Fundamentals of Clinical Research for Radiologists

Nancy A. Obuchowski[1]

ROC Analysis

Received October 28, 2004; accepted after revision November 3, 2004.

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

This is the 14th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, 9500 Euclid Ave., Cleveland, OH. Address correspondence to N. Obuchowski.

AJR 2005;184:364–372

0361–803X/05/1842–364

In this module we describe the standard methods for characterizing and comparing the accuracy of diagnostic and screening tests. We motivate the use of the receiver operating characteristic (ROC) curve, provide definitions and interpretations for the common measures of accuracy derived from the ROC curve (e.g., the area under the ROC curve), and present recent examples of ROC studies in the radiology literature. We describe the basic statistical methods for fitting ROC curves, comparing them, and determining sample size for studies using ROC curves. We briefly describe the MRMC (multiple-reader, multiple-case) ROC paradigm. We direct the interested reader to available software for analyzing ROC studies and to literature on more advanced statistical methods of ROC analysis.

Why ROC?

In module 13 [1], we defined the basic measures of accuracy: sensitivity (the probability the diagnostic test is positive for disease for a patient who truly has the disease) and specificity (the probability the diagnostic test is negative for disease for a patient who truly does not have the disease). These measures require a decision rule (or positivity threshold) for classifying the test results as either positive or negative. For example, in mammography the BI-RADS (Breast Imaging Reporting and Data System) scoring system is used to classify mammograms as normal, benign, probably benign, suspicious, or malignant. One positivity threshold is classifying probably benign, suspicious, and malignant findings as positive (and classifying normal and benign findings as negative). Another positivity threshold is classifying suspicious and malignant findings as positive. Each threshold leads to different estimates of sensitivity and specificity. Here, the second threshold would have higher specificity than the first but lower sensitivity. Also, note that trained mammographers use the scoring system differently. Even the same mammographer may use the scoring system differently on different reviewing occasions (e.g., classifying the same mammogram as probably benign on one interpretation and as suspicious on another), leading to different estimates of sensitivity and specificity even with the same threshold.

Which decision threshold should be used to classify test results? How will the choice of a decision threshold affect comparisons between two diagnostic tests or between two radiologists? These are critical questions when computing sensitivity and specificity, yet the choice for the decision threshold is often arbitrary.

ROC curves, although constructed from sensitivity and specificity, do not depend on the decision threshold. In an ROC curve, every possible decision threshold is considered. An ROC curve is a plot of a test's false-positive rate (FPR), or 1 – specificity (plotted on the horizontal axis), versus its sensitivity (plotted on the veritical axis). Each point on the curve represents the sensitivity and FPR at a different decision threshold. The plotted (FPR, sensitivity) coordinates are connected with line segments to construct an empiric ROC curve. Figure 1 illustrates an empiric ROC curve constructed from the fictitious mammography data in Table 1. The empiric ROC curve has four points corresponding to the four decision thresholds described in Table 1.

An ROC curve begins at the (0, 0) coordinate, corresponding to the strictest decision threshold whereby all test results are negative for disease (Fig. 1). The ROC curve ends at the (1, 1) coordinate, corresponding to the

most lenient decision threshold whereby all test results are positive for disease. An empiric ROC curve has $h - 1$ additional coordinates, where h is the number of unique test results in the sample. In Table 1 there are 200 test results, one for each of the 200 patients in the sample, but there are only five unique results: normal, benign, probably benign, suspicious, and malignant. Thus, $h = 5$, and there are four coordinates plotted in Figure 1 corresponding to the four decision thresholds described in Table 1.

The line connecting the (0, 0) and (1, 1) coordinates is called the "chance diagonal" and represents the ROC curve of a diagnostic test with no ability to distinguish patients with versus those without disease. An ROC curve that lies above the chance diagonal, such as the ROC curve for our fictitious mammography example, has some diagnostic ability. The further away an ROC curve is from the chance diagonal, and therefore, the closer to the upper left-hand corner, the better discriminating power and diagnostic accuracy the test has.

In characterizing the accuracy of a diagnostic (or screening) test, the ROC curve of the test provides much more information about how the test performs than just a single estimate of the test's sensitivity and specificity [1, 2]. Given a test's ROC curve, a clinician can examine the trade-offs in sensitivity versus specificity for various decision thresholds. Based on the relative costs of false-positive and false-negative errors and the pretest probability of disease, the clinician can choose the optimal decision threshold for each patient. This idea is discussed in more detail in a later section of this article. Often, patient management is more complex than is allowed with a decision threshold that classifies the test results into positive or negative. For example, in mammography suspicious and malignant findings are usually followed up with biopsy, probably benign findings usually result in a follow-up mammogram in 3–6 months, and normal and benign findings are considered negative.

When comparing two or more diagnostic tests, ROC curves are often the only valid method of comparison. Figure 2 illustrates two scenarios in which an investigator, comparing two diagnostic tests, could be misled by relying on only a single sensitivity–specificity pair. Consider Figure 2A. Suppose a more expensive or risky test (represented by ROC curve Y) was reported to have the following accuracy: sensitivity = 0.40, specificity = 0.90 (labeled as coordinate 1 in Fig. 2A); a less expensive or less risky test (represented by ROC curve X) was reported to have the following accuracy: sensitivity = 0.80, specificity = 0.65 (labeled as coordinate 2 in Fig. 2A). If the investigator is looking for the test with better specificity, then he or she may choose the more expensive, risky test, not realizing that a simple change in the decision threshold of the less expensive, cheaper test could provide the desired specificity at an even higher sensitivity (coordinate 3 in Fig. 2A).

Now consider Figure 2B. The ROC curve for test Z is superior to that of test X for a narrow range of FPRs (0.0–0.08); otherwise, diagnostic test X has superior accuracy. A comparison of the tests' sensitivities at low FPRs would be misleading unless the diagnostic tests are useful only at these low FPRs.

