Yulei Jiang, MS • Charles E. Metz, PhD • Robert M. Nishikawa, PhD

A Receiver Operating Characteristic Partial Area Index for Highly Sensitive Diagnostic Tests[1]

PURPOSE: **Area under a receiver operating characteristic (ROC) curve (A_z) is widely used as an index of diagnostic performance. However, A_z is not a meaningful summary of clinical diagnostic performance when high sensitivity must be maintained clinically. The authors developed a new ROC partial area index, which measures clinical diagnostic performance more meaningfully in such situations, to summarize an ROC curve in only a high-sensitivity region.**

MATERIALS AND METHODS: **The mathematical formulation of the partial area index was derived from the conventional binormal model. Statistical tests of apparent differences in this index were formulated analogous to that of A_z. One common statistical test involving the partial area index was validated by computer simulations under realistic conditions.**

RESULTS: **An example in mammography illustrates a situation in which the partial area index is more meaningful than A_z in measuring clinical diagnostic performance.**

CONCLUSION: **The partial area index can be used as a more meaningful alternative to the conventional A_z index for highly sensitive diagnostic tests.**

RECEIVER operating characteristic (ROC) analysis is widely used to evaluate diagnostic performance (1,2). An ROC curve provides a concise description of trade-offs available between sensitivity and specificity—the two related but distinct aspects of diagnostic performance. The area under an ROC curve, denoted A_z when the ROC curve is fitted with the conventional binormal model (1,2), is often used to summarize the diagnostic performance described by an entire ROC curve (3). The value of the A_z index can be interpreted as the average value of sensitivity over all possible values of specificity (between 0 and 1) or, alternatively, as the average value of specificity over all possible values of sensitivity (between 0 and 1) (2).

The A_z index, however, may not be an entirely relevant measure of diagnostic performance in some situations. The clinical tasks of some diagnostic tests demand high sensitivity. For these tests, only those operating points on an ROC curve that have high sensitivity values are clinically acceptable; other operating points with lower sensitivity values are clinically irrelevant because, although they can be estimated in an experiment and plotted on an ROC curve, they will not be used clinically. Therefore, the A_z index, which summarizes an entire ROC curve by giving equal weight to operating points at all sensitivity levels, does not measure diagnostic performance meaningfully from a clinical perspective in such situations.

One example of a clinical task that demands high sensitivity is screening mammography, which can help reduce the mortality of breast cancer by depicting asymptomatic and early-stage breast cancers that can be treated effectively. To achieve this benefit, however, sensitivity must be high in clinical practice because women with false-negative findings at mammography cannot benefit from timely treatment of the cancer (4). In actual clinical practice, high sensitivity has been demonstrated in large mammographic series (5,6).

Sensitivity is also crucial in the diagnostic work-up prompted by abnormal findings at screening mammography; in this situation, the cost of false-negative findings far exceeds that of false-positive findings. The diagnostic work-up may include acquisition of additional mammograms and performance of ultrasonography; findings may lead to fine-needle aspiration biopsy and cytologic examination, needle core biopsy, or surgical biopsy. Except for surgical biopsy, these diagnostic tests are performed

Index terms: Diagnostic radiology, observer performance • Receiver operating characteristic curve (ROC) • Statistical analysis

Abbreviation: ROC = receiver operating characteristic.

Radiology 1996; 201:745–750

[1] From the Department of Radiology, Kurt Rossmann Laboratories for Radiologic Image Research, the University of Chicago, 5841 S Maryland Ave, MC2026, Chicago, IL 60637. From the 1995 RSNA scientific assembly. Received April 29, 1996; revision requested June 18; revision received July 1; accepted July 8. Supported in part by grant no. DE FG02-94ER61816 from the U. S. Department of Energy and grant no. RO1 CA 60187 from the National Institutes of Health. **Address reprint requests to** Y.J.

See also the editorial by Dwyer (pp 621–625) in this issue.

primarily to help reduce cost and patient morbidity by obviating surgical biopsy of breast lesions that are likely benign. Therefore, the sensitivity of each diagnostic test is, in fact, an integral component of the sensitivity of screening mammography. Consequently, these diagnostic tests must maintain high sensitivity to meet the objective of screening mammography for the detection of breast cancer. As a result, the positive biopsy yield after mammography is typically 30% or less (7).

