Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models

Kelly H. Zou, PhD; A. James O'Malley, PhD; Laura Mauri, MD, MSc

Receiver-operating characteristic (ROC) analysis was originally developed during World War II to analyze classification accuracy in differentiating signal from noise in radar detection.[1] Recently, the methodology has been adapted to several clinical areas heavily dependent on screening and diagnostic tests,[2–4] in particular, laboratory testing,[5] epidemiology,[6] radiology,[7–9] and bioinformatics.[10]

ROC analysis is a useful tool for evaluating the performance of diagnostic tests and more generally for evaluating the accuracy of a statistical model (eg, logistic regression, linear discriminant analysis) that classifies subjects into 1 of 2 categories, diseased or nondiseased. Its function as a simple graphical tool for displaying the accuracy of a medical diagnostic test is one of the most well-known applications of ROC curve analysis. In *Circulation* from January 1, 1995, through December 5, 2005, 309 articles were published with the key phrase "receiver operating characteristic." In cardiology, diagnostic testing plays a fundamental role in clinical practice (eg, serum markers of myocardial necrosis, cardiac imaging tests). Predictive modeling to estimate expected outcomes such as mortality or adverse cardiac events based on patient risk characteristics also is common in cardiovascular research. ROC analysis is a useful tool in both of these situations.

In this article, we begin by reviewing the measures of accuracy—sensitivity, specificity, and area under the curve (AUC)—that use the ROC curve. We also illustrate how these measures can be applied using the evaluation of a hypothetical new diagnostic test as an example.

Diagnostic Test and Predictive Model

A diagnostic classification test typically yields binary, ordinal, or continuous outcomes. The simplest type, binary outcomes, arises from a screening test indicating whether the patient is nondiseased ($Dx=0$) or diseased ($Dx=1$). The screening test indicates whether the patient is likely to be diseased or not. When >2 categories are used, the test data can be on an ordinal rating scale; eg, echocardiographic grading of mitral regurgitation uses a 5-point ordinal (0, 1+, 2+, 3+, 4+) scale for disease severity. When a particular cutoff level or threshold is of particular interest, an ordinal scale may be dichotomized (eg, mitral regurgitation ≤2+ and >2+), in which case methods for binary outcomes can be used.[7] Test data such as serum markers (brain natriuretic peptide[11]) or physiological markers (coronary lumen diameter,[12] peak oxygen consumption[13]) also may be acquired on a continuous scale.

Gold Standard

To estimate classification accuracy using standard ROC methods, the disease status for each patient is measured without error. The true disease status often is referred to as the gold standard. The gold standard may be available from clinical follow-up, surgical verification, and autopsy; in some cases, it is adjudicated by a committee of experts.

In selection of the gold standard, 2 potential problems arise: verification bias and measurement error. Verification bias results when the accuracy of a test is evaluated only among those with known disease status.[14–16] Measurement error may result when a true gold standard is absent or an imperfect standard is used for comparison.[17,18]

Sensitivity and Specificity

The fundamental measures of diagnostic accuracy are sensitivity (ie, true positive rate) and specificity (ie, true negative rate). For now, suppose the outcome of a medical test results in a continuous-scale measurement. Let t be a threshold (sometimes called a cutoff) value of the diagnostic test used to classify subjects. Assume that subjects with diagnostic test values less than or equal to t are classified as nondiseased and that subjects with diagnostic test values greater than t are classified as diseased, and let m and n denote the number of subjects in each group. Once the gold standard for each subject is determined, a 2×2 contingency table containing the counts of the 4 combinations of classification and true disease status may be formed; the cells consist of the number of true negatives, false negatives, false positives, and true positives (the Table).

The accuracy of such binary-valued diagnostic tests is assessed in terms of the probability that the test correctly classifies a nondiseased subject as negative, namely the specificity (also known as the true negative rate), and the probability that the test correctly classifies a diseased subject

From Children's Hospital Boston (K.H.Z.), Harvard Medical School (K.H.Z., A.J.O., L.M.), Brigham and Women's Hospital (L.M.), and Harvard Clinical Research Institute (L.M.), Boston, Mass.