TABLE 1 **Construction of Receiver Operating Characteristic Curve Based on Fictitious Mammography Data**

Mammography Results (BI-RADs Score)	Pathology/Follow-Up Results		Decision Rules 1–4	
	Not Malignant	Malignant	FPR	Sensitivity
Normal	65	5	(1) 35/100	95/100
Benign	10	15	(2) 25/100	80/100
Probably benign	15	10	(3) 10/100	70/100
Suspicious	7	60	(4) 3/100	10/100
Malignant	3	10		
Total	100	100		

Note.—Decision rule 1 classifies normal mammography findings as negative; all others are positive. Decision rule 2 classifies normal and benign mammography findings as negative; all others are positive. Decision rule 3 classifies normal, benign, and probably benign findings as negative; all others are positive. Decision rule 4 classifies normal, benign, probably benign, and suspicious findings as negative; malignant is the only finding classified as positive. BI-RADS = Breast Imaging Reporting and Data System, FPR = false-positive rate.

To compare two or more diagnostic tests, it is convenient to summarize the tests' accuracies with a single summary measure. Several such summary measures are used in the literature. One is Youden's index, defined as sensitivity + specificity − 1 [2]. Note, however, that Youden's index is affected by the choice of the decision threshold used to define sensitivity and specificity. Thus, different decision thresholds yield different values of the Youden's index for the same diagnostic test.

Another summary measure commonly used is the probability of a correct diagnosis, often referred to simply as "accuracy" in the literature. It can be shown that the probability of a correct diagnosis is equivalent to

$$\text{probability (correct diagnosis)} = \text{PREV}_s \times \text{sensitivity} + (1 - \text{PREV}_s) \times \text{specificity}, \quad (1)$$

where PREV_s is the prevalence of disease in the sample. That is, this summary measure of accuracy is affected not only by the choice of the decision threshold but also by the prevalence of disease in the study sample [2]. Thus, even slight changes in the prevalence of disease in the population of patients being tested can lead to different values of "accuracy" for the same test.

Summary measures of accuracy derived from the ROC curve describe the inherent accuracy of a diagnostic test because they are not affected by the choice of the decision threshold and they are not affected by the prevalence of disease in the study sample. Thus, these summary measures are preferable to Youden's index and the probability of a correct diagnosis [2]. The most popular summary measure of accuracy is the area under the ROC curve, often denoted as "AUC" for area under curve. It ranges in value from 0.5 (chance) to 1.0 (perfect discrimination or accuracy). The chance diagonal in Figure 1 has an AUC of 0.5. In Figure 2A the areas under both ROC curves are the same, 0.841. There are three interpretations for the AUC: the average sensitivity over all false-positive rates; the average specificity over all sensitivities [3]; and the probability that, when presented with a randomly chosen patient with disease and a randomly chosen patient without disease, the results of the diagnostic test will rank the patient with disease as having higher suspicion for disease than the patient without disease [4].

The AUC is often too global a summary measure. Instead, for a particular clinical application, a decision threshold is chosen so that the diagnostic test will have a low FPR

(e.g., FPR < 0.10) or a high sensitivity (e.g., sensitivity > 0.80). In these circumstances, the accuracy of the test at the specified FPRs (or specified sensitivities) is a more meaningful summary measure than the area under the entire ROC curve. The partial area under the ROC curve, PAUC (e.g., the PAUC where FPR < 0.10, or the PAUC where sensitivity > 0.80), is then an appropriate summary measure of the diagnostic test's accuracy. In Figure 2B, the PAUCs for the two tests where the FPR is between 0.0 and 0.20 are the same, 0.112. For interpretation purposes, the PAUC is often divided by its maximum value, given by the range (i.e., maximum–minimum) of the FPRs (or false-negative rates [FNRs]) [5]. The PAUC divided by its maximum value is called the partial area index and takes on values between 0.5 and 1.0, as does the AUC. It is interpreted as the average sensitivity for the FPRs examined (or average specificity for the FNRs examined). In our example, the range of the FPRs of interest is 0.20–0.0 = 0.20; thus, the average sensitivity for FPRs less than 0.20 for diagnostic tests X and Z in Figure 2B is 0.56.

Although the ROC curve has many advantages in characterizing the accuracy of a diagnostic test, it also has some limitations. One criticism is that the ROC curve extends beyond the clinically relevant area of potential clinical interpretation. Of course, the PAUC was developed to address this criticism. Another criticism is that it is possible for a diagnostic test with perfect discrimination between diseased and nondiseased patients to have an AUC of 0.5. Hilden [6] describes this unusual situation and offers solutions. When comparing two diagnostic tests' accuracies, the tests' ROC curves can cross, as in Figure 2. A comparison of these tests based only on their AUCs can be misleading. Again, the PAUC attempts to address this limitation. Last, some [6, 7] criticize the ROC curve, and especially the AUC, for not incorporating the pretest probability of disease and the costs of misdiagnoses.

The ROC Study

Weinstein et al. [1] describe the common features of a study of the accuracy of a diagnostic test. These include samples from both patients with and those without the disease of interest and a reference standard for determining whether positive test results are true-positives or false-positives, and whether negative test results are true-negatives or false-negatives. They also discuss the need to blind reviewers who are interpreting test images and other relevant biases common to these types of studies.

In ROC studies we also require that the test results, or the interpretations of the test images, be assigned a numeric value or rank. These numeric measurements or ranks are the basis for defining the decision thresholds that yield the estimates of sensitivity and specificity that are plotted to form the ROC curve. Some diagnostic tests yield an objective measurement (e.g., attenuation value of a lesion). The decision thresholds for constructing the ROC curve are

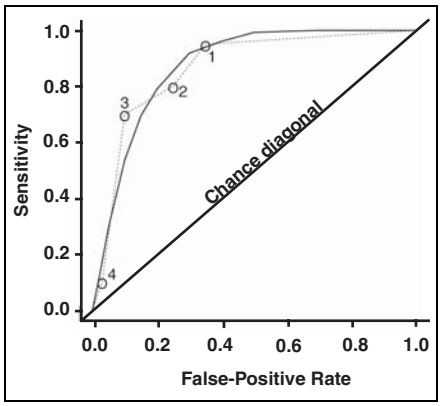

Fig. 1.—Empiric and fitted (or "smooth") receiver operating characteristic (ROC) curves constructed from mammography data in Table 1. Four labeled points on empiric curve (*dotted line*) correspond to four decision thresholds used to estimate sensitivity and specificity. Area under curve (AUC) for empiric ROC curve is 0.863 and for fitted curve (*solid line*) is 0.876.