In screening mammography and the subsequent diagnostic tests, the A_z index does not measure diagnostic performance meaningfully in a clinical sense. For these tests, the only operating points that are clinically relevant are those at high levels of sensitivity, which appear on the upper portion of an ROC curve rather than on the entire ROC curve. One alternative for measuring clinical performance of these tests is to compare a pair of operating points on two ROC curves (8). To use this approach, however, criteria must be established a priori for choosing a single clinically important operating point on an ROC curve. In practice, such criteria may be difficult to establish.

We propose a new partial area index for ROC analysis that can be used to measure diagnostic performance more meaningfully and more accurately in a clinical sense. This new partial area index summarizes an ROC curve specifically in a clinically relevant high-sensitivity region, and it is conceptually similar to the partial area described in detail by McClish (9) that summarizes a portion of an ROC curve within a band of false-positive fractions. The latter approach, however, is not directly relevant in the evaluation of screening mammography and subsequent diagnostic tests, when the emphasis is on sensitivity rather than on false-positive fraction (ie, specificity). We describe the formulation of this new partial area index and the formulation of statistical tests that involve this index. We also present an example in mammography to illustrate a clinical situation in which the partial area index measured diagnostic performance more meaningfully than did the A_z index.

MATERIALS AND METHODS

Definition

We define our partial area index, ${}_{TPF_0}A_z'$, as the area that lies under an ROC curve but above a preselected sensitivity threshold (TPF_o) in a conventional ROC graph, divided by the constant $(1 - TPF_o)$:

$$ {}_{TPF_o}A_Z' \equiv \frac{\int_{TPF_o}^{1} [1 - FPF(TPF)]dTPF}{1 - TPF_o}, \quad (1) $$

where FPF = false-positive fraction and TPF = true-positive fraction. The sensitivity threshold, TPF_o, is a free parameter to be chosen in the clinical context of the diagnostic test of interest. This parameter represents the minimum acceptable sensitivity of the diagnostic test in a particular situation. Notice that the partial area index becomes identical to the conventional A_z index if $TPF_o = 0$.

This partial area index has properties similar to those of the conventional A_z index. The partial area index can be interpreted as the average value of specificity over all values of sensitivity between TPF_o and 1. The numerical value of the partial area index is bounded by 1, for a perfect ROC curve, and by $(1 - TPF_o)/2$, for chance performance (the positive diagonal in a conventional ROC graph, where sensitivity is equal to false-positive fraction); that is,

$$ \frac{1}{2}(1 - TPF_o) \le {}_{TPF_o}A_Z' \le 1. \quad (2) $$

The partial area index defined in Equation (1) represents the partial area under an ROC curve of interest relative to the same partial area under a perfect ROC curve. This normalization offers several advantages. First, it allows convenient interpretation of the partial area index as an average specificity of the diagnostic test when the test is operated to provide the range of clinically relevant high sensitivity values. Second, for reasonably high ROC curves, this normalization expresses the values of the partial area index on a numerical scale similar to that of the A_z index. For example, an ROC curve may have an A_z value of 0.92 and a ${}_{0.90}A_z'$ nonnormalized equivalent of the partial area would equal 0.082. Unlike the A_z index, however, an ROC curve of close to chance performance can have a partial area index of less than 0.5. Third, for reasonably high ROC curves, the choice of TPF_o has a rather weak influence on the numerical values of the normalized partial area index. For example, an ROC curve may have a ${}_{0.90}A_z'$ value of 0.82, and a ${}_{0.75}A_z'$ value of 0.85, whereas the corresponding nonnormalized partial area values are 0.082 and 0.21, respectively.

Computation of the Partial Area Index

An alternative expression for the partial area index can be obtained by changing the variable of integration in Equation (1) from TPF to the corresponding critical value of a latent decision variable, x. Therefore,

$$ {}_{TPF_o}A_Z' = 1 + \frac{1}{1 - TPF_o}\int_{-\infty}^{x(TPF_o)} \times FPF(x)\left|\frac{dTPF(x)}{dx}\right|dx. \quad (3) $$

To derive the relationship between the partial area index and a parametric description of the ROC curve, a particular mathematical model must be adopted. We used the conventional binormal model, which assumes that the actually negative (eg, healthy) population has a standard-normal latent decision-variable distribution with zero mean and unit standard deviation, and that the actually positive (eg, diseased) population has a normal distribution with mean of a/b and standard deviation of $1/b$. This model produces reasonably good fits to a wide variety of ROC curves (10,11). The two parameters in this model, a and b, can be determined from experimental data by means of maximum-likelihood estimation (1,12,13).