Correspondence to Kelly H. Zou, PhD, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115. E-mail kelly.zou@childrens.harvard.edu

(*Circulation.* 2007;115:654-657.)

Circulation is available at http://www.circulationaha.org

DOI: 10.1161/CIRCULATIONAHA.105.594929

Contingency Table of Counts Based on the Diagnostic Test and Gold Standard

	GS		
Dx	Nondiseased (GS=0)	Diseased (GS=1)	Total
Negative (Dx=0)	A=true negatives	B=false negatives	A+B=test negatives
Positive (Dx=1)	C=false positives	D=true positives	C+D=test positives
Total	A+C=nondiseased	B+D=diseased	A+B+C+D=total sample size

GS indicates gold standard; DX, diagnostic test. Specificity=true negative rate=A/(A+C). Sensitivity=true positive rate=D/(B+D). Negative predictive value=A/(A+B). Positive predictive value=D/(C+D). Disease prevalence=(B+D)/(A+B+C+D).

as positive, namely the sensitivity (also known as the true positive rate) (Figure 1).

When evaluating a continuous-scale diagnostic test, we need to account for the changes of specificity and sensitivity when the test threshold t varies. One may wish to report the sum of sensitivity and specificity at the optimal threshold (discussed later in greater detail). However, because the optimal value of t may not be relevant to a particular application, it can be helpful to plot sensitivity and specificity over a range of values of interest, as is done with an ROC curve. This inherent tradeoff between sensitivity and specificity also can be demonstrated by varying the choice of threshold.

ROC Analysis

An ROC curve is a plot of sensitivity on the *y* axis against (1−specificity) on the *x* axis for varying values of the threshold t. The 45° diagonal line connecting (0,0) to (1,1) is the ROC curve corresponding to random chance. The ROC curve for the gold standard is the line connecting (0,0) to (0,1) and (0,1) to (1,1). Generally, ROC curves lie between these 2 extremes. The area under the ROC curve is a summary measure that essentially averages diagnostic accuracy across the spectrum of test values Figure 2).

Estimation Methods

Nonparametric Methods

The empirical method for creating an ROC plot involves plotting pairs of sensitivity versus (1−specificity) at all possible values for the decision threshold when sensitivity and specificity are calculated nonparametrically. An advantage of this method is that no structural assumptions are made about the form of the plot, and the underlying distributions of the outcomes for the 2 groups do not need to be specified.[19] However, the empirical ROC curve is not smooth (Figure 3). When the true ROC curve is a smooth function, the precision of statistical inferences based on the empirical ROC curve is reduced relative to a model-based estimator (at least when the model is correctly specified). Analogous to regression, the specification of a model for the ROC curve enables information to be pooled over all values when estimating sensitivity or specificity at any 1 point. Smooth nonparametric ROC curves may be derived from estimates of density or distribution functions of the test distributions.[20]

Parametric Methods

As an alternative to the nonparametric approach, parametric models such as the binormal model may be assumed (Figure 3).[21–25] The binormal model assumes that both measurements have 2 independent normal distributions with different means and SDs. In our example, the distributions have a mean of 0 and an SD of 1 for the nondiseased population and a mean of 1.87 and an SD of 1.5 for the diseased population. These models have the further advantage of allowing easy incorporation of covariates into the model. By incorporating an optimal transformation, typically a log transformation to normal distributions, the estimated ROC curve may yield a better fit.[26–28]