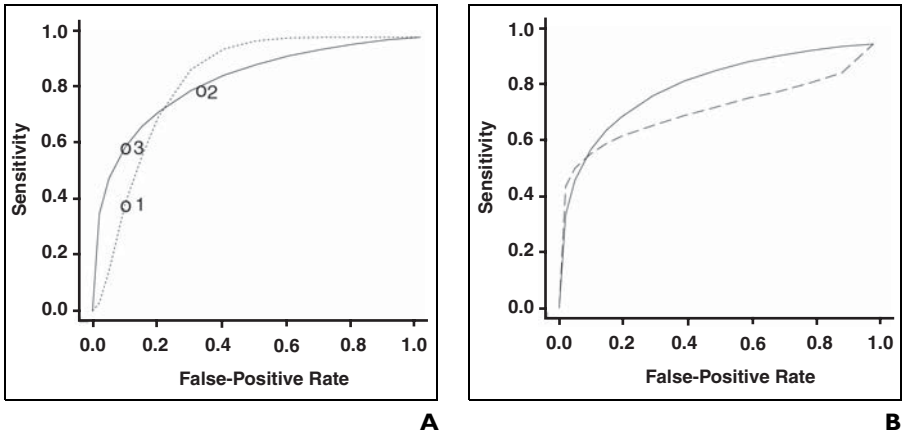

Fig. 2.—Two examples illustrate advantages of receiver operating characteristic (ROC) curves (see text for explanation) and comparing summary measures of accuracy.
A, ROC curve Y (*dotted line*) has same area under curve (AUC) as ROC curve X (*solid line*), but lower partial area under curve (PAUC) when false-positive rate (FPR) is ≤ 0.20, and higher PAUC when false-positive rate > 0.20.
B, ROC curve Z (*dashed line*) has same PAUC as curve X (*solid line*) when FPR ≤ 0.20 but lower AUC.

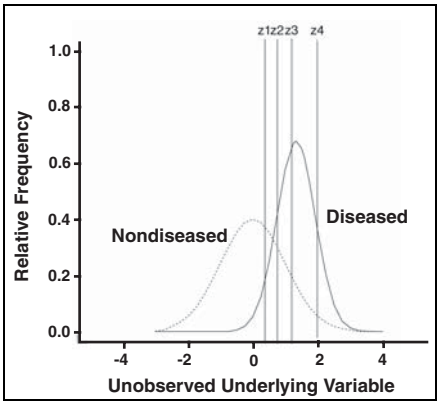

Fig. 3.—Unobserved binormal distribution that was assumed to underlie test results in Table 1. Distribution for nondiseased patients was arbitrarily centered at 0 with SD of 1 (i.e., $\mu_0 = 0$ and $\sigma_0 = 1$). Binormal parameters were estimated to be A = 2.27 and B = 1.70. Thus, distribution for diseased patients is centered at $\mu_1 = 1.335$ with SD of $\sigma_1 = 0.588$. Four cutoffs, z1, z2, z3, and z4, correspond to four decision thresholds in Table 1. If underlying test value is less than z1, then mammographer assigns test result of "normal." If the underlying test value is less than z2 but greater than z1, then mammographer assigns test result of "benign," and so forth.

based on increasing the values of the attenuation coefficient. Other diagnostic tests must be interpreted by a trained observer, often a radiologist, and so the interpretation is subjective. Two general scales are often used in radiology for observers to assign a value to their subjective interpretation of an image. One scale is the 5-point rank scale: 1 = definitely normal, 2 = probably normal, 3 = possibly abnormal or equivocal, 4 = probably abnormal, and 5 = definitely abnormal.

The other popular scale is the 0–100% confidence scale, where 0% implies that the observer is completely confident in the absence of the disease of interest, and 100% implies that the observer is completely confident in the presence of the disease of interest. The two scales have strengths and weaknesses [2, 8], but both are reasonably well suited to radiology research. In mammography a rating scale already exists, the BI-RADS score, which can be used to form decision thresholds from least to most suspicion for the presence of breast cancer.

When the diagnostic test requires a subjective interpretation by a trained reviewer, the reviewer becomes part of the diagnostic process [9]. Thus, to properly characterize the accuracy of the diagnostic test, we must include multiple reviewers in the study. This is the so-called MRMC, multiple-reader multiple-case, ROC study. Much has been written about the design and analysis of MRMC studies [10–20]. We mention here only the basic design of MRMC studies, and in a later subsection we describe their statistical analysis.

The usual design for the MRMC study is a factorial design, in which every reviewer interprets the image (or images if there is more than one test) of every patient. Thus, if there are R reviewers, C patients, and I diagnostic tests, then each reviewer interprets $C \times I$ images, and the study involves $R \times C \times I$ total interpretations. The accuracy of each reviewer with each diagnostic test is characterized by an ROC curve, so $R \times I$ ROC curves are constructed. Constructing pooled or consensus ROC curves is not the goal of these studies. Rather, the primary goals are to document the variability in diagnostic test accuracy between reviewers and report the average, or typical, accuracy of reviewers. In order for the results of the study to be generalizeable to the relevant patient and reviewer populations, representative samples from both populations are needed for the study. Often expert reviewers take part in studies of diagnostic test accuracy, but the accuracy for a nonexpert may be considerably less. An excellent illustration of the issues involved in sampling reviewers for an MRMC study can be found in the study by Beam et al. [21].

