Review Article

Receiver Operating Characteristic Analysis: A Primer[1]

John Eng, MD

Receiver operating characteristic (ROC) analysis is commonly used in clinical radiology research to express the diagnostic accuracy of imaging examinations. Many excellent resources are available that cover the technical and statistical aspects of ROC analysis (1–4). In this article, I take a nonstatistical approach in explaining the definition, interpretation, and construction of ROC curves, in hopes of making them accessible to general readers of the radiology literature and beginning clinical researchers. With this information, the reader should be able to identify data that are amenable to ROC analysis and have an intuitive understanding of the process by which an ROC curve is constructed from such data. Although ROC analysis is far from being a new technique for assessing diagnostic medical tests (5), advances continue to be made. Some of these key advances will be cited, and perhaps these citations could form a basis for further reading.

DEFINING ROC ANALYSIS

In its conventional form, ROC analysis applies to a particular, perhaps simplified, diagnostic situation. In this situation, the diagnostician's task is to correctly assign one of exactly two classifications to a diagnostic case after observing a particular stimulus associated with the case. In radiology, the two classifications are usually two disease states, such as the presence or absence of a particular disease or pathophysiological process. The two states could be simply labeled "normal" versus "abnormal," or "positive" versus "negative." In radiology, the stimulus is the imaging exam. The observer's task is to determine the correct disease classification based on information obtained from the imaging exam.

In ROC analysis, the observer's classification of each case is compared to its true classification according to an appropriate reference standard. This comparison should be familiar to all radiologists because it is the same comparison with which the familiar performance measures of sensitivity and specificity are calculated. Sensitivity is simply the proportion of correctly classified cases among all of those that are truly positive, and specificity is the proportion of correctly classified cases among all of those that are truly negative.

In the real world of medical diagnosis, performing the binary classification of each case is associated with uncertainty, so that sensitivity and specificity are not both 100%. In the interpretation of diagnostic tests, there is usually a trade-off between sensitivity and specificity (Fig. 1a). This trade-off depends on the observer's threshold for calling an exam positive. An observer with a low threshold (tendency to "over-call") will have a high sensitivity but relatively low specificity. In contrast, an observer with a high threshold (tendency to "under-call") will have a low sensitivity but relatively high specificity. The latter observer will miss more positive cases than the former, but fewer negative cases will be mistakenly called positive.

If a plot of sensitivity versus specificity (Fig. 1a) is flipped horizontally, the result is an ROC curve (Fig. 1b). The resulting flipped horizontal axis is the false positive fraction, which is equal to the specificity subtracted from 1. Thus the ROC curve is simply a plot of the intuitive trade-off between sensitivity and specificity, with the horizontal axis flipped for historical reasons. The original aim of ROC analysis was to focus on positive test results, both true positive and false positive.

Acad Radiol 2005; 12:909–916

[1] From the Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, 600 North Wolfe Street, Baltimore, Maryland 21287. Received December 20, 2004; Revised and accepted April 1, 2005. **Address correspondence to** J.E. e-mail: jeng@jhmi.edu