The binormal model can be expressed mathematically as

$$ FPF(x_c) = \Phi(-x_c) \quad (4a) $$

$$ TPF(x_c) = \Phi(a - bx_c), \quad (4b) $$

where

$$ \Phi(x) \equiv \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt \quad (4c) $$

and where x_c represents the critical value of x that distinguishes positive and negative test results. By substitution of Equation (4), Equation (3) can be further simplified to an integral function of a, b, and the constant, TPF_o:

$$ {}_{TPF_o}A_Z' = 1 - \frac{b}{\sqrt{2\pi}(1 - TPF_o)}\int_{x=-\infty}^{x(TPF_o)} \times \Phi(x)e^{-(bx-a)^2/2}dx. \quad (5) $$

Since TPF_o is a user-defined constant, and since a and b are calculated by computer programs that fit ROC curves to confidence-rating data by means of maximum-likelihood estimation (1,12,13), the partial area index can be computed by means of numerical integration.

Computation of Variance

Since the partial area index is a function of the binormal parameters a and b and of the constant TPF_o, first-order approximation (the "delta method") shows that

$$ \mathrm{Var}\{{}_{TPF_o}A_Z'(a, b)\} \cong \left(\frac{\partial {}_{TPF_o}A_Z'}{\partial a}\right)^2 \mathrm{Var}\{a\} + \left(\frac{\partial {}_{TPF_o}A_Z'}{\partial b}\right)^2 \mathrm{Var}\{b\} + 2\left(\frac{\partial {}_{TPF_o}A_Z'}{\partial a}\right)\left(\frac{\partial {}_{TPF_o}A_Z'}{\partial b}\right)\mathrm{Cov}\{a, b\}. \quad (6) $$

The derivatives in Equation (6) are given by

$$ \frac{\partial {}_{TPF_o}A_Z'}{\partial a} = \frac{e^{-a^2/2(1+b^2)}}{(1 - TPF_o)\sqrt{2\pi(1 + b^2)}} \times [1 - \Phi(\lambda)] \quad (7a) $$

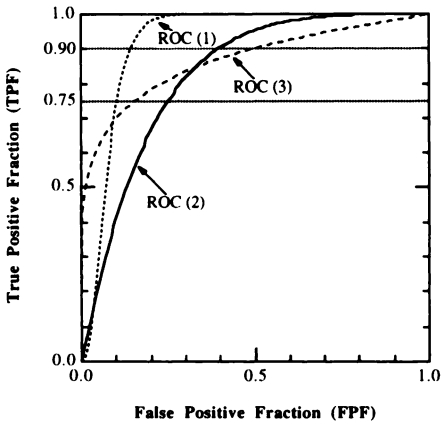


Figure 1. Three empirical ROC curves used in the computer simulation. The values of the A_z index were 0.917, 0.825, and 0.862. The values of the partial area indexes were 0.817, 0.484, and 0.261, respectively, for $TPF_o = 0.90$; the corresponding values for $TPF_o = 0.75$ were 0.852, 0.606, and 0.523, respectively.

and

$$\frac{\partial {}_{TPF_o}A'_Z}{\partial b} = \frac{e^{-a^2/2(1+b^2)-\lambda^2/2}}{2\pi(1+b^2)(1-TPF_o)} - \frac{ab\,e^{-a^2/2(1+b^2)}}{\sqrt{2\pi}(1+b^2)^{3/2}(1-TPF_o)} \times [1-\Phi(\lambda)], \quad (7b)$$

where

$$\lambda = \frac{\sqrt{1+b^2}}{b}\Phi^{-1}(TPF_o) - \frac{a}{b\sqrt{1+b^2}}. \quad (7c)$$

Since computer programs that fit ROC curves by means of maximum-likelihood estimation also calculate the variance and covariance of the curve parameter estimates (1,12,13), the numerical value of the variance of the partial area index can be calculated from Equations (6) and (7).

Statistical Tests

Because the partial area index is conceptually similar to the conventional A_z index, statistical tests for the significance of differences in values of the partial area index are entirely analogous to statistical tests involving the A_z index (14).