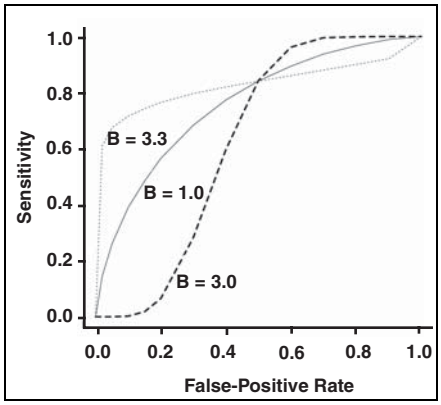

Fig. 4.—Three receiver operating characteristic (ROC) curves with same binormal parameter A (i.e., A = 1.0) but different values for parameter B of 3.0 ($3\sigma_1 = \sigma_0$), 1.0 ($\sigma_1 = \sigma_0$), and 0.33 ($\sigma_1 = 3\sigma_0$).When B = 3.0, ROC curve dips below chance diagonal; this is called an improper ROC curve [2].

Examples of ROC Studies in Radiology

The radiology literature, and the clinical laboratory and more general medical literature, contain many excellent examples of how ROC curves are used to characterize the accuracy of a diagnostic test and to compare accuracies of diagnostic tests. We briefly describe here three recent examples of ROC curves being used in the radiology literature.

Kim et al. [22] conducted a prospective study to determine if rectal distention using warm water improves the accuracy of MRI for preoperative staging of rectal cancer. After MRI, the patients underwent surgical resection, considered the gold standard regarding the invasion of adjacent structures and regional lymph node involvement. Four observers, unaware of the pathology results, independently scored the MR images using 4- and 5-point rating scales. Using statistical methods for MRMC studies [13], the authors determined that typical reviewers' accuracy for determining outer wall penetration is improved with rectum distention, but that reviewer accuracy for determining regional lymph node involvement is not affected.

Osada et al. [23] used ROC analysis to assess the ability of MRI to predict fetal pulmonary hypoplasia. They imaged 87 fetuses, measuring both lung volume and signal intensity. An ROC curve based on lung volume showed that lung volume has some ability to discriminate between fetuses who will have good versus those who will have poor respiratory outcome after birth. An ROC curve based on the combined information from lung volume and signal intensity, however, has superior accuracy. For more information on the optimal way to combine measures or test results, see the article by Pepe and Thompson [24].

In a third study, Zheng et al. [25] assessed how the accuracy of a mammographic computer-aided detection (CAD) scheme was affected by restricting the maximum number of regions that could be identified as positive. Using a sample of 300 cases with a malignant mass and 200 normals, the investigators applied their CAD system, each time reducing the maximum number of positive regions that the CAD system could identify from seven to one. A special ROC technique called "free-response receiver operating characteristic curves" (FROC) was used. The horizontal axis of the FROC curve differs from the traditional ROC curve in that it gives the average number of false-positives per image. Zheng et al. concluded that limiting the maximum number of positive regions that the CAD could identify improves the overall accuracy of CAD in mammography. For more information on FROC curves and related methods, I refer you to other articles [26–29].

Statistical Methods for ROC Analysis

Fitting Smooth ROC Curves

In Figure 1 we saw the empiric ROC curve for the test results in Table 1. The curve was constructed with line segments connecting the observed points on the ROC curve. Empiric ROC curves often have a jagged appearance, as seen in Figure 1, and often lie slightly below the "true," smooth, ROC curve—that is, the test's ROC curve if it were constructed with an infinite number of points (not just the four points in Fig. 1) and an infinitely large sample size. A smooth curve gives us a better idea of the relationship between the diagnos-

tic test and the disease. In this subsection we describe some methods for constructing smooth ROC curves.

The most popular method of fitting a smooth ROC curve is to assume that the test results (e.g., the BI-RADS scores in Table 1) come from two unobserved distributions, one distribution for the patients with disease and one for the patients without the disease. Usually it is assumed that these two distributions can be transformed to normal distributions, referred to as the binormal assumption. It is the unobserved, underlying distributions that we assume can be transformed to follow a binormal distribution, and not the observed test results. Figure 3 illustrates the hypothesized unobserved binormal distribution estimated for the observed BI-RADS results in Table 1. Note how the distributions for the diseased and nondiseased patients overlap.

Let the unobserved binormal variables for the nondiseased and diseased patients have means μ_0 and μ_1, and variances σ_0 [2] and σ_1 [2], respectively. Then it can be shown [30] that the ROC curve is completed described by two parameters:

$$A = (\mu_1 - \mu_0) / \sigma_1 \quad (2)$$

$$B = \sigma_0 / \sigma_1. \quad (3)$$

(See Appendix 1 for a formula that links parameters A and B to the ROC curve.) Figure 4 illustrates three ROC curves. Parameter A was set to be constant at 1.0 and parameter B varies as follows: 0.33 (the underlying distribution of the diseased patients is three times more variable than that of the nondiseased patients), 1.0 (the two distributions have the same SD), and 3.0 (the underlying distribution of the nondiseased patients is three times more variable than that of the diseased patients). As one can see, the curves differ dramatically with changes in parameter B. Parameter A, on the other hand, determines how far the curve is above the chance diagonal (where A = 0); for a constant B parameter, the greater the value of A, the higher the ROC curve lies (i.e., greater accuracy).

Parameters A and B can be estimated from data such as in Table 1 using maximum likelihood methods [30, 31]. For the data in Table 1, the maximum likelihood estimates (MLEs) of parameters A and B are 2.27 and 1.70, respectively; the smooth ROC curve is given in Figure 1. Fortunately, some useful software [32] has been written to perform the necessary calculations of A and B, along with estimation of the area under the smooth curve (see next subsection), its SE and confidence interval (CI), and CIs for the ROC curve itself (see Appendix 1).

Dorfman and Alf [30] suggested a statistical test to evaluate whether the binormal assumption was reasonable for a given data set. Others [33, 34] have shown through empiric investigation and simulation studies that many different underlying distributions are well approximated by the binormal assumption.