doi:10.1016/j.acra.2005.04.005

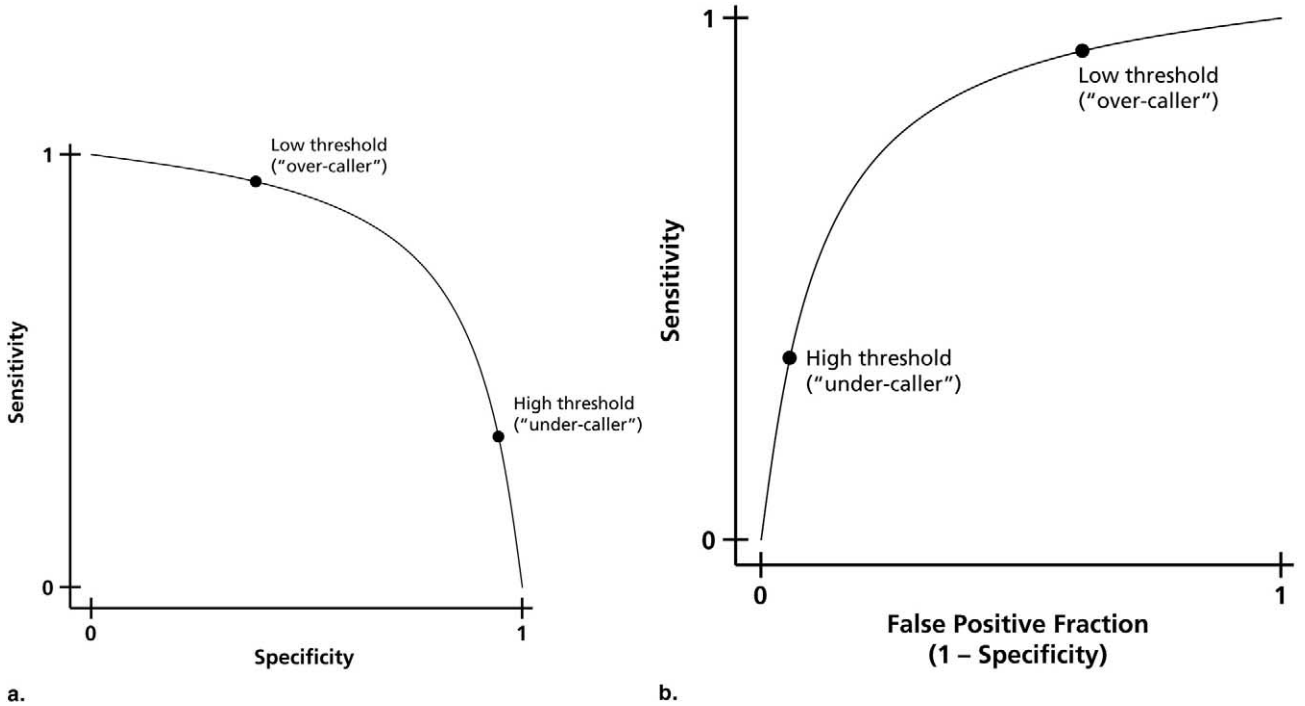

Figure 1. **(a)** Plot of a hypothetical relationship between the sensitivity and specificity of an imaging exam. There is typically a tradeoff between sensitivity and specificity. **(b)** Plot of a hypothetical receiver operating characteristic curve. The receiver operating characteristic curve is merely a simple variation of the sensitivity versus specificity plot.

The area under the ROC curve (often abbreviated AUC) is commonly used as a global indicator of diagnostic performance. It can be shown that the AUC is equal to the probability that the observer will correctly identify the positive case when presented with a randomly chosen pair of cases in which one case is positive and one case is negative (6). The AUC can also be interpreted as the average sensitivity over the entire range of possible specificities, or the average specificity over the entire range of possible sensitivities (1). If the combination of observer and test were perfectly accurate, with 100% sensitivity and 100% specificity, then the ROC curve would consist of two straight line segments encompassing the entire unit square, so the AUC would be 1 (Fig. 2). This curve would be interpreted as the observer having a 100% probability of correctly classifying a random positive–negative case pair. If the observer was completely inexperienced and/or the test was completely indiscriminate, equivalent to blind guessing, then the ROC curve would be a straight line connecting the lower left to upper right corners, and the area under this curve would be 0.5 (Fig. 2). This line corresponds to a 50% probability of the observer correctly classifying a random positive–negative case pair, which is the same as flipping a coin and letting random chance decide the correct classification.

In practice, the AUC can be thought of as representing the "average accuracy" of a diagnostic test. If so, then why bother with this area? Instead, why not simply use the traditional measures of sensitivity, specificity, and accuracy (proportion correct)? The ROC curve and the area under it possess an important property that the other measures do not. That property is independence from the threshold the observer chooses when interpreting the diagnostic exam. The ROC curve in effect adjusts for the variation in sensitivity and specificity (due to the trade-off discussed earlier) that occurs when varying interpretation thresholds exist within the same reader or among a group of readers. Not only does the AUC represent an overall accuracy measure, it also represents an accuracy measure covering all possible interpretation thresholds.

