Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists

Seong Ho Park, MD[1]
Jin Mo Goo, MD[1]
Chan-Hee Jo, PhD[2]

The receiver operating characteristic (ROC) curve, which is defined as a plot of test sensitivity as the y coordinate versus its 1-specificity or false positive rate (FPR) as the x coordinate, is an effective method of evaluating the performance of diagnostic tests. The purpose of this article is to provide a nonmathematical introduction to ROC analysis. Important concepts involved in the correct use and interpretation of this analysis, such as smooth and empirical ROC curves, parametric and nonparametric methods, the area under the ROC curve and its 95% confidence interval, the sensitivity at a particular FPR, and the use of a partial area under the ROC curve are discussed. Various considerations concerning the collection of data in radiological ROC studies are briefly discussed. An introduction to the software frequently used for performing ROC analyses is also presented.

Index terms:
Diagnostic radiology
Receiver operating characteristic (ROC) curve
Software reviews
Statistical analysis

Korean J Radiol 2004;5:11-18
Received January 27, 2004; accepted after revision February 5, 2004.

[1]Department of Radiology, Seoul National University College of Medicine and Institute of Radiation Medicine, SNUMRC; [2]Biostatistics Section, Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR, U.S.A.

Address reprint requests to:
Jin Mo Goo, MD, Department of Radiology, Seoul National University Hospital, 28 Yongon-dong, Chongro-gu, Seoul 110-744, Korea.
Tel. (822) 760-2584
Fax. (822) 743-6385
e-mail: jmgoo@plaza.snu.ac.kr

The receiver operating characteristic (ROC) curve, which is defined as a plot of test sensitivity as the y coordinate versus its 1-specificity or false positive rate (FPR) as the x coordinate, is an effective method of evaluating the quality or performance of diagnostic tests, and is widely used in radiology to evaluate the performance of many radiological tests. Although one does not necessarily need to understand the complicated mathematical equations and theories of ROC analysis, understanding the key concepts of ROC analysis is a prerequisite for the correct use and interpretation of the results that it provides. This article is a nonmathematical introduction to ROC analysis for radiologists who are not mathematicians or statisticians. Important concepts are discussed along with a brief discussion of the methods of data collection to use in radiological ROC studies. An introduction to the software programs frequently used for performing ROC analyses is also presented.

What is the ROC Curve?

Sensitivity and specificity, which are defined as the number of true positive decisions/the number of actually positive cases and the number of true negative decisions/the number of actually negative cases, respectively, constitute the basic measures of performance of diagnostic tests (Table 1). When the results of a test fall into one of two obviously defined categories, such as either the presence or absence of a disease, then the test has only one pair of sensitivity and specificity values. However, in many diagnostic situations, making a decision in a binary mode is both difficult and impractical. Image findings may not be obvious or clean-cut. There may be a considerable variation in the diagnostic confidence levels between the radiologists who interpret the findings. As a result, a single pair of sensitivity and specificity values is

insufficient to describe the full range of diagnostic performance of a test.

Consider an example of 70 patients with solitary pulmonary nodules who underwent plain chest radiography to determine whether the nodules were benign or malignant (Table 2). According to the biopsy results and/or follow-up evaluations, 34 patients actually had malignancies and 36 patients had benign lesions. Chest radiographs were interpreted according to a five-point scale: 1 (definitely benign), 2 (probably benign), 3 (possibly malignant), 4 (probably malignant), and 5 (definitely malignant). In this example, one can choose from four different cutoff levels to define a positive test for malignancy on the chest radiographs: viz. $\geq$2 (i.e., the most liberal criterion), $\geq$3, $\geq$4, and 5 (i.e., the most stringent criterion). Therefore, there are four pairs of sensitivity and specificity values, one pair for each cutoff level, and the sensitivities and specificities depend on the cutoff levels that are used to define the positive and negative test results (Table 3). As the cutoff level decreases, the sensitivity increases while the specificity decreases, and vice versa.

