychophysics, which we describe here briefly for their historical significance. The measures d', d'_e, and $z(A)$ are derived from ROC curves fit to the binormal model. They are best understood when the ROC curve is plotted on normal deviate scales (Fig. 2.10), because normal-fitted ROC curves are straight lines on such scales. By "normal deviate," we mean the value from a standard normal distribution that corresponds to a certain probability. For example, 95% of observations from a standard normal distribution are less than the normal deviate value of 1.645. Thus in Fig. 2.10, instead of indicating on the y axis a sensitivity of 0.95, we indicate a normal deviate value of 1.645.

The first index, d', equals the normal deviate value corresponding to the sensitivity minus the normal deviate value corresponding to the FPR (Green and Swets, 1966). Index d' is applicable only when $b = 1.0$; it can be measured at any point along the ROC curve. In Fig. 2.10, the ROC curve labeled 1

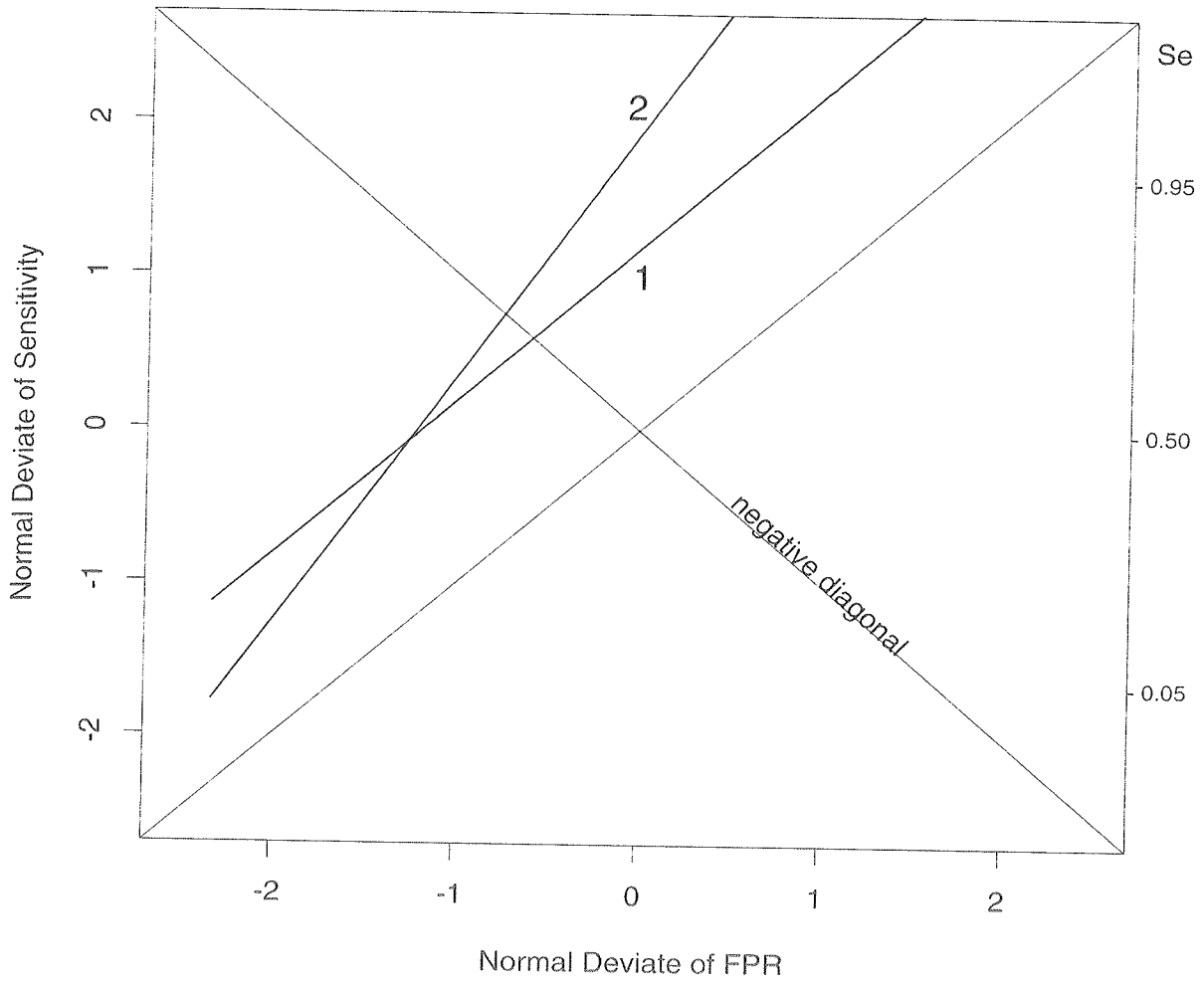

Figure 2.10 Normal fitted ROC curves plotted on normal deviate scales.

has a unit slope. At the normal deviate of the FPR of 0.0, written $z(\text{FPR}) = 0.0$, the $z(\text{TPR}) = 1.19$; thus d' equals $1.19 - 0.0 = 1.19$. The value of d' for the chance diagonal is 0.0; d' equals about 4.0 at near-perfect discrimination (Swets, 1979). The index d'_e can be used when the binormal ROC curve parameter b is not equal to 1.0 (Egan and Clarke, 1966). It is calculated in the same way as d' but measured where the ROC curve crosses the negative diagonal, that is, the diagonal line from (0, 1) to (1, 0). In Fig. 2.10, the ROC curve labeled 2 has a nonunit slope. Where the negative diagonal crosses ROC curve 2, $z(\text{FPR}) = -0.75$ and $z(\text{TPR}) = 0.75$; thus $d'_e = 1.5$. A second measure for when b does not equal one is $z(A)$ (Schulman and Mitchell, 1966; Simpson and Fritter, 1973; Swets, 1979). Index $z(A)$ is equal to the perpendicular distance between the ROC curve and the point $z(\text{FPR}) = z(\text{TPR}) = 0.0$. For ROC curves 1 and 2, $z(A) = 0.84$ and 1.04, respectively. When $b = 1.0$, the quantity $\sqrt{2}z(A)$ is equal to d' and d'_e. These three measures are seldom used in diagnostic medicine because (1) they rely on a normal distribution fit to the test results and (2) they do not have clinically useful interpretations.