CONSTRUCTING AN ROC CURVE

It is instructive to consider how ROC curves are constructed. A typical ROC study begins by asking an observer to interpret a number of cases, some of which are positive

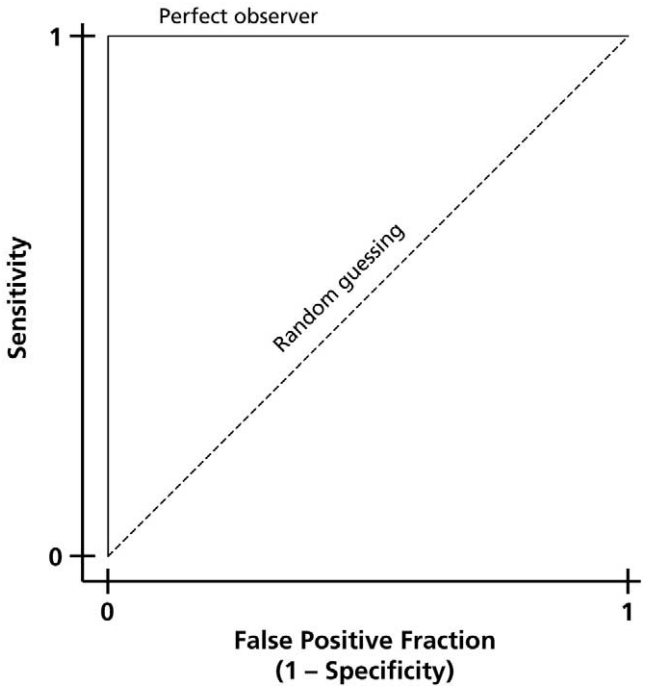

Figure 2. Plot comparing the receiver operating characteristic curves in the presence of perfect accuracy (solid line with area under curve of 1) versus random guessing (dotted line with area under curve of 0.5).

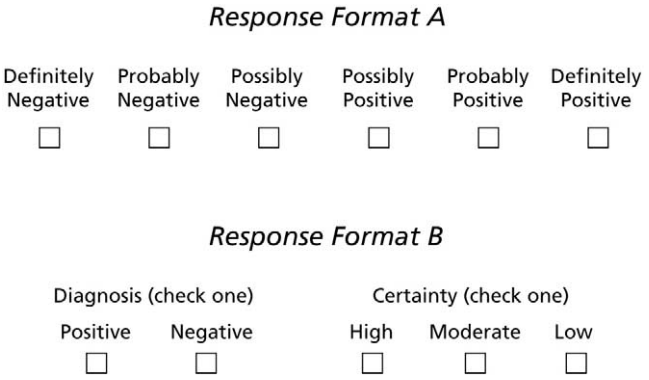

Figure 3. Two formats (A and B) for collecting data suitable for receiver operating characteristic analysis.

and some negative (true classification unknown to observer). Instead of simply asking the observer to classify each case as positive or negative, the observer is asked to rate each case according to how strongly the observer believes the case is positive (or negative). This rating is most commonly done according to an ordinal scale (Fig. 3), such as 1 = definitely negative, 2 = probably negative, 3 = possibly negative, 4 = possibly positive, 5 = probably positive, and 6 = definitely positive. Alternatively, the observer identifies each case as positive or negative and assigns a level of certainty for the positive/negative diagnosis (Fig. 3). The latter method is a bit easier for the observer to understand and can be converted to the 6-point scale in the first step of data analysis. The observer's responses are usually recorded on a data form employing a rating scale similar to the ones shown in Fig. 3.

After collecting the observer's responses, the ratings are tallied according to whether the case was truly positive or negative. Table 1 shows an example of results from an observer interpreting a set of skeletal radiographs to detect fractures. The sensitivity/specificity points for the ROC curve are calculated by considering what would happen if the exam were interpreted as positive using each successive level of certainty on the ordinal scale as the criterion threshold. Each sensitivity/specificity pair is plotted, and the result is a type of ROC curve known as the empirical ROC curve (Fig. 4a). For example, if we consider the case to be positive only if the reader had a confidence level of 4, then the sensitivity would be only 0.58, relatively low because there were many positive cases with lower confidence ratings. On the other hand, the specificity would be 0.92, relatively high because few negative cases were associated with high confidence levels. The points at each end of the curve represent extreme situations in which all cases were considered positive (point A in Table 1 and Fig. 4a) or all were considered negative (point E in Table 1 and Fig. 4a).