To deal with these multiple pairs of sensitivity and specificity values, one can draw a graph using the sensitivities as the y coordinates and the 1-specificities or FPRs as the x coordinates (Fig. 1A). Each discrete point on the graph, called an operating point, is generated by using different cutoff levels for a positive test result. An ROC curve can be estimated from these discrete points, by making the assumption that the test results, or some unknown monotonic transformation thereof, follow a certain distribution. For this purpose, the assumption of a binormal distribution (i.e., two Gaussian distributions: one for the test results of those patients with benign solitary pulmonary nodules and the other for the test results of those patients with malignant solitary pulmonary nodules) is most commonly made (1, 2). The resulting curve is called the fitted or smooth ROC curve (Fig. 1B) (1). The estimation of the smooth ROC curve based on a binormal distribution uses a statistical method called maximum likelihood estimation (MLE) (3). When a binormal distribution is used, the shape of the smooth ROC curve is entirely determined by two parameters. The first one, which is referred to as *a*, is the standardized difference in the means of the distributions of the test results for those subjects with and without the condition (Appendix) (2, 4). The other parameter, which is referred to as *b*, is the ratio of the standard deviations of the distributions of the test results for those subjects without versus those with the condition (Appendix) (2, 4). Another way to construct an ROC curve is to connect all the points obtained at all the possible cutoff levels. In the previous example, there are four pairs of FPR and sensitivity values (Table 3), and the two endpoints on the ROC curve are 0, 0 and 1, 1 with each pair of values

Table 1. The Decision Matrix. Sensitivity and Specificity of a Test are Defined as TP/D+ and TN/D −, Respectively

	True Condition Status		
Test Result	Positive	Negative	Total
Positive	TP	FP	T+
Negative	FN	TN	T−
Total	D+	D−	

Note.—TP: true positive = test positive in actually positive cases, FP: false positive = test positive in actually negative cases, FN: false negative = test negative in actually positive cases, TN: true negative = test negative in actually negative cases

Table 3. Sensitivity, Specificity, and FPR for the Diagnosis of Malignant Solitary Pulmonary Nodules at Each Cutoff Level from the Plain Chest Radiography Study

Test Positive If Greater Than or Equal To	Sensitivity	Specificity	FPR
2: Probably benign	0.912 (31/34)	0.222 (8/36)	0.778
3: Possibly malignant	0.794 (27/34)	0.528 (19/36)	0.472
4: Probably malignant	0.676 (23/34)	0.750 (27/36)	0.250
5: Definitely malignant	0.206 (7/34)	0.944 (34/36)	0.056

Note.—These data are obtained from the results in Table 2. FPR is 1-specificity.

Table 2. Results from Plain Chest Radiography of 70 Patients with Solitary Pulmonary Nodules

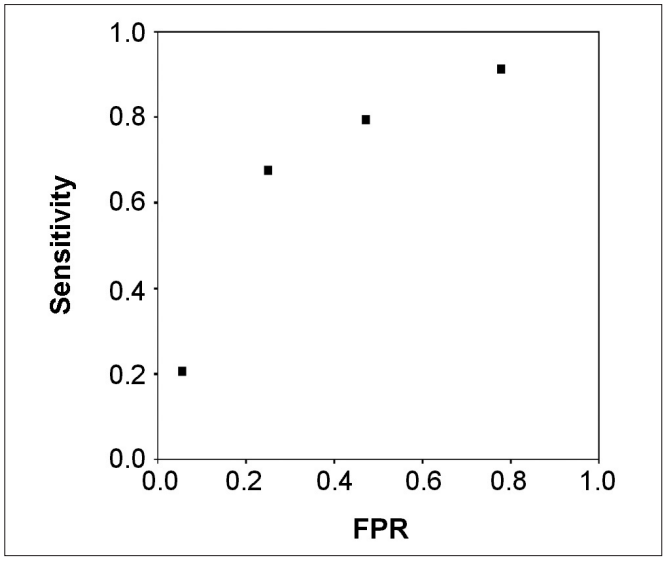
Reference Standard Result	Radiologist's Interpretation					
	Definitely Benign	Probably Benign	Possibly Malignant	Probably Malignant	Definitely Malignant	Total
Benign	8	11	8	7	2	36
Malignant	3	4	4	16	7	34
Total	11	15	12	23	9	70

Note.—Data are numbers of patients with the given result in a fictitious study of plain chest radiography in which 34 patients had malignancies and 36 had benign lesions.