Two new measures have been proposed as alternatives to the ROC curve

area: the *projected length of the ROC curve (PLC)* and the *area swept out by the ROC curve (ASC)* (Lee and Hsiao, 1996). They are defined geometrically but have probabilistic interpretations. The PLC is the sum of all of the projected lengths of the segments making up the ROC curve onto the negative diagonal. To illustrate these two measures, we use the ROC curve in Fig. 2.5; it comprises the following four line segments: a, b, c, and d. The PLC $= a' + b' + c' + d' = 4(\sqrt{2}/4) = 1.414$. The ASC is the sum of the areas swept out by a ray emanating from the origin (0, 0) to each point on the ROC curve. In Fig. 2.5, the ASC $= A + 2B + C = 0.5$. The corresponding probabilistic interpretations are given in the example that follows.

Suppose a patient with equal chances of having and not having the condition presents for diagnosis. After testing, we compare the probabilities of having versus not having the condition; the diagnosis is assigned according to the higher probability. Lee and Hsiao (1996) refer to this scenario as *strategy A*. The probability of correctly diagnosing a patient with the condition, in addition to the probability of correctly diagnosing a patient without the condition, by using strategy A is a linear function of the PLC. For example, in Fig. 2.5 this probability is 1.0. The ASC index is related to a different testing scenario, referred to as *strategy B*. Suppose that two patients—one with and one without the condition—present for diagnosis. We first rank their test results without determining the actual values of those test results; then we ascertain the actual test result of at least one of the patients. If we ascertain the test result of the patient with the lower rank, we denote this the *low in-value*. We compare the probabilities of having versus not having the condition; the diagnosis is assigned according to the higher probability. The probability of correctly diagnosing a pair of patients when the low–in-value patient has the condition, in addition to the probability of correctly diagnosing a pair of patients when the low–in-value patient does not have the condition, is a linear function of the ASC. For example, in Fig. 2.5 this probability is 1.0.

The main advantage of the PLC and ASC indices over the ROC curve area is that they do not require any transformation of the test results in the rare situation, such as that in Fig. 2.5, in which the ROC curve area is 0.5 and yet the test discriminates perfectly between patients with and without the condition. The disadvantages of the PLC and ASC are that

1. like the ROC curve area, they are global measures of accuracy and thus are not useful for particular clinical applications;
2. their interpretations are no more meaningful (perhaps even less meaningful) clinically than the probabilistic interpretation of the ROC area;
3. their estimation, SEs, and statistical properties have not been well-studied;
4. they are difficult to estimate for tests with results on a continuous scale.

2.9 THE LOCALIZATION AND DETECTION OF MULTIPLE ABNORMALITIES

Some diagnostic tasks are more complicated than simple detection of a single occurrence of the condition. For example, mammography patients can have multiple lesions that must be correctly located prior to follow-up procedures, such as biopsies, and surgery. Another example is the detection of infarcts in a patient suspected of having a stroke. Multiple infarcts can occur, making detecting and locating them in the correct brain hemisphere especially critical. Several modifications to the ROC curve have been proposed to describe accuracy when it involves the localization and/or detection of multiple abnormalities. In this section, we briefly review these ideas.

Starr et al. (1975) proposed the idea of *location-ROC curves (LROCs)*, where a TP requires both detection and correct localization of the condition. Starr et al. also developed equations to predict the performance of a reader who must detect and correctly locate a condition; the equations are based on the reader's conventional (i.e., detection-only) ROC curve. They assume that the decision variables in each subregion of an image are independent. This assumption is highly restrictive; perhaps it is the reason why LROC curves have not been used widely.

In 1976, Metz, Starr, and Lusted proposed a modification to the ROC curve for describing accuracy when there are potentially multiple occurrences of the condition (no localization). Unfortunately, this modification also assumes that subregions are independent of one another.

Egan, Greenberg, and Schulman (1961) and, later, Bunch et al. (1978) proposed the idea of *free-response ROC curves (FROC)*, which handle the task of detecting and locating multiple occurrences of the condition. The y axis of the FROC curve is the probability of both detecting and correctly locating the condition; the x axis is the average number of FPs per case. The summary index of the FROC curve is interpreted as the average fraction of occurrences detected on each image before the reader makes one FP error. Chakraborty (1989) and, later, Chakraborty and Winter (1990) developed methods to estimate the FROC curve. However, they, too, assume independence between multiple positive findings on the same image.

Obuchowski et al. (2001) proposed an alternative to FROC curves that does not make the independence assumption. They proposed that the image be divided a priori into multiple mutually exclusive regions, each of which the reader must diagnose separately. The authors proposed the use of statistical methods that consider the correlation in test results between regions of the same patient.

2.10 INTERPRETATION OF DIAGNOSTIC TESTS

In this section, we address one of the most important questions to clinicians: What does this test result mean? For a patient with a positive test result, we

want to know the probability of the patient having the condition; for a patient with a negative test result, we want to know the probability of the patient not having the condition. In symbols, these probabilities are $P(D = 1|T = 1)$ and $P(D = 0|T = 0)$, respectively. Determining these probabilities is tricky because they depend on not only the intrinsic accuracy of the test but also the probability of the condition before the test is performed.