The area under the empirical ROC curve, known as the empirical AUC, can be simply calculated by adding up the areas of the trapezoidal sections under each segment of the curve (Fig. 4a), a method known as the trapezoidal rule. Because the segments only approximate a smooth ROC curve, the empirical AUC will slightly underestimate the actual value for the area. However, this underestimation is usually negligible (7).

A potentially more accurate and visually pleasing ROC curve can be obtained by fitting a smooth curve to the individual sensitivity/specificity data points (Fig. 4b). As with any curve-fitting procedure, one must assume an underlying model for the fitted curve. For example, in linear regression, the underlying model is that of a straight line. In ROC analysis, the underlying model is more complicated: the binormal model (1,4,7). The binormal model hypothesizes that the observer's ratings for the cases form two normal distributions, one distribution for the positive cases and one for the negative cases (Fig. 5). Because of uncertainties and imperfections in the diagnostic exam, the two distributions overlap.

Table 1
Tally of Observer Ratings From an Experiment in Detecting Fractures From a Set of Skeletal Radiographs

	Observer rating				
	1	2	3	4	
Number of truly positive cases	3	1	6	14	
Number of truly negative cases	12	7	3	2	
	↑	↑	↑	↑	↑
Threshold rating between negative and positive	Below 1	Between 1 and 2	Between 2 and 3	Between 3 and 4	Above 4
Sensitivity at indicated threshold	1.00	0.88	0.83	0.58	0.00
Specificity at indicated threshold	0.00	0.50	0.79	0.92	1.00
Point in Fig. 4a	A	B	C	D	E

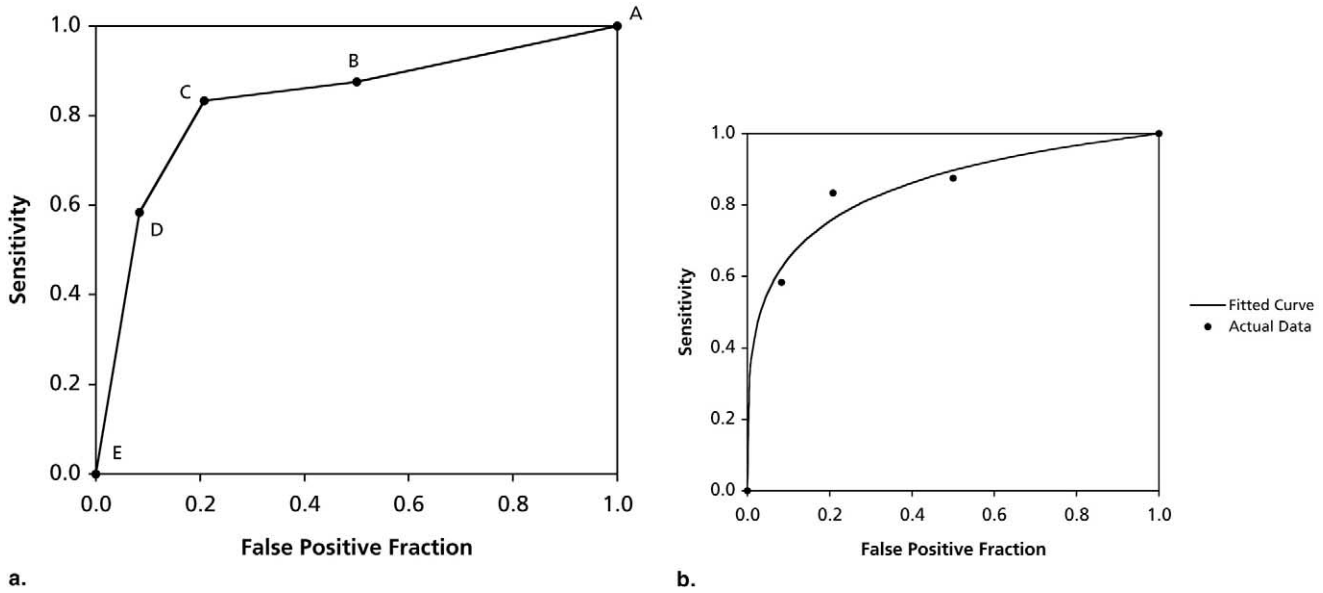

Figure 4. **(a)** Receiver operating characteristic curve plotted from the data in Table 1. **(b)** Smooth receiver operating characteristic curve fitted to the data in Table 1 using the binormal model.