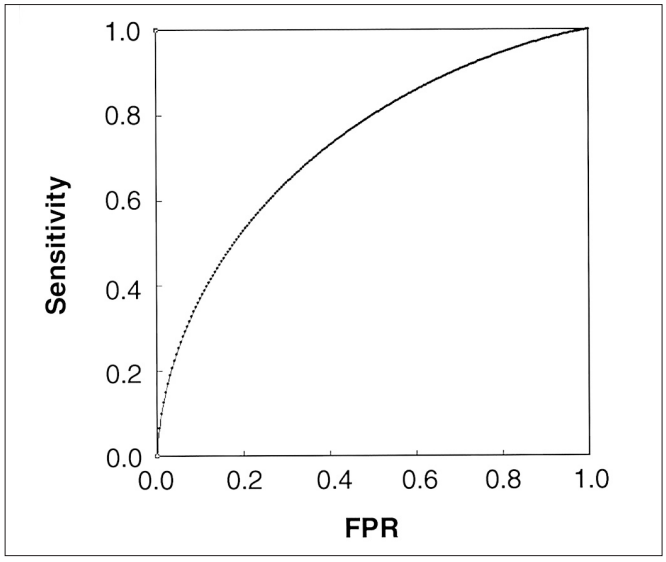
corresponding to the FPR and sensitivity, respectively. The resulting ROC curve is called the empirical ROC curve (Fig. 1C) (1). The ROC curve illustrates the relationship between sensitivity and FPR. Because the ROC curve displays the sensitivities and FPRs at all possible cutoff levels, it can be used to assess the performance of a test independently of the decision threshold (5).

Area Under the ROC Curve: a Measure of Overall Diagnostic Performance

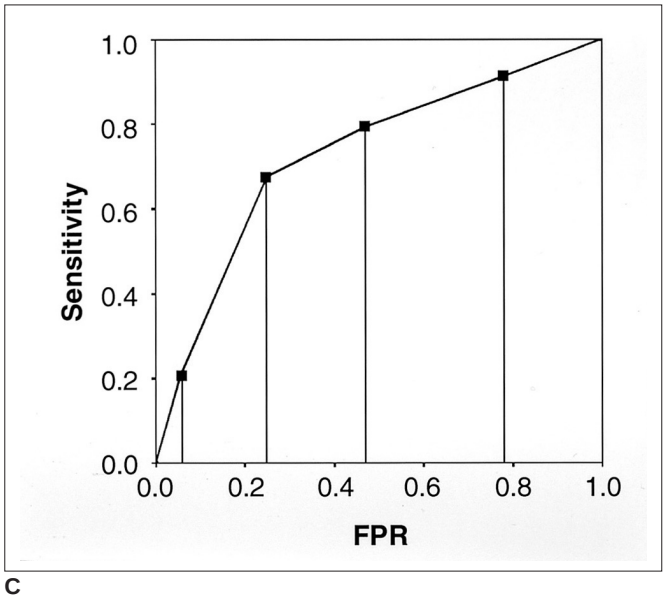
Several summary indices are associated with the ROC curve. One of the most popular measures is the area under the ROC curve (AUC) (1, 2). AUC is a combined measure of sensitivity and specificity. AUC is a measure of the overall performance of a diagnostic test and is interpreted as the average value of sensitivity for all possible values of specificity (1, 2). It can take on any value between 0 and 1, since both the x and y axes have values ranging from 0 to 1. The closer AUC is to 1, the better the overall diagnostic performance of the test, and a test with an AUC value of 1 is one that is perfectly accurate (Fig. 2). The practical lower limit for the AUC of a diagnostic test is 0.5. The line segment from 0, 0 to 1, 1 has an area of 0.5 (Fig. 2). If we were to rely on pure chance to distinguish those subjects with versus those without a particular disease, the resulting ROC curve would fall along this diagonal line, which is referred to as the chance diagonal (Fig. 2) (1, 2). A diagnos-