Consider as an example a 65-year-old woman who has undergone a screening mammogram, the result of which is positive. What is the probability that this patient has breast cancer? Suppose that Table 2.7 describes the results of a prospective study of 3000 65-year-old women who have undergone screening mammography. We can compute the probability of breast cancer after a positive mammogram directly from these data. The number of patients who test positive is 1910, and of these patients, only 29 actually have breast cancer; thus $P(D = 1|T = 1) = 29/1910$, or 0.015. The probability of the condition given a positive test result, is the *positive predictive value* or *PPV*. The probability that the patient does not have breast cancer following a positive mammogram, $P(D = 0|T = 1)$, is $1881/1910 = 0.985$, or simply 1 − PPV.

Suppose that this patient has a negative mammogram. The probability that the patient does not have breast cancer following a negative test result, $P(D = 0|T = 0)$, is the *negative-predictive value (NPV)*. Here NPV = 1089/1090, or 0.999. The probability of breast cancer after a negative test result is 1 − NPV, or 0.001.

Recall that the sensitivity and specificity calculated from Tables 2.3 and 2.7 were identical: 0.967 and 0.367. However, the PPV and NPV calculated from these two tables are not identical; from Table 2.3, the PPV = 0.604 and the NPV = 0.917, as compared with 0.015 and 0.999 from Table 2.7. The discrepancy is due to the different prevalence rates. The PPV and NPV are *not* measures of the intrinsic accuracy of a test; they are functions of both the intrinsic accuracy and the prevalence of the condition. Both the study design and sampling scheme affect the prevalence rate in a study sample. (See Chapter 3.) These factors must be considered when estimating the PPV and NPV.

Continuing with this example, suppose that the 65-year-old woman with the positive mammogram differs from the patients in Table 2.7 because she has a family history of breast cancer. The probability of breast cancer in women with a family history of the disease is higher than in the general population. Because PPV and NPV are functions of the prevalence of the condition, we cannot compute them directly from Table 2.7. However, we can still use the intrinsic accuracy estimates from Table 2.7 (or Table 2.3) to compute the PPV and NPV using *Bayes' theorem.*

Bayes' theorem, named after the Reverend and mathematician who developed it (Bayes, 1763), is a method of determining both the PPV and NPV, given both the intrinsic accuracy of a test and the probability of the condition before the test is applied. The latter probability is the *pre-test probability* and is based on the patient's history, signs and symptoms, and results of any diagnostic tests performed previously. The PPV and NPV are the *post-test probabilities*

of the condition (also called *revised* or *posterior probabilities*), because they represent the probability of the condition after the test result is known. Bayes' theorem, then, gives us the post-test probability of the condition as a function of the pre-test probability of the condition and the sensitivity and specificity of the test. Bayes' theorem is expressed as

$$P(D = d|T = t) = \frac{P(T = t|D = d)P(D = d)}{P(T = t|D = 0)P(D = 0) + P(T = t)|D = 1)P(D = 1)} \quad (2.3)$$

For example, to compute the PPV and NPV,

$$\text{PPV} = P(D = 1|T = 1) = \frac{Se \times P(D = 1)}{Se \times P(D = 1) + (1 - Sp) \times P(D = 0)} \quad (2.4)$$

and

$$\text{NPV} = P(D = 0|T = 0) = \frac{Sp \times P(D = 0}{Sp \times P(D = 0) + (1 - Se) \times P(D = 1)} \quad (2.5)$$

Bayes' theorem can be proven using the statistical definition of conditional probability. Let A and B denote two events. The conditional probability $P(A|B)$ is equal to $P(A \text{ and } B)/P(B)$. The numerator on the right side of Eq. (2.3) is equal to $P(A \text{ and } B)$ and the denominator is equal to $P(B)$; thus the theorem is proven.

Figure 2.11 illustrates the relationship between the pre- and post-test probabilities after a positive test result. Here, the sensitivity is constant at 0.95, and the FPR is 0.01, 0.10, or 0.25. When the pre-test probability is very low, a positive test greatly increases the probability of the condition. In contrast, when the pre-test probability is very high, a positive test has little effect on the probability of the condition. A positive test has its greatest impact when the FPR is low. In contrast, the sensitivity has a large impact when a test result is negative—the greater the sensitivity, the larger the impact.

It is important to note that one cannot properly assess the results of a diagnostic test without knowing the probability of the condition before the test is performed (Sox, Jr. et al., 1989). A good description is given by Diamond and Forrester (1979), who applied Bayes' theorem to compute the probability of coronary artery disease occurring after stress electrocardiography. They present a table of post-test probabilities according to the test result (depression of the S-T segment in millimeters) and to each of three pre-test conditions (patient age, gender, and symptoms). For the same depression of the S-T segment, the post-test probability varies from 0.938 for a 60- to 69-year-old male with typical angina to 0.003 for a 30- to 39-year-old woman with no symptoms.