It is important to note that the fitted ROC curve is based on an underlying model, and the curve's smoothness should not imply a higher degree of precision and continuity than exists in the underlying data. Nevertheless, most ROC curves are published showing only their fitted form. Therefore, editors and readers of the literature should request that the individual data points be plotted along with fitted ROC curves (Fig. 4b).

FITTING AND STATISTICAL ANALYSIS OF ROC CURVES

The mathematical transformation of the binormal model into an ROC curve (1,4) is a topic that is beyond the introductory scope of this article. Even more complicated are the numerous statistical methods that have been developed—and continue to be developed—to fit an ROC curved based on the binormal and other models. However, some appreciation for these statistical methods may be obtained by considering the different situations in which they are applied.

Up to this point, we have considered the simplest possible case: the ROC curve of one reader performing a single task and rating his or her confidence according to an ordinal scale. Despite its simplicity, this case corresponds to one of the most common ways to calculate an ROC curve, the ROCFIT program (8), which is an implementation of a popular maximum likelihood method (9).

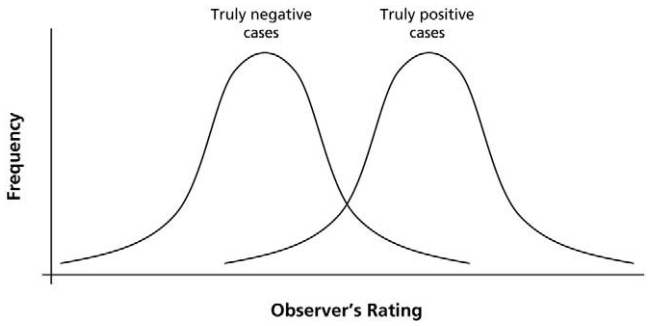

Figure 5. The binormal model of the distribution of observer ratings.

The results of many diagnostic tests, however, are expressed on a continuous numerical scale rather than a discrete ordinal one. Examples include serum levels expressed as mg/dL or tests that yield a numerical probability. An empirical ROC curve can still be constructed from such continuous data, but ROCFIT cannot be used to fit a smooth ROC curve because the program's algorithm assumes ordinal data. Two major approaches have been developed to handle continuous data. First, the continuous data can be grouped into a number of ordinal categories, transforming the data into a form that can be analyzed by the ROCFIT method (the LABROC program) (10). A more recent approach recasts the ROC curve into a form that can be directly fitted using generalized linear modeling (11). In the latter approach, ROC analysis is made more tractable because generalized linear modeling, essentially a generalized form of linear regression, is already an established and well-understood statistical method.

For all ROC calculation methods, it is possible to compare statistically two ROC curves, such as the ROC curves associated with two different imaging modalities (1–3,8,12). As with any statistical comparison method, methods for comparing ROC curves rely on estimates of the statistical variability (analogous to a standard deviation) of the fitted curves. The difference between the fitted curves is calculated, and this difference is compared with the variability estimates to see if the difference is statistically significant. Selection of an appropriate method depends upon whether the same cases were used for both ROC curves (with the curves differing by reader or modality). If the same cases are used, some statistical correlation will exist between the two curves. Accurate comparison of the two curves must account for this correlation.

Thus, ROC analysis can handle the comparison of two ROC curves, whether they are derived from two different readers or two different modalities. What if the clinical situation involves both multiple readers and multiple modalities? Analysis of this situation represents another jump in statistical complexity. Several methods have been proposed (13–18). The methods employ various strategies to account for the complex statistical correlation that arises when using a fixed set of observers to interpret a fixed number of imaging modalities using a fixed set of cases. The methods have differing statistical assumptions, and when used to analyze the same data, the methods can produce slightly different results (19,20).

An ultimate layer of complexity is to consider how external factors, which statisticians call covariates, affect the ROC curve. For example, one might want to know if certain reader or patient characteristics affect the ROC curve in addition to the main effect (eg, the imaging modality) being examined. You might also want to improve the efficiency of an ROC study by adjusting for differences in the readers or patients that would otherwise interfere with the results. A number of the methods for analyzing data from multiple readers and modalities can also be used to analyze covariates (11,13–15).