A



B



C

Fig. 1. ROC curves from a plain chest radiography study of 70 patients with solitary pulmonary nodules (Table 3).
A. A plot of test sensitivity (y coordinate) versus its false positive rate (x coordinate) obtained at each cutoff level.
B. The fitted or smooth ROC curve that is estimated with the assumption of binormal distribution. The parametric estimate of the area under the smooth ROC curve and its 95% confidence interval are 0.734 and 0.602 ~ 0.839, respectively.
C. The empirical ROC curve. The discrete points on the empirical ROC curve are marked with dots. The nonparametric estimate of the area under the empirical ROC curve and its 95% confidence interval are 0.728 and 0.608 ~ 0.827, respectively. The nonparametric estimate of the area under the empirical ROC curve is the summation of the areas of the trapezoids formed by connecting the points on the ROC curve.

tic test with an AUC value greater than 0.5 is, therefore, at least better than relying on pure chance, and has at least some ability to discriminate between subjects with and without a particular disease (Fig. 2). Because sensitivity and specificity are independent of disease prevalence, AUC is also independent of disease prevalence (1, 5).

AUC can be estimated both parametrically, with the assumption that either the test results themselves or some unknown monotonic transformation of the test results follows a binormal distribution, and nonparametrically from the empirical ROC curve without any distributional assumption of the test results (Figs. 1B, C). Several nonparametric methods of estimating the area under the empirical ROC curve and its variance have been described (6−8). The nonparametric estimate of the area under the empirical ROC curve is the summation of the areas of the trapezoids formed by connecting the points on the ROC curve (Fig. 1C) (6, 7). The nonparametric estimate of the area under the empirical ROC curve tends to underestimate AUC when discrete rating data (e.g., the five-point scale in the previous example) are collected, whereas the parametric estimate of AUC has negligible bias except when extremely small case samples are employed (2, 4). For discrete rating data, the parametric method is, therefore, preferred (2). However, when discrete rating data are collected, if the test results are not well distributed across the possible response categories (e.g., in the previous example, those patients with actually benign lesions and those patients with actually malignant lesions tend to be rated at each end of the scale, 1 = definitely benign and 5 = definitely malignant, respectively), the data may be degenerate and, consequently, the parametric method may not work well (2, 4). Using the nonparametric method is an option in this case, but may provide even more biased results than it normally would (2). For continuous or quasi-continuous data (e.g., a percent-confidence scale from 0% to 100%), the parametric and nonparametric estimates of AUC will have very similar values and the bias is negligible (2). Therefore, using either the parametric or nonparametric method is fine in this case (2). In most ROC analyses of radiological tests, discrete rating scales with five or six categories (e.g., definitely absent, probably absent, possibly present, probably present and definitely present) are used, for which the parametric method is recommended unless there is a problem with degenerate data. Data collection in radiological ROC studies is further discussed in a later section.

AUC is often presented along with its 95% confidence interval (CI). An AUC of a test obtained from a group of patients is not a fixed, true value, but a value from a sample that is subject to statistical error. Therefore, if one performs the same test on a different group of patients with the same characteristics, the AUC which is obtained may be different. Although it is not possible to specifically define a fixed value for the true AUC of a test, one can choose a range of values in which the true value of AUC lies with a certain degree of confidence. The 95% CI gives the range of values in which the true value lies and the associated degree of confidence. That is to say, one can be 95% sure that the 95% CI includes the true value of AUC (9, 10). In other words, if one believes that the true value of AUC is within the 95% CI, there is a 5% chance of its being wrong. Therefore, if the lower bound of the 95% CI of AUC for a test is greater than 0.5, then the test is statistically significantly better (with a 5% chance of being wrong or a significance level of 0.05) than making the diagnostic decision based on pure chance, which has an AUC of 0.5.