For example, to test the statistical significance of any apparent differences in the partial area index between two ROC curves, the test statistic

$$Z = \frac{{}_{TPF_o}A'_{Z_1} - {}_{TPF_o}A'_{Z_2}}{\sqrt{\mathrm{Var}\{{}_{TPF_o}A'_{Z_1} - {}_{TPF_o}A'_{Z_2}\}}} \quad (8)$$

is computed, which, for sufficiently large case samples, may be expected to follow a standard-normal distribution when the null hypothesis (that ${}_{TPF_o}A'_{z1} = {}_{TPF_o}A'_{z2}$) is true. The variance of the difference between the partial area index estimates is given by

$$\mathrm{Var}\{{}_{TPF_o}A'_{Z_1} - {}_{TPF_o}A'_{Z_2}\} = \mathrm{Var}\{{}_{TPF_o}A'_{Z_1}\} + \mathrm{Var}\{{}_{TPF_o}A'_{Z_2}\} - 2\mathrm{Cov}\{{}_{TPF_o}A'_{Z_1}, {}_{TPF_o}A'_{Z_2}\}. \quad (9)$$

If the two ROC curves are estimated from independent samples (eg, radiographs obtained in different patients), the covariance term in Equation (9) is equal to zero. If the two ROC curve estimates are correlated (eg, if they are obtained from radiographs obtained in the same group of patients), however, the covariance term is nonzero. In this latter case, the first-order approximation of the covariance term is given by

$$\mathrm{Cov}\{{}_{TPF_o}A'_{Z_1}, {}_{TPF_o}A'_{Z_2}\} \cong \left(\frac{\partial {}_{TPF_o}A'_{Z_1}}{\partial a_1}\right) \cdot \left(\frac{\partial {}_{TPF_o}A'_{Z_2}}{\partial a_2}\right)\mathrm{Cov}\{a_1, a_2\} + \left(\frac{\partial {}_{TPF_o}A'_{Z_1}}{\partial a_1}\right)\left(\frac{\partial {}_{TPF_o}A'_{Z_2}}{\partial b_2}\right)\mathrm{Cov}\{a_1, b_2\} + \left(\frac{\partial {}_{TPF_o}A'_{Z_1}}{\partial b_1}\right)\left(\frac{\partial {}_{TPF_o}A'_{Z_2}}{\partial a_2}\right)\mathrm{Cov}\{b_1, a_2\} + \left(\frac{\partial {}_{TPF_o}A'_{Z_1}}{\partial b_1}\right)\left(\frac{\partial {}_{TPF_o}A'_{Z_2}}{\partial b_2}\right)\mathrm{Cov}\{b_1, b_2\}. \quad (10)$$

Equation (10) can be evaluated by using the values of the covariances of the bivariate binormal parameters that are calculated by computer programs that estimate correlated ROC curves from paired data (13,15,16).

Validation of a Statistical Test

A computer simulation study was conducted to check the validity of the test results of statistical significance of apparent differences in the partial area index between two ROC curves that are estimated from a single sample of patients. In our simulation study, two sets of continuously distributed confidence rating data, which were correlated with correlation coefficients of .4560 for normal cases and of .6894 for abnormal cases, were drawn randomly from a bivariate binormal distribution (15). These rating data were used as input to a modified version of the CLABROC program (a version of the CORROC algorithm [15] that has been modified to analyze continuously distributed data [17 [Metz CE, the University of Chicago, Chicago, Ill]]), which estimates two ROC curves from correlated continuously distributed data. In addition to its usual computations, the modified CLABROC program also estimates the partial area index values for the two ROC curves and uses Equations (8)–(10) to test the statistical significance of any apparent differences between the two partial area index values. This procedure of estimating binormal ROC curves on the basis of simulated confidence rating data with fixed correlation and testing for statistical significance of the differences in the partial area index was repeated for 5,000 data sets sampled independently from the same population. In our simulation, the null hypothesis was, in fact, true, because the population values of the a and b values were equal for each of the marginal distributions from which the correlated pairs of confidence rating data were drawn. Consequently, the 5,000 two-tailed P values we obtained allowed comparison of the expected type I error rate (the critical two-tailed P value α [ie, the expected probability of incorrect rejection of the null hypothesis]) with the empirical type I error rate found in the simulation (calculated by thresholding the cumulative histogram of the P values at each value of α).

We studied a variety of parameter values in our simulation. Three empirical ROC curves with a broad range of performance level were employed, with the following bivariate binormal parameters: $a = 4.7017$ and $b = 3.2410$ for ROC 1; $a = 1.6857$ and $b = 1.5049$ for ROC 2; and $a = 1.2766$ and $b = 0.6061$ for ROC 3. Two values of TPF_o, 0.90 and 0.75, were investigated. The corresponding values of the A_z index and of the partial area index are listed in the caption for Figure 1, where the three ROC curves are plotted. Confidence rating data that represented 50 normal and 50 abnormal cases were drawn from the bivariate binormal distribution for each pair of ROC curve estimates. These numbers of confidence rating data are typical in radiologic imaging studies. Additional simulations were performed in which 500 normal and 500 abnormal confidence ratings were drawn for each pair of ROC curve estimates; the purpose of these larger simulated data sets was to verify that the empirical type I error rate approaches the expected type I error rate as the case sample size increases.