A variety of computer programs are available for fitting ROC curves and performing ROC analysis. These programs are primarily intended for those with some statistical experience. Perhaps the most widely used programs are those in the ROCKIT package (containing ROCFIT and LABROC) by Metz and colleagues for calculating single ROC curves and certain comparisons of two ROC curves (21). A web version of ROCFIT is available (22). ROC curve fitting can also be performed by a number of dedicated commercial programs (23) and some standard statistical packages (Stata, version 8.0, Stata Corp., College Station, TX). Software for more complex situations, such as multiple readers with multiple modalities (21,24) and the analysis of covariates (25), is available.

CAVEATS AND OTHER ISSUES

While ROC analysis addresses the variance of sensitivity and specificity due to variance in interpretation thresholds, it is still subject to some of the limitations affecting studies of sensitivity and specificity. First, only binary diagnostic states can be considered, such as the presence or absence of a particular disease. Conventional ROC

analysis cannot be applied to situations in which there are more than two possible outcomes, such as a screening procedure for abdominal pain, which could reveal one of many possible conditions. Second, ROC analysis still requires a reference standard that indicates the true state (diagnosis) of each case. An imperfect reference standard would introduce potential inaccuracy in the ROC analysis, just as it would for the determination of sensitivity and specificity. In the case of sensitivity and specificity, mathematical methods exist that adjust for an imperfect reference standard (2,3). In the case of ROC analysis, adjustment methods are less well developed (3).

An imperfect reference standard is an important type of bias, a term defined by epidemiologists as any deviation of study results or inferences from the truth as a consequence of how a study is designed or conducted (26). As implied by this broad definition, clinical research is potentially susceptible to countless types of bias. Studies of diagnostic tests, such as those generating data for ROC analysis, are particularly susceptible to certain types of bias (27,28), such as an imperfect reference standard. Another potential source of bias in these studies is verification bias, where the reference standard is not applied to all participants undergoing the diagnostic test being examined. For example, the reference standard test may be invasive or expensive so that there may be a tendency to obtain it only in high-risk patients or those who have a positive diagnostic screening test. As a result, the apparent accuracy of the screening test may be distorted. Mathematical methods are available to adjust for verification bias in certain situations (2,3).

If the ROC analysis is based on an ordinal rating scale, care should be take to ensure the scale is truly ordinal and represents a monotonic progression of disease suspicion. For example, the 6 assessment categories (0 to 5) of the Breast Imaging Reporting and Data System (BI-RADS) (29) seem to be readily amenable to ROC analysis. However, there is some controversy because category 0 indicates a need for more information rather than the level of confidence for malignancy. Categories 1 (negative) and 2 (benign finding) both indicate no mammographic evidence for malignancy, which might be considered the same level of confidence for disease.

ROC curves derived from several readers of the same cases are sometimes averaged to yield an ROC curve representing the overall performance of the diagnostic exam in question (Fig. 6) (30). A subtle, but important assumption is sometimes made when this pooling is done. The assumption is that all observers have the same underlying

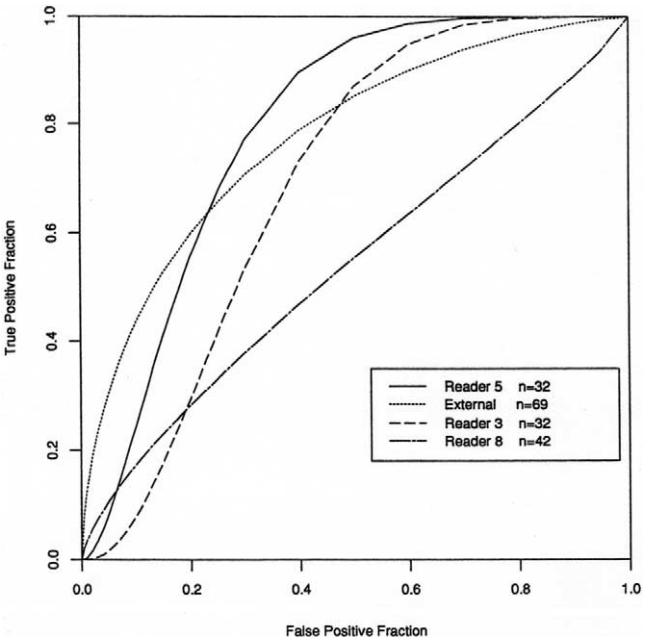

Figure 6. Receiver operating characteristic curves of several representative readers in a multi-institutional study to detect periprostatic invasion of prostate cancer with MR imaging (30). The receiver operating characteristic curves from these and other readers in the study were pooled and an area of 0.61 was reported.