Comparing the Areas Under the ROC Curves: Comparing Overall Diagnostic Performance

Since AUC is a measure of the overall performance of a diagnostic test, the overall diagnostic performance of different tests can be compared by comparing their AUCs. The bigger its AUC is, the better the overall performance of the diagnostic test. When comparing the AUCs of two tests, equal AUC values mean that the two tests yield the

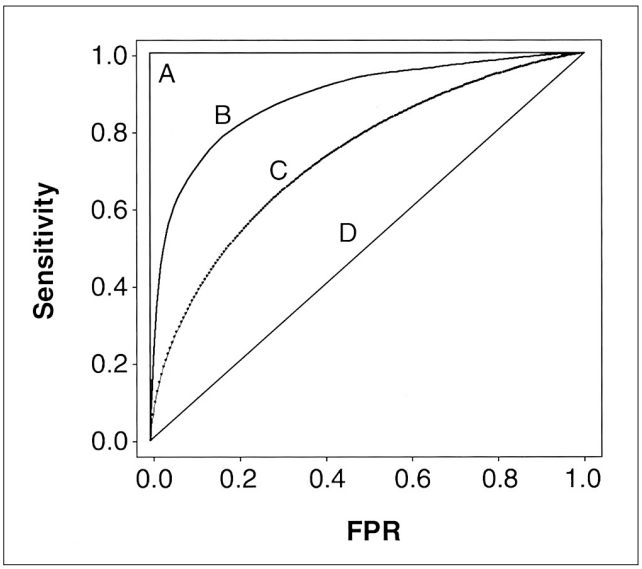


Fig. 2. Four ROC curves with different values of the area under the ROC curve. A perfect test (A) has an area under the ROC curve of 1. The chance diagonal (D, the line segment from 0, 0 to 1, 1) has an area under the ROC curve of 0.5. ROC curves of tests with some ability to distinguish between those subjects with and those without a disease (B, C) lie between these two extremes. Test B with the higher area under the ROC curve has a better overall diagnostic performance than test C.

same overall diagnostic performance, but does not necessarily mean that the two ROC curves of the two tests are identical (3). Figure 3 illustrates two ROC curves with equal AUCs. The curves are obviously not identical. Although the AUCs and, therefore, the overall performances of the two tests are the same, test B is better than test A in the high FPR range (or high sensitivity range), whereas test A is better than test B in the low FPR range (or low sensitivity range) (Fig. 3). The equality of two ROC curves can be tested by using the two parameters, a and b, instead. Because the shape of a binormal smooth ROC curve can be completely specified by the two parameters, a and b, the equality of the two ROC curves under the binormal assumption can be assessed by testing the equality of the two sets of parameters, a and b, i.e. by comparing the two sets of values from the two ROC curves. The null hypothesis and alternative hypothesis of the test are H_0: $a_1 = a_2$ and $b_1 = b_2$ versus H_1: $a_1 \neq a_2$ or $b_1 \neq b_2$, respectively, where 1 and 2 denote the two different ROC curves (2, 3). According to this method, the ROC curves and, consequently, the diagnostic performances of different tests are considered to be different, unless the ROC curves are identical: in other words, unless they yield equal sensitivities for every specificity between 0 and 1 or equal specificities for every sensitivity between 0 and 1 (4).

Sensitivity at a Particular FPR and Partial Area Under the ROC Curve

In some clinical settings, when comparing the performances of different diagnostic tests, one may be interested in only a small portion of the ROC curve and comparing the AUCs and the overall diagnostic performance may be misleading. When screening for a serious disease in a high-risk group (e.g., breast cancer screening), the cutoff range for a positive test should be chosen in such a way as to provide good sensitivity, even if the FPR is high, because false negative test results may have serious consequences. On the other hand, in screening for a certain disease, whose prevalence is very low and for which the subsequent confirmatory tests and/or treatments are very risky, a high specificity and low FPR is required. If the cutoff range for a positive test is not adjusted accordingly, almost all of the positive decisions will be false positive decisions, resulting in many unnecessary, risky follow-up examinations and/or treatments. In Figure 3, although the AUCs and overall performances of the two tests are the same, in the former diagnostic situation requiring high sensitivity, test B would be better than test A, whereas in the latter situation requiring a low FPR, test A would be better than test B. AUC, as a measure of the overall diagnostic performance, is not helpful in these specific diagnostic situations. The diagnostic performance of a test should be judged in the context of the diagnostic situation to which the test is applied. And, depending on the specific diagnostic situation, only a portion of the overall ROC curve may need to be considered.