When the diagnostic test results are themselves a continuous measurement (e.g., CT attenuation values, or measured lesion diameter), it may not be necessary to assume the existence of an unobserved, underlying distribution. Sometimes continuous-scale test results themselves follow a binormal distribution, but caution should be taken that the fit is good (see the article by Goddard and Hinberg [35] for a discussion of the resulting bias when the distribution is not truly binormal yet the binormal distribution is assumed). Zou et al. [36] suggest using a Box-Cox transformation to transform data to

TABLE 2 Estimating Area Under Empirical Receiver Operating Characteristic Curve

Test Results		Score	No. of Pairs	Score x No. of Pairs
Nondiseased	Diseased			
Normal	Normal	1/2	65 x 5 = 325	162.5
Normal	Benign	1	65 x 15 = 975	975
Normal	Probably benign	1	65 x 10 = 650	650
Normal	Suspicious	1	65 x 60 = 3,900	3,900
Normal	Malignant	1	65 x 10 = 650	650
Benign	Normal	0	10 x 5 = 50	0
Benign	Benign	1/2	10 x 15 = 150	75
Benign	Probably benign	1	10 x 10 = 100	100
Benign	Suspicious	1	10 x 60 = 600	600
Benign	Malignant	1	10 x 10 = 100	100
Probably benign	Normal	0	15 x 5 = 75	0
Probably benign	Benign	0	15 x 15 = 225	0
Probably benign	Probably benign	1/2	15 x 10 = 150	75
Probably benign	Suspicious	1	15 x 60 = 900	900
Probably benign	Malignant	1	15 x 10 = 150	150
Suspicious	Normal	0	7 x 5 = 35	0
Suspicious	Benign	0	7 x 15 = 105	0
Suspicious	Probably benign	0	7 x 10 = 70	0
Suspicious	Suspicious	1/2	7 x 60 = 420	210
Suspicious	Malignant	1	7 x 10 = 70	70
Malignant	Normal	0	3 x 5 = 15	0
Malignant	Benign	0	3 x 15 = 45	0
Malignant	Probably benign	0	3 x 10 = 30	0
Malignant	Suspicious	0	3 x 60 = 180	0
Malignant	Malignant	1/2	3 x 10 = 30	15
Total			10,000 pairs	8,632.5

TABLE 3 Fictitious Data Comparing the Accuracy of Two Diagnostic Tests

	ROC Curve	
	X	Y
Estimated AUC	0.841	0.841
Estimated SE of AUC	0.041	0.045
Estimated PAUC where FPR < 0.20	0.112	0.071
Estimated SE of PAUC	0.019	0.014
Estimated covariance	0.00001	
Z test comparing PAUCs	$Z = [0.112 - 0.071] / \sqrt{[0.019^2 + 0.014^2 - 0.00002]}$	
95% CI for difference in PAUCs	$[0.112 - 0.071] \pm 1.96 \times \sqrt{[0.019^2 + 0.014^2 - 0.00002]}$	

Note.—AUC = area under the curve, PAUC = partial area under the curve, CI = confidence interval.

binormality. Alternatively, one can use software like ROCKIT [32] that will bin the test results into an optimal number of categories and apply the same maximum likelihood methods as mentioned earlier for rating data like the BI-RADS scores.

More elaborate models for the ROC curve that can take into account covariates (e.g., the patient's age, symptoms) have also been developed in the statistics literature [37–39] and will become more accessible as new software is written.

Estimating the Area Under the ROC Curve

Estimation of the area under the smooth curve, assuming a binormal distribution, is described in Appendix 1. In this subsection, we describe and illustrate estimation of the area under the empiric ROC curve. The process of estimating the area under the empiric ROC curve is nonparametric, meaning that no assumptions are made about the distribution of the test results or about any hypothesized underlying distribution. The estimation works for tests scored with a rating scale, a 0–100% confidence scale, or a true continuous-scale variable.

The process of estimating the area under the empiric ROC curve involves four simple steps: First, the test result of a patient with disease is compared with the test result of a patient without disease. If the former test result indicates more suspicion of disease than the latter test result, then a score of 1 is assigned. If the test results are identical, then a score of 1/2 is assigned. If the diseased patient has a test result indicating less suspicion for disease than the test result of the nondiseased patient, then a score of 0 is assigned. It does not matter which diseased and nondiseased patient you begin with. Using the data in Table 1 as an illustration, suppose we start with a diseased patient assigned a test result of "normal" and a nondiseased patient assigned a test result of "normal." Because their test results are the same, this pair is assigned a score of 1/2.

Second, repeat the first step for every possible pair of diseased and nondiseased patients in your sample. In Table 1 there are 100 diseased patients and 100 nondiseased patients, thus 10,000 possible pairs. Because there are only five unique test results, the 10,000 possible pairs can be scored easily, as in Table 2.

Third, sum the scores of all possible pairs. From Table 2, the sum is 8,632.5.

Fourth, divide the sum from step 3 by the number of pairs in the study sample. In our example we have 10,000 pairs. Dividing the sum from step 3 by 10,000 gives us 0.86325, which is our estimate of the area under the empiric ROC curve. Note that this method of estimating the area under the empiric ROC curve gives the same result as one would obtain by fitting trapezoids under the curve and summing the areas of the trapezoids (so-called trapezoid method).

The variance of the estimated area under the empiric ROC curve is given by DeLong et al. [40] and can be used for constructing CIs; software programs are available for estimating the nonparametric AUC and its variance [41].

Comparing the AUCs or PAUCs of Two Diagnostic Tests

To test whether the AUC (or PAUC) of one diagnostic test (denoted by AUC_1) equals the AUC (or PAUC) of another diagnostic test (AUC_2), the following test statistic is calculated:

$$Z = [AUC_1 - AUC_2] / \sqrt{[var_1 + var_2 - 2 \times cov]}, \quad (4)$$

where var_1 is the estimated variance of AUC_1, var_2 is the estimated variance of AUC_2, and cov is the estimated covariance between AUC_1 and AUC_2. When different samples of patients undergo the two diagnostic tests, the covariance equals zero. When the same sample of patients undergoes both diagnostic tests (i.e., a paired study design), then the covariance is not generally equal to zero and is often positive. The estimated variances and covariances are standard output for most ROC software [32, 41].

The test statistic Z follows a standard normal distribution. For a two-tailed test with significance level of 0.05, the critical values are –1.96 and +1.96. If Z is less than –1.96, then we conclude that the accuracy of diagnostic test 2 is superior to that of diagnostic test 1; if Z exceeds +1.96, then we conclude that the accuracy of diagnostic test 1 is superior to that of diagnostic test 2.