ROC curve and only differ from each other because of variations in interpretation thresholds and random statistical variation. The pooling of ROC curves may not be valid if one or more of the observers is better or worse than the others in a nonrandom way. For example, one could wonder if the rather large difference between the readers in Fig. 6 were due to nonrandom factors such as clinical skill or uncontrolled institutional differences. Methods have been proposed to combine ROC data properly (31).

The AUC is the most commonly used index of performance associated with ROC analysis, but it suffers from a major limitation: It is a global indicator of diagnostic performance representing the average performance over the entire range of possible sensitivities and specificities. The AUC may be insensitive to significant differences in performance for isolated regions of the ROC curve because two curves may differ in their shape but encompass the same total area (7). Furthermore, not all regions of the ROC curve have equal clinical importance. At the ends of the ROC curve, for example, either the sensitivity or specificity is nearly zero. It is unlikely that a diagnostic test with a near-zero sensitivity or specificity would be clinically useful. Clinically relevant sensitivities or speci-

ficities are often somewhere away from the ends of the ROC curve. Therefore, some investigators have proposed using the area under part of the ROC curve, such as between two arbitrarily defined values of the false positive fraction (along the horizontal axis) (32) or sensitivity (along the vertical axis) (33). Methods for analyzing the entire ROC curve have been extended to the statistical analysis of the partial area (34). The choice of the appropriate range along the horizontal or vertical axis depends on the clinical setting. In a clinical setting in which it is important to "rule out" a disease (eg, a fatal disease if untreated), a range of relatively high false positive fractions—corresponding to high sensitivities—would be chosen. In a clinical setting in which it is important to "rule in" a disease (eg, a disease whose treatment has major side effects), a range of relatively low false positive fractions—equivalent to high specificities—would be chosen. The range of false positive fractions can also be narrowed to the extreme of selecting one particular value of false positive fraction at which to compare the sensitivity between two ROC curves (2).

A situation for which conventional ROC analysis is not well-equipped is when the diagnostic task involves determining the location of the abnormality or disease in addition to determining its presence. Obviously, this situation is a common one in medical imaging. The problem arises because ROC analysis only allows the reader to make a diagnosis that applies to the entire case rather than a particular location within an image. As a result, ROC analysis may overestimate the observer's performance because the observer is credited for all truly positive cases that the observer calls positive, even if the observer identified the wrong part of the image as being positive. Techniques to address this problem have been proposed, most notably the localization ROC (LROC) curve (8,35,36). In this approach, a correct interpretation of a truly positive case requires the reader to detect the abnormality and correctly describe its location on the image. The equations describing the LROC curve are based on the reader's conventional ROC curve. LROC curves are not common in the radiology literature, perhaps because software for performing this analysis is not widely available and because statistical procedures for fitting LROC curves are a relatively recent development (37).

Another situation not well handled by conventional ROC analysis is when more than one occurrence of the disease or abnormality can occur within an image (eg, lung nodules in a chest radiograph). To handle this situation, free-response ROC (FROC) analysis has been proposed (8,36,38). The vertical axis of a FROC curve is the sensitivity over all truly positive locations in the case set (where each case can contribute more than one truly positive location), and the horizontal axis is the average number of false positives per case (since each case can have more than one location for which a diagnosis must be specified). Like LROC analysis, FROC analysis is not commonly found in the radiology literature. Software to perform FROC analysis is not widely available. FROC analysis also suffers from the problem of assuming all possible locations of the multiple abnormalities to be independent of each other. This is unlikely to be true in most clinical settings. For example, if a lung mass is found in one location of the lung, it is more likely that others will be present at other locations. Therefore, all locations within the patient are statistically correlated, not independent.