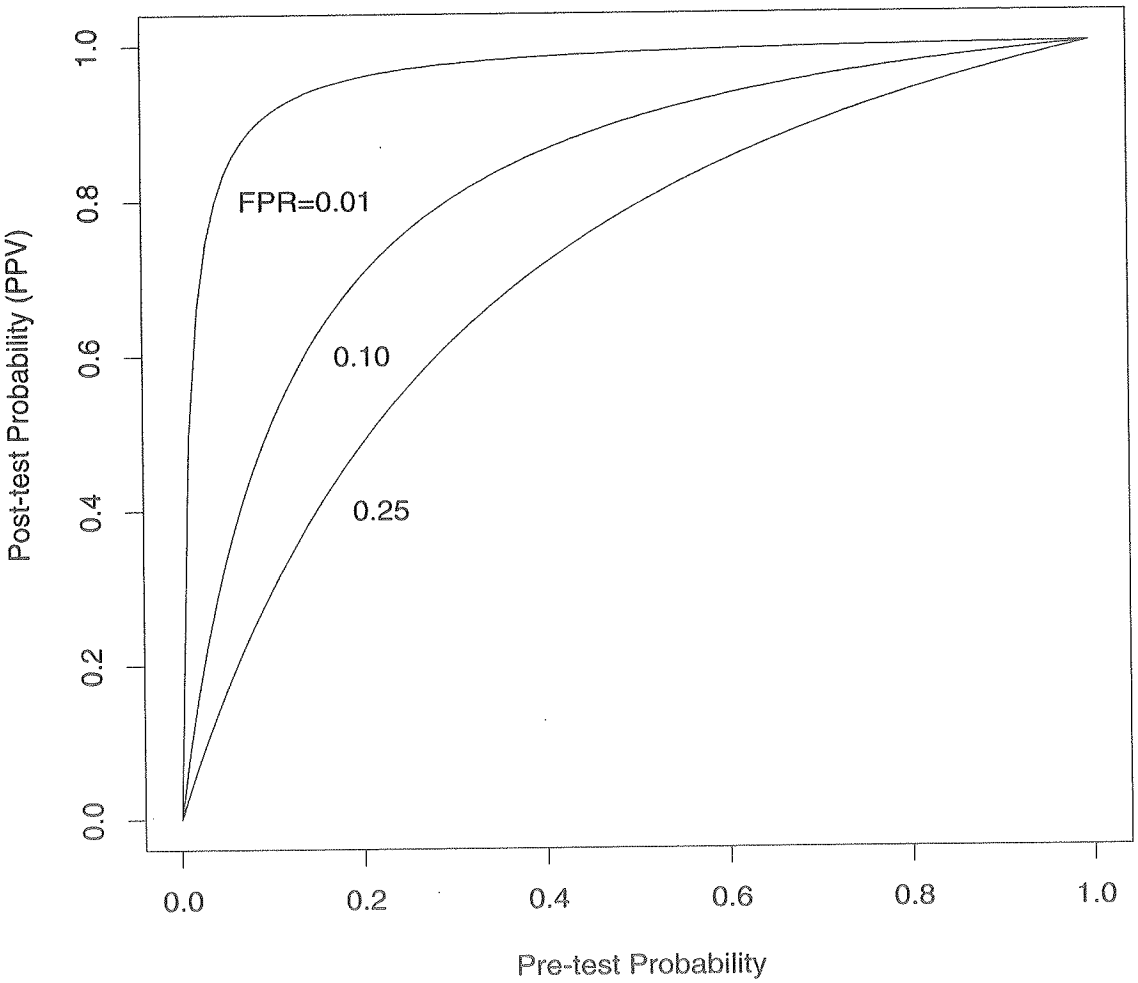

Figure 2.11 The relationship between pre-test probability and the PPV.

An alternative form of Bayes' theorem uses odds ratios and LRs (Sox, Jr. et al., 1989):

$$\text{post-test odds} = \text{pre-test odds} \times \text{LR} \tag{2.6}$$

The odds are formed by dividing a probability by its complement: odds = $P/(1-P)$. For example, if we divide Eq. (2.4) by $P(D=0|T=1)$, then

$$\frac{\text{PPV}}{(1-\text{PPV})} = \frac{P(D=1)}{P(D=0)} \times \frac{Se}{(1-Sp)}$$

Similarly, $\text{NPV}/(1-\text{NPV}) = P(D=0)/P(D=1) \times Sp/(1-Se)$. The probability is then calculated from the odds by $P = \text{odds}/(\text{odds}+1)$.

Suppose that the pre-test probability of breast cancer is 10% for a 65-year-old woman with a family history of breast cancer. The pre-test odds are 0.10/0.90 to 1, or 0.1111 to 1. From Table 2.7 (or Table 2.3), the LR(+) is 1.53, in which case the post-test odds are $0.1111 \times 1.53 = 0.17$, which is equivalent to a probability (i.e., a PPV) of 0.145.

The version of Bayes' theorem given in Eq. (2.6) is convenient because the value of the LR conveys the impact of the test result. An LR of 1.0 does not alter the odds, a large LR increases the odds, and a small LR decreases the odds. For example, if a patient has a creatine kinase concentration of 250—that is, LR(241 – 360) = 4.15—and a pre-test odds of an AMI of 0.25, the post-test odds will be 1.04 (i.e., the probability of an AMI revised from 20% to 51%). If another patient has a creatine kinase concentration of 450—that is, LR(361 – 480) = 7.31—with the same pre-test odds, the post-test odds will be 1.83 (i.e., the probability of an AMI revised from 20% to 65%). If a single LR from a single cutoff value, such as LR(>241), had been reported instead of LRs from several ranges of creatine kinase concentrations, the different post-test probabilities of these patients would not have been appreciated (Radack et al., 1986).

Several assumptions are made when applying Bayes' theorem (Sox, Jr. et al., 1989). One is that sensitivity and specificity are constant, regardless of the pre-test probability. This assumption is violated if, for example, a test is less sensitive in detecting a condition in its early stages, when its pre-test probability is low. An example is a chest x-ray for detecting lung cancer. In the condition's early stages, when the lesion is small, the sensitivity of a chest x-ray is low. If the patient has no early symptoms, the pre-test probability also will be low. Later, when the lesion is larger and easier to detect, the test's sensitivity increases and, at the same time, symptoms develop; thus the pre-test probability increases. The post-test probability calculated from Bayes' theorem would be misleading if it were based on data from a study of patients having both early and late stages of lung cancer.

A second assumption is important when calculating the probability of a condition after a sequence of tests (Sox, Jr. et al., 1980). To use Bayes' theorem, the sensitivity and specificity of a test must be independent of the results of other tests—meaning that if two tests are to be performed in sequence, the sensitivity of the second test must be equivalent for patients who test positive and for patients who test negative on the first test. We can write this assumption as

$$P(T_2 = 1 | T_1 = 1, D = 1) = P(T_2 = 1 | T_1 = 0, D = 1)$$

where T_1 and T_2 denote the results of the first and second test, respectively. The foregoing assumption also applies to specificity, as follows:

$$P(T_2 = 0 | T_1 = 1, D = 0) = P(T_2 = 0 | T_1 = 0, D = 0)$$

If this assumption is met, then the post-test probability of the first test in the sequence is the pre-test probability of the second test (and so forth). If Eq. (2.6) is used, the LRs of multiple tests can be multiplied. (See Exercise 2.7 at the end of this chapter.)