One way to consider only a portion of an ROC curve is to use the ROC curve to estimate the sensitivity at a particular FPR, and to compare the sensitivities of different ROC

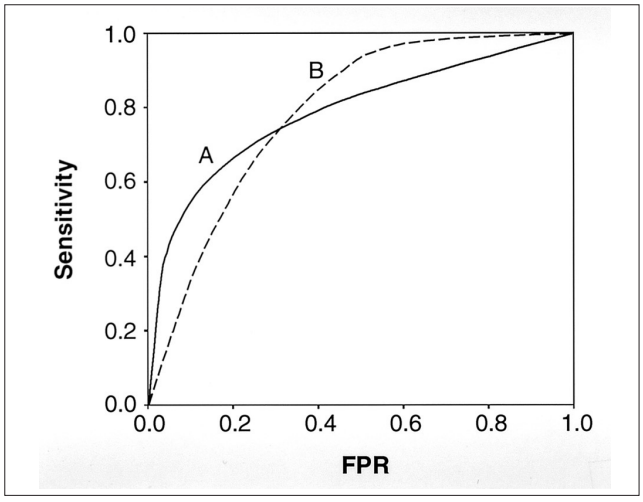


Fig. 3. Two ROC curves (A and B) with equal area under the ROC curve. However, these two ROC curves are not identical. In the high false positive rate range (or high sensitivity range) test B is better than test A, whereas in the low false positive rate range (or low sensitivity range) test A is better than test B.

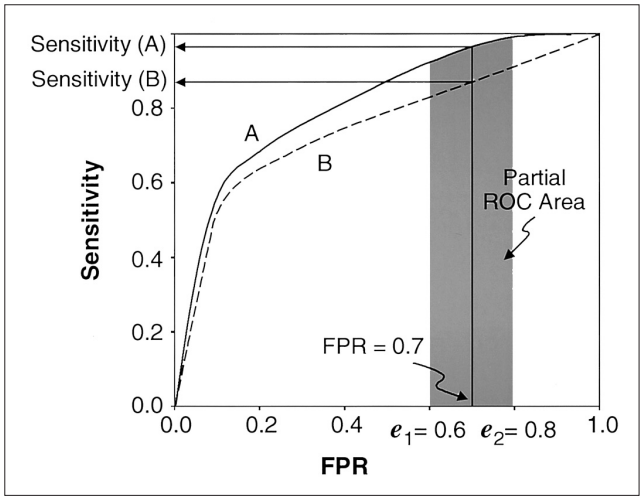


Fig. 4. Schematic illustration of a comparison between the sensitivities of two ROC curves (A and B) at a particular false positive rate and comparison between two partial ROC areas. For this example, the false positive rate and partial range of false positive rate ($e_1 - e_2$) are arbitrarily chosen as 0.7 and 0.6 ~ 0.8, respectively.

curves at a particular FPR (Fig. 4). Another way is to use the partial area under the ROC curve (Fig. 4) (11, 12). Partial ROC area is defined as the area between two FPRs or between two sensitivities. The partial area under the ROC curve between two FPRs, $FPR_1 = e_1$ and $FPR_2 = e_2$, can be denoted as $A(e_1 \leq FPR \leq e_2)$ (2). Unlike AUC, whose maximum possible value is always 1, the magnitude of the partial area under the ROC curve is dependent on the two FPRs chosen. Therefore, the standardization of the partial area by dividing it by its maximum value is recommended and Jiang et al. (12) referred to this standardized partial area as the partial area index. The maximum value of the partial area between $FPR_1 = e_1$ and $FPR_2 = e_2$ is equal to the width of the interval, $e_2 - e_1$. The partial area index is interpreted as the average sensitivity for the range of FPRs or specificities chosen (1, 2).