SUMMARY

In summary, the ROC curve has found many useful applications in radiology. While the statistics and mathematics behind ROC analysis can be complex, the ROC curve is fundamentally just a plot of the trade-off between sensitivity and specificity. In fact, many of the assumptions of ROC analysis (binary classification, reliance on a reference standard) are the same as those necessary to calculate sensitivity and specificity. An indication of an observer's degree of diagnostic certainty is the key additional data element that must be collected to calculate an ROC curve. Ongoing developments in ROC analysis will address more complex types of diagnostic situations and will likely expand the applicability of ROC analysis.

REFERENCES

1. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
2. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York, NY: Wiley & Sons, 2002.
3. Pepe MS. Statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press, 2003.
4. Obuchowski NA. ROC analysis. Am J Roentgenol 2005; 184:364–372.
5. Lusted LB. Signal detectability and medical decision-making. Science 1971; 171:1217–1219.
6. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
7. Dwyer AJ. In pursuit of a piece of the ROC. Radiology 1996; 201:621–625.
8. Metz CE. Practical issues in ROC studies. Invest Radiol 1989; 24:234–245.
9. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. J Math Psychol 1969; 6:487–496.

10. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. Stat Med 1998; 17:1033–1053.
11. Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. Biometrics 2000; 56:352–359.
12. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. Biometrics 1988; 44:837–845.
13. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27:723–731.
14. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. Commun Stat Simul Comput 1995; 24:285–308.
15. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. Stat Med 1996; 15:1807–1826.
16. Song HH. Analysis of correlated ROC areas in diagnostic testing. Biometrics 1997; 53:370–382.
17. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. Acad Radiol 2000; 7:341–349.
18. Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. Can J Stat 2000; 28: 731–750.
19. Toledano AY. Three methods for analyzing correlated ROC curves: a comparison of real data sets from multi-reader, multi-case studies with a factorial design. Stat Med 2003; 22:2919–2933.
20. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. Acad Radiol 2004; 11:980–995.
21. ROCKIT and LABMRMC [computer programs]. Chicago: University of Chicago [updated 2004 Aug 3; cited 2005 Mar 28]. Avalable from: http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index.htm.
22. Eng J. ROC analysis: web-based calculator for ROC curves [computer program]. Baltimore: Johns Hopkins University [updated 2004 Mar 31; cited 2005 Mar 28]. Available from: http://www.rad.jhmi.edu/roc.
23. Stephan C, Wesseling S, Schink T, Jung K. Comparison of eight computer programs for receiver operating characteristic analysis. Clin Chem 2003; 49:433–439.
24. Obuchowski NA. OBUMRM [computer program]. Cleveland: Cleveland Clinic Foundation [cited 2005 Mar 28]. Available from: http://www.bio.ri.ccf.org/html/obumrm.html.
25. Pepe MS. The statistical evaluation of medical tests for classification and prediction [web site]. Seattle: University of Washington [updated 2003 Aug 29; cited 2005 Mar 28]. Available from: http://www.fhcrc.org/labs/pepe/book/.
26. Last JM, ed. Dictionary of epidemiology. 4th ed. Oxford: Oxford University Press, 2001; 14.
27. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. Radiology 1988; 167:565–569.
28. Obuchowski NA. Special topics III: bias. Radiology 2003; 229:617–621.
29. American College of Radiology. Breast imaging reporting and data system. 4th ed. Reston, VA: American College of Radiology; 2003.
30. Tempany CM, Zhou X, Zerhouni EA, et al. Staging of prostate cancer: results of radiology diagnostic oncology group project comparison of three MR imaging techniques. Radiology 1994; 192:47–54.
31. Zhou XH. Empirical Bayes combination of estimated areas under ROC curves using estimating equations. Med Decis Making 1996; 16:24–28.
32. McClish DK. Analyzing a portion of the ROC curve. Med Dec Making 1989; 9:190–195.
33. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745–750.
34. Dodd LE, Pepe MS. Partial AUC estimation and regression. Biometrics 2003; 59:614–623.
35. Star SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. Radiology 1975; 116:533–538.
36. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Med Phys 1996; 23: 1709–1725.
37. Swensson RG. Using localization data from image interpretations to improve estimates of performance accuracy. Med Decision Making 2000; 20:170–185.
38. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A free response approach to the measurement and characterization of radiographic observer performance. Proc SPIE 1977; 127:124–135.