The use of Bayes' theorem to interpret diagnostic tests has an interesting

analogy in the interpretation of statistical tests in clinical research (Browner and Newman, 1987). As with diagnostic testing, errors occur in statistical hypothesis testing. The results of statistical tests, as with diagnostic tests, cannot be interpreted properly without knowledge of the prior probability of the research hypothesis. Although difficult to quantify, the prior probability of the research hypothesis can be used in Bayes' theorem to calculate the probability that the research hypothesis is true.

2.11 OPTIMAL DECISION THRESHOLD ON THE ROC CURVE

In this section, we present a simplistic approach to determining the *optimal decision threshold* on the ROC curve for a particular application. Loosely defined, the optimal decision threshold for a particular application is the point on the ROC curve where, on average, the financial and/or health effects (i.e., the "costs") are minimized. Our approach to determining the optimal decision threshold is simplistic, because some complicated issues are treated casually and the costs needed for the determination are assumed known when, actually, they are difficult to estimate (Metz, 1978).

We begin with two basic assumptions needed in the derivation of the optimal threshold (Dwyer, 1997). First, we assume that two options for managing the patient exist: give treatment when the condition is present or withhold treatment when the condition is not present. Second, we assume that the decision to give or withhold treatment is based on the results of the test; positive results imply that treatment should be given, negative results imply that treatment should be withheld.

The optimal decision threshold for a particular application depends on the costs of performing the test and the cost of the consequences of the test's results (the "downstream costs"). These costs—financial and/or health—can be viewed from the perspective of the patient and his or her care providers, insurers, and dependents, as well as the perspective of society (Zweig and Campbell, 1993). The costs of performing the test are denoted by C_0. Here, C_0 may include the technical and professional costs of performing the test, as well as any health costs caused by test complications. The costs of each diagnostic decision's consequences are denoted by C_{TP}, C_{FP}, C_{TN}, and C_{FN}, where, for example, C_{TP} denotes the cost of a true-positive result. We weigh each of these costs by the probability of its occurrence. The average overall cost of performing a test, C, is

$$C = C_0 + P(\text{TP}) \times C_{TP} + P(\text{FP}) \times C_{FP} + P(\text{TN}) \times C_{TN} + P(\text{FN}) \times C_{FN} \quad (2.7)$$

where $P(\text{TP})$ denotes the probability of a true-positive result and is equal to $Se \times P(D = 1)$. Thus the cost of performing a test depends on the sensitivity and specificity of the test, the pre-test probability of the condition, and the consequences of the test decisions.

The location on the ROC curve where the average overall cost is at a minimum for a particular application is the optimal operating point on the curve (Metz, 1978). The slope m, of the ROC curve at the optimal operating point is given by the following equation. [See Metz (1978) for proof.]

$$m = \frac{P(D=0)}{P(D=1)} \times \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}} \tag{2.8}$$

If the ROC curve is smooth, the optimal operating point is where a line with this slope is tangent to the curve. When the empirical ROC curve is used, the optimal operating point is where a line—with the slope calculated from Eq. (2.8)—moves down from above and to the left to intersect the ROC curve plot (Zweig and Campbell, 1993). Another way to find the optimal operating point is to find the sensitivity and specificity pair that maximizes the function [sensitivity $- m(1 -$ specificity)], where m is from Eq. (2.8) (Zweig and Campbell, 1993).

Note that the best operating point on the ROC curve does not depend on C_0. Instead, it depends on the consequences of the test's results only in terms of the difference in costs between FPs and TNs relative to the difference in costs between FNs and TPs.

The slope of the ROC curve is steep in the lower left, where both the TP and FP rates are low, and it is flat near the upper right, where the TP and FP rates are high. The best operating point is near the lower left if the condition is rare and/or if treatment for the condition is harmful to healthy patients and of little benefit to patients with the condition. In these situations, we want to minimize the number of FPs, so the best operating point is in the lower left (Metz, 1978). In contrast, when the condition is common and/or when treatment is highly beneficial and poses little harm to healthy patients, the best operating point is toward the upper right. In these situations, we want to minimize FNs.

Somoza and Mossman (1991) use Eq. (2.8) to determine the optimal operating point for a biological marker used to detect depression. The biological marker is rapid eye movement (REM) latency, the time between sleep onset and the start of the first rapid eye movement period. REM latency is shorter in patients with depression. Somoza and Mossman fit ROC curves to the data of four studies of REM latency in patients suspicious for depression. They use patient "utility" values to describe the relative costs of the test's decision, with values ranging from 0.0 (the lowest health value) to 1.0 (the highest health value). Somoza and Mossman also assigned a utility value of 1.0 to patients in whom depression was correctly diagnosed and for whom treatment could be offered (TPs); 0.9 to patients in whom depression was correctly ruled out but for whom no treatment could be offered (TNs); 0.7 to patients for whom an incorrect diagnosis of depression was made and, consequently, an unnecessary treatment regimen was given (with needless exposure to treatment side effects) (FPs); and 0.0 to depressed patients in whom depression went undetected and for whom an effective treatment was not given (FNs). If the prevalence of

depression in the presenting population is 0.10, then the slope of the ROC curve at the optimal operating point will be 1.8. The optimal decision threshold is between 47 and 60 minutes, depending on which ROC curves of the four studies are used. Patients with a REM latency of less than this decision threshold are diagnosed with depression and are treated; otherwise, the patient is considered negative for depression and is not treated.

The financial and health costs used in determining the optimal decision threshold must be calculated with great care. Estimation of these costs is a specialized field in medicine. A few relevant references are Pauker and Kassirer (1975), (1980); Weinstein et al. (1980), (1996); Gold et al. (1996); and Russell et al. (1996).