Data Collection in Radiological ROC Studies

Unlike in the case of many laboratory tests, the interpretation of most radiological tests is qualitative and there are several ways to express the reader's confidence in the presence of a disease, namely a binary result which is either positive or negative for the disease, a discrete rating scale such as a five-point scale, and a continuous or quasi-continuous scale such as a percent-confidence scale from 0% to 100% (2). The first approach is inadequate for ROC analysis, however, the second and third approaches are appropriate (2). In most of the ROC analyses of radiological tests which have been conducted to date, a discrete rating scale with five or six categories has been used. Rockette et al. (13) performed a study to assess how the estimates of performance on ROC curves are affected by the use of a discrete five-point scale versus a continuous percent-confident scale. They compared the AUCs obtained with the two different scales in the case of abdominal CTs used for detecting abdominal masses and suggested that the discrete rating or continuous scales are often not significantly different, and can be used interchangeably in image-evaluation ROC studies, although they recommended continuous scales for routine use in radiological ROC studies, because of their potential advantages in some situations (13). Having as many categories as possible or using a continuous or quasi-continuous scale is desirable theoretically (14) and has been shown to produce results essentially equivalent to those of discrete scales, when the latter produce well-distributed operating points (15).

Software for ROC Analysis

Several software programs that are frequently used for ROC analysis are available on the Internet.

ROCKIT, which is available at http://xray.bsd.uchicago.edu/krl/roc_soft.htm (accessed December 31, 2003), is a program for parametric ROC analysis that combines the features of ROCFIT, LABROC, CORROC2, CLABROC and INDROC. It estimates the smooth ROC curve and its AUC, 95% CI of AUC, and the parameters *a* and *b* on the basis of a binormal distribution. ROCKIT tests the statistical significance of the differences between two paired (i.e., two ROC curves from the same group of patients), partially paired, or unpaired (i.e., two ROC curves from two different groups of patients, viz. one curve each from each group of patients) ROC curves. The difference between two AUCs (i.e., the difference in the overall diagnostic performance of the two tests) is tested with the z test. Differences in the parameters *a* and *b* of two ROC curves (i.e., the equality of the two ROC curves) are tested using the bivariate chi-square test, as presented by Metz et al (2, 4). ROCKIT also estimates the sensitivity at a particular FPR and tests the statistical significance of the difference between the two sensitivities on the two curves at a particular FPR by means of the z test.

PlotROC.xls, which is available at http://xray.bsd.uchicago.edu/krl/roc_soft.htm (accessed December 31, 2003), is a Microsoft Excel 5.0 (Microsoft, Redmond, WA, U.S.A.) macro sheet which takes the *a* and *b* parameter values based on the assumption of a binormal distribution to plot a smooth ROC curve.

MedCalc (MedCalc Software, Mariakerke, Belgium), which is available at http://www.medcalc.be (accessed December 31, 2003), is a statistical package that offers nonparametric ROC analysis. It provides the empirical ROC curve and nonparametric estimate of the area under the empirical ROC curve with its 95% CI, based on the method developed by Hanley et al. (7). A comparison between two paired ROC curves is available and the statistical significance of the difference between two AUCs is calculated with the z test, as described by Hanley et al. (16). SPSS version 10.0 (SPSS Inc., Chicago, IL, U.S.A.) also provides the empirical ROC curve and nonparametric estimate of the area under the empirical ROC curve and its 95% CI, which are calculated using a method similar to that of Medcalc. However, it does not provide a statistical comparison between ROC curves.

Partarea.for, which is available at http://www.bio.ri.ccf.org/Research/ROC (accessed December 31, 2003), is a

FORTRAN program designed to estimate the partial area under the smooth ROC curve between two FPRs, based on the method developed by McClish (11). It also tests the statistical significance of the difference between the two partial areas of two ROC curves using the z test. This program should be used in conjunction with a parametric program such as ROCKIT. To estimate the partial area, it requires the a and b parameter estimates, along with the variances (a) and (b) and the covariance (a, b) of an ROC curve, which can be obtained by means of a parametric program. When comparing two partial areas of two ROC curves it also requires the covariances (a_1, a_2), (a_1, b_2), (b_1, a_2) and (b_1, b_2), which can be obtained using a parametric program (note : the subscripts 1 and 2 denote two different ROC curves). This program needs to be compiled before it can be used on a DOS or Windows-based computer.