RESULTS

Simulation Results

Results of the simulation study for samples of 50 normal and 50 abnormal cases are shown in Figure 2a . The observed type I error rates tended to be slightly higher than the expected type I error rates. The observed type I error rate for ROC 3 differed from an ideal test the most when TPF_o was 0.90. In that situation, at an ideal type I error rate of 0.05, the observed type I error rate was 0.065, whereas the upper limit of the 95% confidence interval of an ideal statistical test was 0.056.

Figure 2b shows the results of the simulation study for samples of 500 normal and 500 abnormal cases. In this situation, most data points lay within the 95% confidence band of an ideal test. This indicates that the statistical test is more accurate when a large case sample is used to estimate

the ROC curves. In radiologic research, however, the available case samples are often small.

We then investigated a nonlinear transformation of the partial area index that can be used to replace the partial area index in calculating the test statistic prescribed by Equation (8). This transformation of the partial area index, which is equivalent to the Fischer "r-to-Z" transformation (18,19), is given by

$$\theta = \frac{1}{2}\ln\left[\frac{1 + {}_{TPF_o}A'_Z}{1 - {}_{TPF_o}A'_Z}\right]. \quad (11)$$

Under this transformation, the test statistic becomes

$$Z = \frac{\theta_1 - \theta_2}{\sqrt{\mathrm{Var}\{\theta_1 - \theta_2\}}} = \frac{\theta_1 - \theta_2}{\sqrt{\mathrm{Var}\{\theta_1\} + \mathrm{Var}\{\theta_2\} - 2\mathrm{Cov}\{\theta_1, \theta_2\}}}, \quad (12)$$

with the variance and covariance given by

$$\mathrm{Var}\{\theta\} \cong \frac{1}{[1 - ({}_{TPF_o}A'_Z)^2]^2} \cdot \mathrm{Var}\{{}_{TPF_o}A'_Z\} \quad (13)$$

and

$$\mathrm{Cov}\{\theta_1, \theta_2\} \cong 1/\{1 - ({}_{TPF_o}A'_{Z_1})^2] \cdot [1 - ({}_{TPF_o}A'_{Z_2})^2\} \cdot \mathrm{Cov}\{{}_{TPF_o}A'_{Z_1}, {}_{TPF_o}A'_{Z_2}\}. \quad (14)$$

Figure 3a shows the simulation results of the statistical test on the transformed partial area index for a case sample of 50 normal and 50 abnormal cases. Compared with Figure 2, most data points in Figure 3a appear to be within or lower than the 95% confidence band of an ideal test, which indicates that in some situations this test produces conservative estimates. At an expected type I error rate of 0.05, the observed type I error rate that differed the most from the result of an ideal test (ROC 2 with TPF_o = 0.90) was 0.058, whereas the upper bound of the 95% confidence interval of an ideal test was 0.056. Figure 3b shows the simulation results of the same test on the transformed partial area index for a case sample of 500 normal and 500 abnormal cases. Similar to the results obtained on the original partial area index for a large case sample (Fig 2b), most of the results in Figure 3b lie within the 95% confidence band of an ideal test.

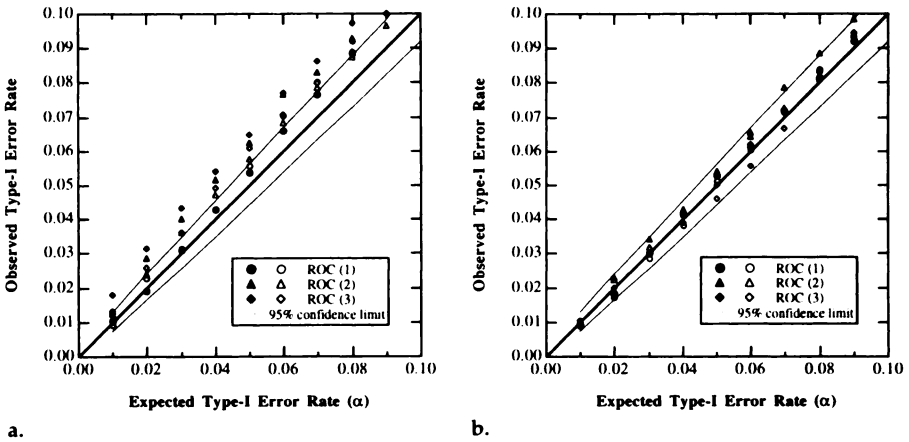


Figure 2. Results from 5,000 simulated experiments that tested the statistical significance of apparent differences in the partial area index, ${}_{TPF_o}A'_Z$, between two ROC curves that were estimated from correlated confidence-rating data. The sensitivity thresholds in the partial area index, TPF_o, were 0.90 (solid symbols) and 0.75 (open symbols). **(a)** The two ROC curves in each simulated experiment were estimated from a single case sample of 50 normal and 50 abnormal cases. **(b)** The two ROC curves in each simulated experiment were estimated from a single case sample of 500 normal and 500 abnormal cases.