2.12 MULTIPLE TESTS

Few diagnostic tests are both highly sensitive and specific. To diagnose patients, clinicians often order two or more tests, which can be performed *in parallel* (i.e., at the same time and interpreted in combination) or *serially* (i.e., the results of the first test determine whether the second test is performed). The advantage of serial testing is its cost-effectiveness, because some patients receive only one test. The potential disadvantage is the delay in treatment while one awaits the results of the second test (Hershey, Cebul, and Williams, 1986). We talk briefly about these two scenarios, beginning with parallel testing.

Griner et al. (1981) gave hypothetical data for two tests, A and B, for diagnosing pancreatic cancer. We assume that the sensitivity and specificity of the tests are independent of the results of the other tests. (See Section 2.10.) Individually, test A has a sensitivity and specificity of 0.8 and 0.6, respectively; test B, 0.9 and 0.9, respectively. There are two ways in which the tests can be interpreted in parallel:

1. The OR rule, in which the diagnosis is positive if either A or B is positive. Both A and B must be negative for the diagnosis to be negative.
2. The AND rule, in which the diagnosis is positive only if both A and B are positive. Either A or B can be negative for the diagnosis to be negative.

Using the OR rule, the sensitivity of the combined result is $Se_A + Se_B - (Se_A \times Se_B) = 0.8 + 0.9 - (0.8 \times 0.9) = 0.98$. The specificity is $(Sp_A \times Sp_B) = 0.54$. With the OR rule, the sensitivity of the combined result is higher than either test individually, but the specificity is lower than either test individually. Using the AND rule, the combined sensitivity is $(Se_A \times Se_B) = 0.72$, whereas the specificity is $Sp_A + Sp_B - (Sp_A \times Sp_B) = 0.96$. Thus with the AND rule, specificity is higher than either test individually, but the sensitivity is lower than either test individually.

An example of parallel testing is given by Beam, Sullivan, and Layde (1996)

in a study of the effect of double-reading mammograms. Here, two readers interpreted each mammogram, and their results were combined using the OR rule. The result was generally an increase in sensitivity offset by an increase in the FPR.

An alternative to parallel testing is serial testing. The common decision rules in serial testing are as follows:

1. For the OR rule, if the first test is positive, the diagnosis will be positive; otherwise, perform the second test. If the second test is positive, the diagnosis will be positive; otherwise, the diagnosis will be negative.
2. For the AND rule, if the first test is positive, apply the second test. If the second test is also positive, the diagnosis will be positive; otherwise, the diagnosis will be negative.

Again using the hypothetical data from Griner et al. (1981), suppose that test A is the first test applied. Using the AND rule, the sensitivity is $Se_A \times Se_B = 0.72$ and the specificity is $Sp_A + (1 - Sp_A) \times Sp_B = 0.96$; the accuracy is the same as the AND rule for parallel testing. Using the OR rule, the sensitivity is $Se_A + (1 - Se_A) \times Se_B = 0.98$ and the specificity is $Sp_A \times Sp_B = 0.54$; the accuracy is the same as it is in the OR rule for parallel testing.

Serial testing is particularly cost-efficient when screening patients for a rare condition. Exercise 2.7 at the end of this chapter describes a 3-tier serial-test approach, using the AND rule, to screen for preclinical Parkinson's disease.

EXERCISES

2.1 A study was conducted to assess the accuracy of "Cine" MRI for the detection of thoracic aortic dissection (VanDyke et al., 1993). There were 45 patients with a dissection and 69 patients without a dissection studied. The reader used the following confidence scale: 1 = "definitely not dissection," 2 = "probably not dissection," 3 = "possible dissection," 4 = "probable dissection," and 5 = "definite dissection." The test results are summarized in the table that follows. Compute the *Se* and FPR for each possible decision threshold; then plot the ROC curve.

Dissection Status	1	2	3	4	5
Present	7	7	3	5	23
Absent	39	19	9	1	1

2.2 Design an experiment to mimic a diagnostic test that lacks the ability to discriminate between patients with versus without the condition. You may use coins, dice, or other suitable objects. Construct an ROC curve from the results of your experiment. Describe the curve.

2.3 There are five outcomes of a test: $T = a$, b, c, d, or e. Their relative frequencies for patients with and without the condition are given in the table that follows. Compute the LR associated with each potential outcome. Construct a set of decision rules that provides a proper ROC curve.

t	$P(T = t \mid D = 1)$	$P(T = t \mid D = 0)$
a	0.1	0.4
b	0.1	0.1
c	0.2	0.1
d	0.2	0.3
e	0.4	0.1

2.4 The investigators of the study described in Exercise 2.1 hypothesized that the new Cine imaging sequence would improve accuracy over the standard "spin-echo" imaging sequence. The reader's confidence scores for the same 114 patients using spin-echo imaging are given in the table that follows. Plot the ROC curve on the same axes as those in Exercise 2.1. Discuss the relative strengths and weaknesses of the ROC curve area and the partial area for comparing these two curves.

Dissection Status	1	2	3	4	5
Present	1	4	10	4	26
Absent	21	39	9	0	0

2.5 Magnetic resonance angiography (MRA) is a noninvasive test used to detect cerebral aneurysms; its *Se* and *Sp* are each ≥0.80. Asymptomatic patients with a family history of aneurysms have a 20% pre-test probability of aneurysms. For these patients, assuming that the *Se* and *Sp* of MRA are equal to 0.80, what is the probability of an aneurysm after a positive test?

The probability of an aneurysm in the general population (i.e., subjects without a family history) is 0.02. What must the *Se* and *Sp* of MRA be to achieve the same post-test probability for a patient without a family history as for a patient with a family history? (Assume that $Se = Sp$.)

2.6 Under what scenarios is the PPV equal to zero? Equal to one? Under what scenarios is the NPV equal to zero? Equal to one?