Summary

- The ROC curve is a plot of test sensitivity along the y axis versus its 1-specificity or FPR along the x axis.
- In ROC analyses of radiological tests, discrete rating scales with five or six categories are widely used, however, it would be preferable to have as many categories as possible or to use a continuous or quasi-continuous scale for data collection.
- AUC, which is interpreted as the average value of sensitivity for all possible values of specificity, is a measure of the overall performance of a diagnostic test. AUC can take on any value between 0 and 1, where a bigger value suggests the better overall performance of a diagnostic test.
- The nonparametric estimate of the area under the empirical ROC curve tends to underestimate AUC when discrete rating data are collected, whereas the parametric estimate of AUC has negligible bias, except when extremely small case samples are employed. Therefore, when discrete rating scales are employed, the use of a parametric method is recommended.
- The diagnostic performance of a test should be judged in the context of the diagnostic situation to which the test is applied. The partial ROC area and sensitivity at a particular FPR are useful indicators, when only a portion of the entire ROC curve needs to be considered.

Appendix

Parameters a and b under assumption of binormal distribution (2)

If the data are actually binormal or if a known function can transform the data so that it follows a binormal distribution, parameters a (the standardized difference in the means of the distributions of the test results for those subjects with and without the condition) and b (the ratio of the standard deviations of the distributions of the test results for those subjects without versus those with the condition) can be estimated directly from the means and standard deviations of the distributions of those subjects with and without the condition. Thus, we will have

$$a = (u_1 - u_0) / \sigma_1; \qquad b = \sigma_0 / \sigma_1$$

where u_i is the mean and σ_i is the standard deviation of the test results, i = 0 (without the condition), 1 (with the condition).

For discrete rating data, we hypothesize discrete rating scale test results, T_0 (without the condition) and T_1 (with the condition) as a categorization of two latent continuous scale random variables, T^*_0 and T^*_1, respectively, each of which has a normal distribution. For a discrete rating scale test result, T_i, which can take on one of the K-ordered values, where i = 0 (without the condition) or 1 (with the condition), we assume that there are $K - 1$ unknown decision thresholds c_1, c_2, ..., c_{K-1}, so that

If $T^*_i \leq c_1$,	then $T_i = 1$
If $c_{j-1} < T^*_i \leq c_j$,	then $T_i = j$, j = 2, 3, ..., K - 1
If $T^*_i > c_{K-1}$,	then $T_i = K$

Because we assume that both T^*_0 and T^*_1 have normal distributions, then

$$T^*_0 \sim N(\mu_0, \sigma_0^2); \qquad T^*_1 \sim N(\mu_1, \sigma_1^2)$$

where μ_0, μ_1 are the means and σ_0^2, σ_1^2 are the variances of the normal distributions. Therefore, we will have

$$a = (\mu_1 - \mu_0) / \sigma_1; \qquad b = \sigma_0 / \sigma_1$$

Acknowledgements

The authors wish to thank Charles E. Metz, PhD at Kurt Rossmann Laboratories, Department of Radiology, University of Chicago, IL, USA for reviewing the manuscript and providing helpful comments, and Frank Schoonjans at MedCalc Software, Mariakerke, Belgium for providing the information on MedCalc.

References

1. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* 2003;229:3-8

2. Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine*, 1st ed. New York: John Wiley & Sons, 2002:15-164
3. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol* 1986;21:720-733
4. Metz CE. Some practical issues of experimental design and data analysis in radiologic ROC studies. *Invest Radiol* 1989;24:234-245
5. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283-298
6. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol* 1975;12:387-415
7. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29-36
8. 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
9. Motulsky H. *Intuitive biostatistics*, 1st ed. New York: Oxford University Press, 1995:9-60
10. Metz CE. Quantification of failure to demonstrate statistical significance: the usefulness of confidence intervals. *Invest Radiol* 1993;28:59-63
11. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989;9:190-195
12. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 1996;201:745-750
13. Rockette HE, Gur D, Metz CE. The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. *Invest Radiol* 1992;27: 169-172
14. Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. *Acad Radiol* 2001;8:328-334
15. Rockette HE, Gur D, Cooperstein LA, et al. Effect of two rating formats in multi-disease ROC study of chest images. *Invest Radiol* 1990;25:225-229
16. Hanley JA, McNeil BJ. A method comparing the areas under receiver operator characteristic curves derived from the same cases. *Radiology* 1983;148:839-843