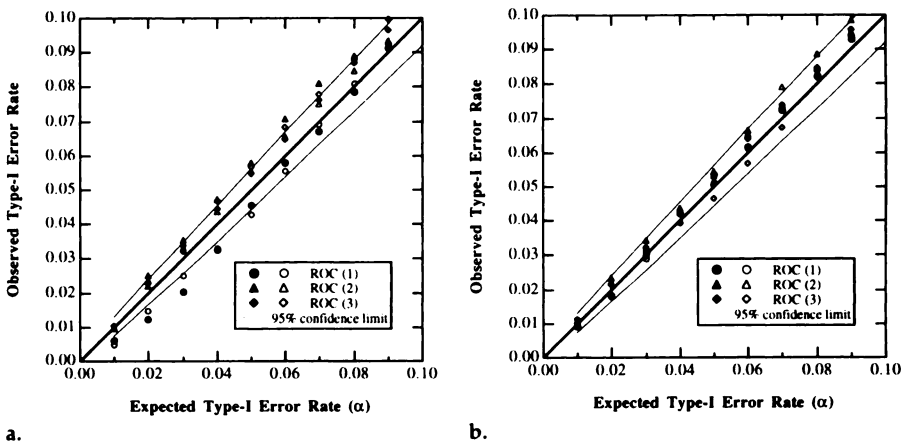
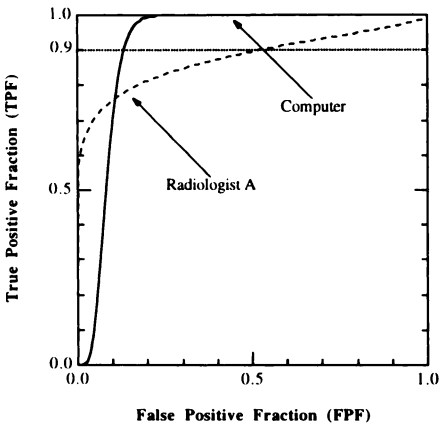


Figure 3. Results from 5,000 simulated experiments that tested the statistical significance of apparent differences in the transformed partial area index described by Equation (11) between two ROC curves that were estimated from correlated confidence-rating data. The sensitivity thresholds in the partial area index, TPF_o, were 0.90 (solid symbols) and 0.75 (open symbols). **(a)** The two ROC curves in each simulated experiment were estimated from a single case sample of 50 normal and 50 abnormal cases. **(b)** The two ROC curves in each simulated experiment were estimated from a single case sample of 500 normal and 500 abnormal cases.

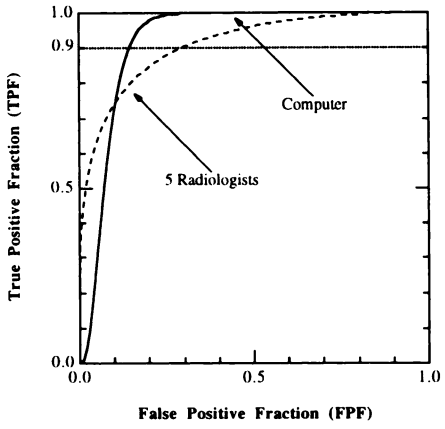
An Example

We describe an example to illustrate a situation in which the partial area index can be more meaningful than the A_z index in measuring diagnostic performance from a clinical perspective. We used the partial area index to analyze data we obtained in an experiment to compare the performance of a computer-aided diagnostic scheme with the performance of a group of five radiologists. The diagnostic task studied in this experiment was to differentiate between malignant and benign clustered microcalcifications in mammograms. The details of this experiment are described elsewhere (20).