A two-sided CI for the difference in AUC (or PAUC) between two diagnostic tests can be calculated from

$$LL = [AUC_1 - AUC_2] - z_{\alpha/2} \times \sqrt{[var_1 + var_2 - 2 \times cov]} \quad (5)$$

$$UL = [AUC_1 - AUC_2] + z_{\alpha/2} \times \sqrt{[var_1 + var_2 - 2 \times cov]}, \quad (6)$$

where LL is the lower limit of the CI, UL is the upper limit, and $z_{\alpha/2}$ is a value from the standard normal distribution corresponding to a probability of $\alpha/2$. For example, to construct a 95% CI, $\alpha = 0.05$, thus $z_{\alpha/2} = 1.96$.

Consider the ROC curves in Figure 2A. The estimated areas under the smooth ROC curves of the two tests are the same, 0.841. The PAUCs where the FPR is greater than 0.20, however, differ. From the estimated variances and covariance in Table 3, the value of the Z statistic for comparing the PAUCs is 1.77, which is not statistically significant. The 95% CI for the difference in PAUCs is more informative: (–0.004 to 0.086); the CI for the partial area index is (–0.02 to 0.43). The CI contains large positive differences, suggesting that more research is needed to investigate the relative accuracies of these two diagnostic tests for FPRs less than 0.20.

Analysis of MRMC ROC Studies

Multiple published methods discuss performing the statistical analysis of MRMC studies [13–20]. The methods are used to construct CIs for diagnostic accuracy and statistical tests for assessing differences in accuracy between tests. A statistical overview of the methods is given elsewhere [10]. Here, we briefly mention some of the key issues of MRMC ROC analyses.

Fixed- or random-effects models.—The MRMC study has two samples, a sample of patients and a sample of reviewers. If the study results are to be generalized to patients similar to ones in the study sample and to reviewers similar to ones in the study sample, then a statistical analysis that treats both patients and reviewers as random effects should be used [13, 14, 17–20]. If the study results are to be generalized to just patients similar to ones in the study sample, then the patients are treated as random effects but the reviewers should be treated as fixed effects [13–20]. Some of the statistical methods can treat reviewers as either random or fixed, whereas other methods treat reviewers only as fixed effects.

Parametric or nonparametric.—Some of the methods rely on models that make strong assumptions about how the accuracies of the reviewers are correlated and distributed (parametric methods) [13, 14], other methods are more flexible [15, 20], and still others make no assumptions [16–19] (nonparametric methods). The parametric methods may be more powerful when their assumptions are met, but often it is difficult to determine if the assumptions are met.

Covariates.—Reviewers' accuracy may be affected by their training or experience or by characteristics of the patients (e.g., age, sex, stage of disease, comorbidities). These variables are called covariates. Some of the statistical methods [15, 20] have models that can include covariates. These models provide valuable insight into the variability between reviewers and between patients.

Software.—Software is available for public use for some of the methods [32, 42, 43]; the authors of the other methods may be able to provide software if contacted.

Determining Sample Size for ROC Studies

Many issues must be considered in determining the number of patients needed for an ROC study. We list several of the key issues and some useful references here, followed by a simple illustration. Software is also available for determining the required sample size for some ROC study designs [32, 41].

1. Is it a MRMC ROC study? Many radiology studies include more than one reviewer but are not considered MRMC studies. MRMC studies usually involve five or more reviewers and focus on estimating the average accuracy of the reviewers. In contrast, many radiology studies include two or three reviewers to get some idea of the interreviewer variability. Estimation of the required sample size for MRMC studies requires balancing the number of reviewers in the reviewer sample with the number of patients in the patient sample. See [14, 44] for formulae for determining sample sizes for MRMC studies and [45] for sample size tables for MRMC studies. Sample size determination for non-MRMC studies is based on the number of patients needed.

2. Will the study involve a single diagnostic test or compare two or more diagnostic tests? ROC studies comparing two or more diagnostic tests are common. These studies focus on the difference between AUCs or PAUCs of the two (or more) diagnostic tests. Sample size can be based on either planning for enough statistical power to detect a clinically important difference, or constructing a CI for the difference in accuracies that is narrow enough to make clinically relevant conclusions from the study. In studies of one diagnostic test, we often focus on the magnitude of the test's AUC or PAUC, basing sample size on the desired width of a CI.

3. If two or more diagnostic tests are being compared, will it be a paired or unpaired study design, and are the accuracies of the tests hypothesized to be different or equivalent? Paired designs almost always require fewer patients than an unpaired design, and so are used whenever they are logistically, ethically, and financially feasible. Studies that are performed to determine whether two or more tests have the same accuracy are called equivalency studies. Often in radiology a less invasive diagnostic test, or a quicker imaging sequence, is developed and compared with the standard test. The investigator wants to know if the test is similar in accuracy to the standard test. Equivalency studies often require a larger sample size than studies in which the goal is to show that one test has superior accuracy to another test. The reason is that to show equivalence the investigator must rule out all large differences between the tests—that is, the CI for the difference must be very narrow.

4. Will the patients be recruited in a prospective or retrospective fashion? In prospective designs, patients are recruited based on their signs or symptoms, so at the time of recruitment it is unknown whether the patient has the disease of interest. In contrast, in retrospective designs patients are recruited based on their known true disease status (as determined by the gold or reference standard) [2]. Both studies are used commonly in radiology. Retrospective studies often require fewer patients than prospective designs.

5. What will be the ratio of nondiseased to diseased patients in the study sample? Let k denote the ratio of the number of nondiseased to diseased patients in the study sample. For retrospective studies k is usually decided in the design phase of the study. For prospective designs k is unknown in the design phase but can be estimated by $(1 - PREV_p) / PREV_p$, where $PREV_p$ is the prevalence of disease in the relevant population. A range of values for $PREV_p$ should be considered when determining sample size.