2.7 Parkinson's disease—here, abbreviated as PD—is a debilitating disorder affecting the neurologic system by depleting the brain of dopamine neurons. Currently, there are no known risk factors, and the disease is often so difficult to diagnose that a substantial dopamine-neuron loss can occur before any treatment is begun. The goal of this study is to identify patients with preclinical PD so that their treatment can begin earlier in the course

of the disease. Because the prevalence of preclinical PD is low (approximately 1%), no single test has adequate sensitivity and specificity. Thus a 3-tier diagnostic test strategy is proposed (E. Montgomery, MD, Cleveland Clinic Foundation, OH; personal communication, 1998). The first test is a simple questionnaire in which patients are asked about problems with daily living; it has a sensitivity of 95% but a specificity of only 20%. If a patient tests positive on the questionnaire, then he or she will undergo the second test, which consists of olfactory, motor, and mood assessments; its sensitivity is 72% and its specificity is 86%. The third test is a nuclear-imaging single-photon emission computed tomography (SPECT) study in which dopamine neurons are examined; its sensitivity and specificity are both 80%. If a patient tests positive on both of the first two tests, he or she will undergo the SPECT imaging. Thus the patient's result is positive if all three tests are positive (i.e., AND serial testing); otherwise, the result is negative. What is the probability of PD after a positive SPECT study—that is, $P(D = 1|$positive results on all three tests)? What is the probability of no PD after a negative SPECT study—that is, $P(D = 0|$ positive results on first two tests and negative result on third test)? What is the probability of no PD after a negative second test—that is, $P(D = 0|$ test + on first test and minus on second test)? What is the sensitivity and FPR of this 3-tier approach?

REFERENCES

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating graph, *J. Math. Psychol.* **12**: 387–415.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances, *Philos. Trans. Royal Soc. Lond.* **53**: 370–418.

Beam, C. A., Sullivan, D. C., and Layde, P. M. (1996). Effect of human variability on independent double reading in screening mammography, *Acad. Radiol.* **3**: 891–897.

Browner, W. S. and Newman, T. B. (1987). Are all significant p values created equal? The analogy between diagnostic tests and clinical research, *JAMA* **257**: 2459–2463.

Bunch, P. C., Hamilton, J. F., Sanderson, G. K., and Simmons, A. H. (1978). A free-response approach to the measurement and characterization of radiographic-observer performance, *J. Appl. Photogr. Eng.* **4**: 166–171.

Campbell, G. (1994). General methodology I: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests, *Stat. Med.* **13**: 499–508.

Chakraborty, D. P. (1989). Maximum-likelihood analysis of free-response receiver operating characteristic (FROC) data, *Med. Phys.* **16**: 561–568.

Chakraborty, D. P. and Winter, L. H. L. (1990). Free-response methodology: Alternative analysis and a new observer-performance experiment, *Radiology* **174**: 873–881.

Chol, B. C. K. (1998). Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test, *Am. J. Epidemiol.* **148:** 1127–1132.

Diamond, G. A. and Forrester, J. S. (1979). Analysis of probability as an aid in the clinical diagnosis of coronary artery disease, *N. Engl. J. Med.* **300:** 1350–1358.

Dwyer, A. J. (1997). In pursuit of a piece of the ROC, *Radiology* **202:** 621–625.

Egan, J. P. and Clarke, F. R. (1966). *Experimental Methods and Instrumentation in Psychology*, McGraw-Hill, New York.

Egan, J. P., Greenberg, G. Z., and Schulman, A. I. (1961). Operating characteristics, signal detectability, and the method of free-response, *J. Acoust. Soc. Am.* **33:** 993–1007.

Gilbert, G. K. (1885). Finley's tornado predictions, *Am. Meteorol. J.* **1:** 167.

Gold, M. R., Siegel, J. E., Russell, L. B., and Weinstein, M. C. (1996). *Cost-effectiveness in Health and Medicine*, Oxford University Press, New York.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*, John Wiley and Sons, New York.

Griner, P. R., Mayewski, R. J., Mushlin, A. I., and Greenland, P. (1981). Selection and interpretation of diagnostic tests and procedures, *Ann. Intern. Med.* **94:** 553–592.

Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: The state of the art, *Crit. Rev. Diagn. Imaging.* **29:** 307–335.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143:** 29–36.

Hershey, J. C., Cebul, R. D., and Williams, S. V. (1986). Clinical guidelines for using two dichotomous tests, *Med. Decis. Making* **6:** 68–78.

Hilden, J. (1991). The area under the ROC curve and its competitors, *Med. Decis. Making* **11:** 95–101.

Hunink, M. G., Polak, J. B., Barlan, M. M., and O'Leary, D. H. (1993). Detection and quantification of carotid artery stenosis: Efficacy of various Doppler velocity parameters. *AJR Am. J. Roentgenol.* **160:** 619–625.

Hunink, M. G., Richardson, D. K., Doubilet, P. M., and Begg, C. B. (1990). Testing for fetal pulmonary maturity: ROC analysis involving covariate, verification bias, and combination testing, *Med. Decis. Making* **10:** 201–211.

Jiang, Y., Metz, C. E., and Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests, *Radiology* **201:** 745–750.

Lee, W. C. and Hsiao, C. K. (1996). Alternative summary indices for the receiver operating characteristic curve, *Epidemiology* **7:** 605–611.

Lusted, L. B. (1971). Signal detectability and medical decision-making, *Science* **171:** 1217–1219.

McClish, D. K. (1989). Analyzing a portion of the ROC curve, *Med. Decis. Making* **9:** 190–195.

Metz, C. E. (1978). Basic principles of ROC analysis, *Semin. Nucl. Med.* **8:** 283–298.

Metz, C. E. (1986). ROC methodology in radiologic imaging, *Invest. Radiol.* **21:** 720–733.

Metz, C. E. (1989). Some practical issues of experimental design and data analysis in radiologic ROC studies, *Invest. Radiol.* **24:** 234–245.