ROC analysis was used in this experiment to measure diagnostic performance. An ROC curve was estimated for the performance of the computerized classification scheme from its calculated estimates of likelihood of malignancy, which were on a continuous scale of values between 0 and 1. An ROC curve was also estimated for each of the five radiologists, who read the same mammograms and estimated the likelihood of malignancy on a numerical scale from 0 to 100 (20). Unless otherwise stated, all ROC curves were obtained with the CLABROC algorithm developed by Metz et al (17). In addition, an ROC curve that represents the combined performance of the five radiologists was obtained by averaging the binormal parameters, *a* and *b*, of the ROC curve for each radiologist.



4.



5.

Figures 4, 5. **(4)** ROC curves of the computerized classification scheme and of radiologist A. The values of the A_z index and the partial area index $_{0.90}A'_z$ for these ROC curves are listed in Table 1. **(5)** ROC curves of the computerized classification scheme and of the five radiologists combined. The values of the A_z index and the partial area index $_{0.90}A'_z$ for these ROC curves are listed in Table 2.

Table 1
Individual Comparison of the A_Z Index and the Partial Area Index $_{0.90}A'_Z$ for the Computerized Classification Scheme and Five Radiologists

	A_Z			$_{0.90}A'_Z$		
Radiologist	Computer	Radiologist	*P* Value	Computer	Radiologist	*P* Value
A	0.91	0.87	.55	0.84	0.20	.002
B	0.91	0.84	.23	0.84	0.29	<.001
C	0.91	0.85	.41	0.83	0.10	<.001
D	0.91	0.89	.65	0.83	0.63	.09
E	0.91	0.96	.21	0.83	0.75	.58

Note.—The two-tailed *P* values were calculated by using the modified CLABROC algorithm.

Table 2
Comparison of the A_Z Index and the Partial Area Index $_{0.90}A'_Z$ for the Computerized Classification Scheme and the Five Radiologists as a Group

Reader	A_Z	$_{0.90}A'_Z$
Computer	0.92	0.82
Five radiologists combined	0.89	0.42
P value	0.21	0.03

Note.—The two-tailed *P* values were calculated with the Student *t* test for paired data. Values for the computer are slightly different from their corresponding values listed in Table 1 because the LABROC algorithm was used to estimate the computer's ROC curve in this comparison.

In this experiment, we compared the performance of the computerized classification scheme with the performance of individual radiologists. Figure 4 shows the ROC curves of the computerized classification scheme and of radiologist A. Because the same set of mammograms was used to obtain both ROC curves in this comparison, the CLABROC algorithm developed by Metz et al (15,17) was used. Table 1 lists the values of the A_z index and the partial area index as well as the statistical significance levels (*P* values) of the differences in these two indexes. The sensitivity threshold of the partial area index, TPF_o, was chosen to be 0.90 in this comparison because a high sensitivity must be maintained clinically when differentiating malignant from benign clustered microcalcifications, since the cost of delayed detection of a breast cancer is much higher than the cost of breast biopsy in a benign lesion. Figure 4 shows that the computer's ROC curve is substantially higher than radiologist A's ROC curve in the clinically important high-sensitivity region. Despite this substantial difference in the ROC curves, however, the difference in their A_z values was not statistically significant ($P = .55$), because although the computer's ROC curve is substantially higher than radiologist A's ROC curve at high sensitivity values, it is lower at low sensitivity values. From a clinical perspective, however, the part of the ROC curve with low sensitivity values is not relevant. Therefore, we compared the partial area index values of the two ROC curves and found that the difference in the partial area index was highly significant ($P = .002$). Table 1 also shows two similar comparisons of performance between the computerized classification scheme and radiologists B and C; highly significant differences were found in the values of the partial area index but not in the values of the A_z index.

We also compared the performance of the computerized classification scheme with the combined performance of the five radiologists by employing the Student *t* test for paired data in a way analogous to that used by Wu et al in testing differences in A_z (21). Figure 5 shows the ROC curves of the computerized classification scheme and of the five radiologists combined. Table 2 lists the values of the A_z index, the partial area index, and the results of the Student *t* tests for paired data for the statistical significance of the differences in these two indexes. As shown in Figure 5, the computer's ROC curve is clearly higher than the radiologists' ROC curve in the clinically important high-sensitivity region. However, similar to the earlier performance comparison between the computer and radiologist A (Fig 4), the difference in the values of the partial area index was statistically significant ($P = .03$), but the difference in the A_z values was not ($P = .21$).