6. What summary measure of accuracy will be used? In this article we have focused mainly on the AUC and PAUC, but others are possible (see [2]). The choice of summary measures determines which variance function formula will be used in calculating sample size. Note that the variance function is related to the variance by the following formula: variance = VF / N, where VF is the variance function and N is the number of study patients with disease.

7. What is the conjectured accuracy of the diagnostic test? The conjectured accuracy is needed to determine the expected difference in accuracy between two or more diagnostic tests. Also, the magnitude of the accuracy affects the variance function. In the following example, we present the variance function for the AUC; see Zhou et al. [2] for formulae for other variance functions.

Consider the following example. Suppose an investigator wants to conduct a study to determine if MRI can distinguish benign from malignant breast lesions. Patients with a suspicious lesion detected on mammography will be prospectively recruited to undergo MRI before biopsy. The pathology results will be the reference standard. The MR images will be interpreted independently by two reviewers; they will score the lesions using a 0–100% confidence scale. An ROC curve will be constructed for each reviewer; AUCs will be estimated, and 95% CIs for the AUCs will be constructed. If MRI shows some promise, the investigator will plan a larger MRMC study.

The investigator expects 20–40% of patients to have pathologically confirmed breast cancer ($PREV_p$ = 0.2–0.4); thus, k = 1.5–4.0. The investigator expects the AUC of MRI to be approximately 0.80 or higher. The variance function of the AUC often used for sample size calculations is as follows:

$$VF = (0.0099 \times e^{-A \times A/2}) \times [(5 \times A^2 + 8) + (A^2 + 8) / k], \quad (7)$$

where A is the parameter from the binormal distribution. Parameter A can be calculated from $A = \phi^{-1}(\text{AUC}) \times 1.414$, where ϕ^{-1} is the inverse of the cumulative normal distribution function [2]. For our example, AUC = 0.80; thus $\phi^{-1}(0.80) = 0.84$ and $A = 1.18776$. The variance function, VF, equals (0.00489) × [(15.05387) + (9.41077) / 4.0] = 0.08512, where we have set k = 4.0. For k = 1.5, the VF = 0.10429.

Suppose the investigator wants a 95% CI no wider than 0.10. That is, if the estimated AUC from the study is 0.80, then the lower bound of the CI should not be less than 0.75 and the upper bound should not exceed 0.85. A formula for calculating the required sample size for a CI is

$$N = [z_{\alpha/2}^2 \times VF] / L^2 \quad (8)$$

where $z_{\alpha/2}$ = 1.96 for a 95% CI and L is the desired half-width of the CI. Here, L = 0.05. N is the number of patients with disease needed for the study; the total number of patients needed for the study is $N \times (1 + k)$. For our example, N equals $[1.96^2 \times 0.08512] / 0.05^2$ = 130.8 for k = 4.0, and 160.3 for k = 1.5. Thus, depending on the unknown prevalence of breast cancer in the study sample, the investigator needs to recruit perhaps as few as 401 total patients (if the sample prevalence is 40%) but perhaps as many as 654 (if the sample prevalence is only 20%).

Finding the Optimal Point on the Curve

Metz [46] derived a formula for determining the optimal decision threshold on the ROC curve, where "optimal" is in terms of minimizing the overall costs. "Costs" can be defined as monetary costs, patient morbidity and mortality, or both. The slope, m, of the ROC curve at the optimal decision threshold is

$$m = (1 - \text{PREV}_p) / \text{PREV}_p \times [C_{FP} - C_{TN}] / [C_{FN} - C_{TP}] \quad (9)$$

where C_{FP}, C_{TN}, C_{FN}, and C_{TP} are the costs of false-positive, true-negative, false-negative, and true-positive results, respectively. Once m is estimated, the optimal decision threshold is the one for which sensitivity and specificity maximize the following expression: [sensitivity – m(1 – specificity)] [47].

Examining the ROC curve labeled X in Figure 2, we see that the slope is very steep in the lower left where both the sensitivity and FPR are low, and is close to zero at the upper right where the sensitivity and FPR are high. The slope takes on a high value when the patient is unlikely to have the disease or the cost of a false-positive is large; for these situations, a low FPR is optimal. The slope takes on a value near zero when the patient is likely to have the disease or treatment for the disease is beneficial and carries little risk to healthy patients; in these situations, a high sensitivity is optimal [3]. A nice example of a study using this equation is given in [48]. See also work by Greenhouse and Mantel [49] and Linnet [50] for determining the optimal decision threshold when a desired level for the sensitivity, specificity, or both is specified a priori.

Conclusion

Applications of ROC curves in the medical literature have increased greatly in the past few decades, and with this expansion many new statistical methods of ROC analysis have been developed. These include methods that correct for common biases like verification bias and imperfect gold standard bias, methods for combining the information from multiple diagnostic tests (i.e., optimal combinations of tests) and multiple studies (i.e., meta-analysis), and methods for analyzing clustered data (i.e., multiple observations from the same patient). Interested readers can search directly for these statistical methods or consult two recently published books on ROC curve analysis and related topics [2, 39]. Available software for ROC analysis allows investigators to easily fit, evaluate, and compare ROC curves [41, 51], although users should be cautious about the validity of the software and check the underlying methods and assumptions.

Acknowledgments

I thank the two series' coeditors and an outside statistician for their helpful comments on an earlier draft of this manuscript.