Metz, C. E. and Kronman, H. B. (1980). Statistical significance tests for binormal ROC curves, *J. Math. Psychol.* **22:** 218–243.

Metz, C. E., Starr, S. J., and Lusted, L. B. (1976). Observer performance in detecting multiple radiographic signals, *Radiology* **121:** 337–347.

Mushlin, A. I., Detsky, A. S., Phelps, C. E., O'Connor, P. W., Kido, D. K., Kucharczyk, W., Giang, D. W., Mooney, C., Tansey, C. M., and Hall, W. J. (1993). The accuracy of magnetic resonance imaging in patients with suspected multiple sclerosis, *JAMA* **269:** 3146–3151.

Obuchowski, N. A. and McClish, D. K. (1997). Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices, *Stat. Med.* **16:** 1529–1542.

Obuchowski, N. A., Lieber, M. L., and Powell, K. A. (2001). Statistical analysis for detecting and locating multiple abnormalities with application to mammography, *Acad. Radiol.* **7:** 516–525.

Pan, X. and Metz, C. E. (1997). The "proper" binormal model: Parametric receiver operating characteristic curve estimation with degenerate data, *Acad. Radiol.* **4:** 380–389.

Pauker, S. G. and Kassirer, J. P. (1975). Therapeutic decision making: A cost-benefit analysis, *N. Engl. J. Med.* **293:** 229–234.

Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making, *N. Engl. J. Med.* **302:** 1109–1117.

Powell, K., Obuchowski, N., Chilcote, W. A., Barry, M. W., Ganobcik, S. N., and Cardenosa, G. (1999). Clinical evaluation of digital versus film-screen mammograms: Diagnostic accuracy and patient management, *AJR Am. J. Roentgenol.* (submitted).

Powell, K., Obuchowski, N., Mueller, K., Hwang, C., Ganobcik, S., Strum, B., LaPresto, E., Hirsch, J., Selzer, R., Nissen, J., and Cornhill, J. F. (1996). Quantitative detection and classification of single-leg fractures in the outlet struts of Bjork–Shiley convex-concave heart valves, *Circulation* **94:** 3251–3256.

Radack, K. L., Rouan, G., and Hedges, J. (1986). The likelihood ratio: An improved measure for reporting and evaluating diagnostic test results, *Arch. Pathol. Lab. Med.* **110:** 689–693.

Remer, E. M., Obuchowski, N., Ellis, J. D., Rice, T. W., Adelstein, D. J., and Baker, M. E. (2000). Adrenal mass evaluation in patients with lung carcinoma: A cost-effectiveness analysis, *AJR Am. J. Roentgenol.* **174:** 1033–1039.

Russell, L. B., Gold, M. R., Siegel, J. E., Daniels, N., and Weinstein, M. C. (1996). The role of cost-effectiveness analysis in health and medicine, *JAMA* **276:** 1172–1177.

Schapira, R. M., Schapira, M. M., Funahashi, A., McAuliffe, T. L., and Varkey, B. (1993). The value of the forced expiratory ti me in the physical diagnosis of obstructive airways disease, *JAMA* **270:** 731–736.

Schulman, A. I. and Mitchell, R. R. (1966). Operating characteristics from yes–no and forced-choice procedures, *J. Acoust. Soc. Am.* **40:** 473.

Simpson, A. J. and Fritter, M. J. (1973). What is the best index of detectability? *Psychol. Bull.* **80:** 481.

Somoza, E. and Mossman, D. (1991). Biological markers and psychiatric diagnosis: Risk-benefit balancing using ROC analysis, *Biol. Psychiatry* **29:** 811–826.

Sox, Jr., H., Stern, S., Owens, D., and Abrams, H. L. (1989). *Assessment of diagnostic technology in health care. Rationale, methods, problems, and directions*, National Academy Press, Washington, DC.

Starr, S. J., Metz, C. E., Lusted, L. B., and Goodenough, D. J. (1975). Visual detection and localization of radiographic images, *Radiology* **116:** 533–538.

Swets, J. A. (1979). ROC analysis applied to the evaluation of medical imaging techniques, *Invest. Radiol.* **14:** 109–121.

Thornbury, J. R., Fryback, D. G., Turski, P. A., Javid, M. J., McDonald, J. V., Beinlich, B. R., Gentry, L. R., Sackett, J. F., Dasbach, E. J., and Martin, P. A. (1993). Disk-caused nerve compression in patients with acute low-back pain: Diagnosis with MR, CT myelography, and plain CT, *Radiology* **186:** 731–738.

Turner, D. A. (1978). An intuitive approach to receiver operating characteristic curve analysis, *J. Nucl. Med.* **19:** 213–220.

VanDyke, C. W., White, R. D., Obuchowski, N. A., Geisinger, M. A., Lorig, R. J., and Meziane, M. A. (1993). Cine MRI in the diagnosis of thoracic aortic dissection, *Annual meeting of the Radiological Society of North America* (presented).

Webb, W. R. Gatsonis, C., Zerhouni, E. A., Hellan, R. T., Glazer, G. M., Francis, I. R., and McNeilm, B. J. (1991). CT and MRI imaging in staging non-small cell bronchogenic carcinoma: Report of the Radiologic Diagnostic Oncology Group, *Radiology* **178:** 705–713.

Weinstein, M. C., Fineberg, H. V., Elstein, A. S., Frazier, H. S., Neuhauser, D., Neutra, R. R., and McNeil, B. J. (1980). *Clinical decision analysis*, WB Saunders, Philadelphia.

Weinstein, M. C., Siegel, J. E., Gold, M. R., Kamlet, M. S., and Russell, L. B. (1996). Recommendations of the panel on cost-effectiveness in health and medicine, *JAMA* **276:** 1253–1258.

Zhou, X. H. (1995). Testing an underlying assumption on a ROC curve based on rating data, *Med. Decis. Making* **15:** 276–282.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine, *Clin. Chem.* **39:** 561–577.