DISCUSSION

In our example, the ROC curve of the computerized classification scheme was substantially higher in high-sensitivity regions than the ROC curves of four of the five radiologists or the summary curve for the five radiologists as a group. This superior performance of the computerized classification scheme demonstrates its potential clinical utility for improving radiologists' ability to distinguish between malignant and benign clustered microcalcifications, thereby potentially helping radiologists reduce the number of biopsies performed in benign breast lesions and increase the positive predictive value of screening mammography without compromising its overall sensitivity (20). This potentially important difference in performance was not quantified accurately with the A_z index, however, and the differences in values were not statistically significant. This is not surprising, because the A_z index measures diagnostic performance over the full range of sensitivity between 0 and 1 and does not focus on any particu-

lar portion of an ROC curve that may be relevant in clinical practice. On the other hand, clinical relevance is emphasized with the partial area index because diagnostic performance is measured in only a particular high-sensitivity region. Therefore, values of the partial area index, $_{TPF_0}A'_z$, should be more meaningful than values of the A_z index in measuring diagnostic performance when high sensitivity is clinically necessary. In our experiment, differences in the values of the partial area index, $_{0.90}A'_z$, were found to be statistically significant, whereas differences in the values of the A_z index were not.

Our example also illustrates that the partial area index can be used to compare diagnostic performance even when two ROC curves cross, as in Figures 4 and 5. In such cases, comparison of the A_z index values is practically meaningless because, in practice, all points on the curve will not have the same clinical relevance. Because the curves cross, one ROC curve is higher at high sensitivity levels whereas the other curve is higher at high specificity levels (low false-positive fractions). If two ROC curves cross at a point where the sensitivity is lower than TPF_o, the partial area index can be used to assess diagnostic performance unambiguously at high levels of sensitivity, because the partial area index then focuses on a narrow range of sensitivity within which the ROC curves do not cross. If two ROC curves cross at a point where sensitivity is higher than TPF_o, the partial area index will have the same problem as the A_z index. Since the partial area index is defined a priori for sensitivity levels of clinical interest, however, there is less ambiguity in the partial area index to the extent that the band of sensitivity levels is well chosen.

Although we used screening mammography for illustration purposes in this article, the potential usefulness of the partial area index, $_{TPF_0}A'_z$, is not limited to that application but extends to the evaluation of any diagnostic test that must maintain a high sensitivity level clinically. The partial area index, $_{TPF_0}A'_z$, is more meaningful than the conventional A_z index in such situations because it reflects the portion of the ROC curve that is clinically relevant. ■

References

1. Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York, NY: Academic, 1982.
2. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
4. Kopans DB. Mammography screening for breast cancer. Cancer 1993; 72:1809–1812.
5. Sickles EA , Ominsky SH, Sollitto RA , Galvin HB, Monticciolo DL. Medical audit of a rapid-throughput mammography screening practice: methods and results of 27,114 examinations. Radiology 1990; 175:323–327.
6. Bird RE, Wallace TW , Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992; 184:613–617.
7. Kopans DB. The positive predictive value of mammography. AJR 1992; 158:521–526.
8. Halpern EJ, Albert M, Krieger AM, Metz CE, Maidment AD. Comparison of receiver operating characteristic curves on the basis of optimal operating points. Acad Radiol 1996; 3:245–253
9. McClish DK . Analyzing a portion of the ROC curve. Med Decision Making 1989; 9:190–195
10. Swets JA. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psychol Bull 1986; 99:181–198.
11. Hanley JA. The robustness of the "binormal" assumptions used in fitting ROC curves. Med Decision Making 1988; 8:197–203.
12. Dorfman DD, Alf E Jr. Maximum likelihood estimation of parameters of signal detection theory: a direct solution. Psychometrika 1968; 33:117–124.
13. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
14. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decision Making 1984; 4:137–150.
15. Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck F, ed. Information processing in medical imaging. The Hague, The Netherlands: Nijhoff , 1984; 432–445.
16. Metz CE. Statistical analysis of ROC data in evaluating diagnostic performance. In: Herbert D, Myers R, eds . Multiple regression analysis: application in the health sciences. New York, NY: American Institute of Physics, 1986; 365–384.
17. Metz CE, Shen JH, Herman BA. New methods for estimating a binormal ROC curve from continuously distributed test results. Presented at the 1990 Joint Statistical Meeting of the American Statistical Society and the Biometric Society, Anaheim, Calif, August 1990.
18. Kendall M, Stuart A. The advanced theory of statistics. Vol 1. 4th ed. New York, NY: Macmillan, 1977; 419–420.
19. Hays WL. Statistics. 4th ed. Chicago, Ill: Holt, Rinehart, & Winston, 1988; 590.
20. Jiang Y, Nishikawa RM, Wolverton DE, et al. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 1996; 198: 671–678.
21. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology 1993; 187:81–87.