References

1. Weinstein S, Obuchowski NA, Lieber ML. Clinical evaluation of diagnostic tests. *AJR* 2005;184:14–19
2. Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine.* New York, NY: Wiley-Interscience, 2002
3. Metz CE. Some practical issues of experimental design and data analysis in radiologic ROC studies. *Invest Radiol* 1989;24:234–245
4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36
5. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989;9:190–195
6. Hilden J. The area under the ROC curve and its competitors. *Med Decis Making* 1991;11:95–101
7. Hilden J. Prevalence-free utility-respecting summary indices of diagnostic power do not exist. *Stat Med* 2000;19:431–440
8. Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. *Acad Radiol* 2001;8:328–334
9. Beam CA, Baker ME, Paine SS, Sostman HD, Sullivan DC. Answering unanswered questions: proposal for a shared resource in clinical diagnostic radiology research. *Radiology* 1992;183:619–620
10. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader multicase receiver operating characteristic analysis: an empirical comparison of five methods. *Acad Radiol* 2004;11:980–995
11. Obuchowski NA. Multi-reader ROC studies: a comparison of study designs. *Acad Radiol* 1995;2:709–716
12. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. *Acad Radiol* 1997;4:587–600
13. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest Radiol* 1992;27:723–731
14. Obuchowski NA. Multi-reader multi-modality ROC studies: hypothesis testing and sample size estimation using an ANOVA approach with dependent observations. with rejoinder. *Acad Radiol* 1995;2:S22–S29
15. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Stat Med* 1996;15:1807–1826
16. Song HH. Analysis of correlated ROC areas in diagnostic testing. *Biometrics* 1997;53:370–382
17. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Acad Radiol* 2000;7:341–349
18. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structure across modalities. *Acad Radiol* 2001;8:605–615
19. Beiden SV, Wagner RF, Campbell G, Chan HP. Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis. *Acad Radiol* 2001;8:616–622
20. Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Can J Stat* 2000;28:731–750
21. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. *Arch Intern Med* 1996;156:209–213
22. Kim MJ, Lim JS, Oh YT, et al. Preoperative MRI of rectal cancer with and without rectal water filling: an intraindividual comparison. *AJR* 2004;182:1469–1476
23. Osada H, Kaku K, Masuda K, Iitsuka Y, Seki K, Sekiya S. Quantitative and qualitative evaluations of fetal lung with MR imaging. *Radiology* 2004;231:887–892
24. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics* 2000;1:123–140

25. Zheng B, Leader JK, Abrams G, et al. Computer-aided detection schemes: the effect of limiting the number of cued regions in each case. *AJR* 2004;182:579–583
26. Chakraborty DP, Winter LHL. Free-response methodology: alternative analysis and a new observer-performance experiment. *Radiology* 1990;174:873–881
27. Chakraborty DP. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med Phys* 1989;16:561–568
28. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996;23:1709–1725
29. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. *Acad Radiol* 2000;7:516–525
30. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory: a direct solution. *Psychometrika* 1968;33:117–124
31. Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal detection theory and determination of confidence intervals: rating method data. *J Math Psychol* 1969;6:487–496
32. ROCKIT and LABMRMC. Available at: xray.bsd.uchicago.edu/krl/KRL_ROC software_index.htm. Accessed December 13, 2004
33. Swets JA. Empirical RO. Cs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychol Bull* 1986;99:181–198
34. Hanley JA. The robustness of the binormal assumption used in fitting ROC curves. *Med Decis Making* 1988;8:197–203
35. Goddard MJ, Hinberg I. Receiver operating characteristic (ROC) curves and non-normal data: an empirical study. *Stat Med* 1990;9:325–337
36. Zou KH, Tempany CM, Fielding JR, Silverman SG. Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. *Acad Radiol* 1998;5:680–687
37. Pepe MS. A regression modeling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika* 1997;84:595–608
38. Pepe MS. An interpretation for the ROC curve using GLM procedures. *Biometrics* 2000;56:352–359
39. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Oxford University Press, 2003
40. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–844
41. ROC analysis. Available at: www.bio.ri.ccf.org/Research/ROC/index.html. Accessed December 13, 2004
42. OBUMRM. Available at: www.bio.ri.ccf.org/OBUMRM/OBUMRM.html. Accessed December 13, 2004
43. The University of Iowa Department of Radiology: The Medical Image Perception Laboratory. MRMC 2.0. Available at: perception.radiology.uiowa.edu. Accessed December 13, 2004
44. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. *Acad Radiol* (in press)
45. Obuchowski NA. Sample size tables for receiver operating characteristic studies. *AJR* 2000;175:603–608
46. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283–298
47. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561–577
48. Somoza E, Mossman D. "Biological markers" and psychiatric diagnosis: risk-benefit balancing using ROC analysis. *Biol Psychiatry* 1991;29:811–826
49. Greenhouse SW, Mantel N. The evaluation of diagnostic tests. *Biometrics* 1950;6:399–412
50. Linnet K. Comparison of quantitative diagnostic tests: type I error, power, and sample size. *Stat Med* 1987;6:147–158
51. Stephan C, Wesseling S, Schink T, Jung K. Comparison of eight computer programs for receiver-operating characteristic analysis. *Clin Chem* 2003;49:433–439
52. Ma G, Hall WJ. Confidence bands for receiver operating characteristic curves. *Med Decis Making* 1993;13:191–197

APPENDIX 1. Area Under the Curve and Confidence Intervals with Binormal Model

Under the binormal assumption, the receiver operating characteristic (ROC) curve is the collection of points given by

$$[1 - \phi(c),\ 1 - \phi(B \times c - A)]$$

where c ranges from $-\infty$ to $+\infty$ and represents all the possible values of the underlying binormal distribution, and ϕ is the cumulative normal distribution evaluated at c. For example, for a false-positive rate of 0.10, $\phi(c)$ is set equal to 0.90; from tables of the cumulative normal distribution, we have $\phi(1.28) = 0.90$. Suppose $A = 2.0$ and $B = 1.0$; then the sensitivity $= 1 - \phi(-0.72) = 1 - 0.2358 = 0.7642$.

ROCKIT [32] gives a confidence interval (CI) for sensitivity at particular false-positive rates (i.e., pointwise CIs). A CI for the entire ROC curve (i.e., simultaneous CI) is described by Ma and Hall [52].

Under the binormal distribution assumption, the area under the smooth ROC curve (AUC) is given by

$$\text{AUC} = \phi[A / \sqrt{(1 + B^2)}].$$

For the example above, $\text{AUC} = \phi[2.0 / \sqrt{(2.0)}] = \phi[1.414] = 0.921$.

The variance of the full area under the ROC curve is given as standard output in programs like ROCKIT [32]. An estimator for the variance of the partial area under the curve (PAUC) was given by McClish [5]; a Fortran program is available for estimating the PAUC and its variance [41].

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

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005