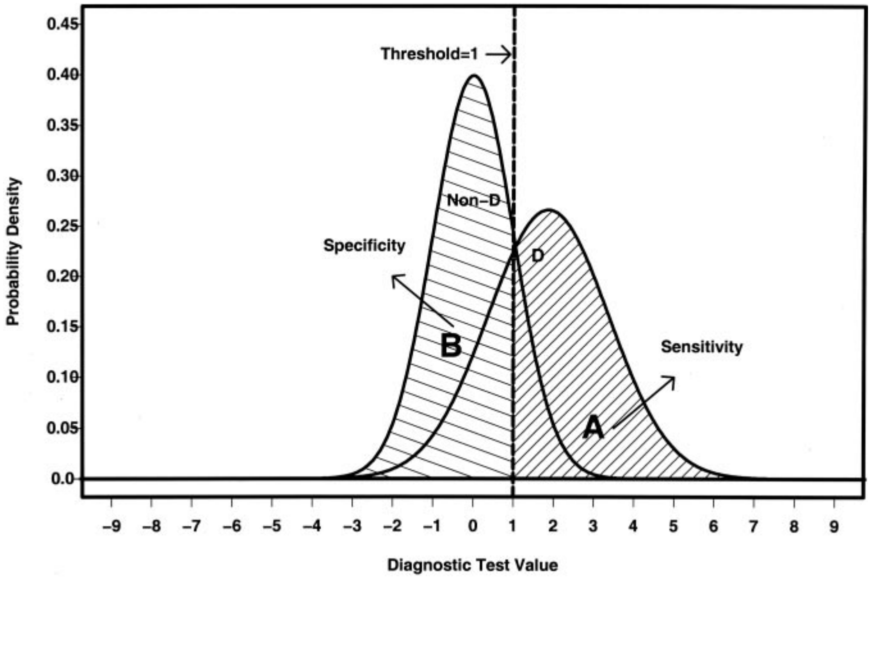

Figure 1. Probability density functions of a hypothetical diagnostic test that gives values on the real line. The density of the diagnostic test is plotted for each of 2 populations, nondiseased (Non-D) and diseased (D), assumed to follow the binormal model with a mixture of N(0,1) and N(1.87,1.5^2), respectively. The specificity of the diagnostic test is represented as the shaded area under the nondiseased distribution (A) above the arbitrary threshold t=1. Sensitivity is represented as the shaded area under the diseased distribution (B) below the same threshold of 1. For example, when the threshold value t=1, (sensitivity, specificity)=(0.72, 0.84). When the test is dichotomized (eg, positive if test value is greater than the threshold), both the sensitivity and specificity vary accordingly, with lower sensitivity and higher specificity as the threshold increases. In practice, a log transformation is often applied to positive-valued marker data to obtain symmetric density functions like those depicted above.[12]

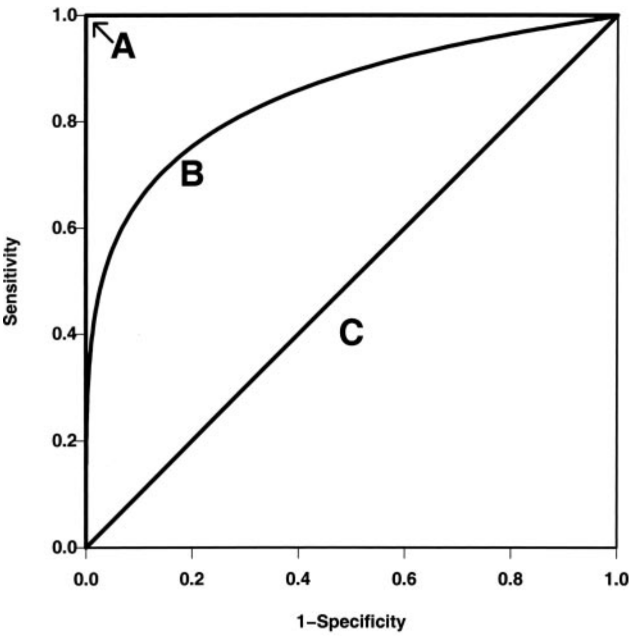

Figure 2. Three hypothetical ROC curves representing the diagnostic accuracy of the gold standard (lines A; AUC=1) on the upper and left axes in the unit square, a typical ROC curve (curve B; AUC=0.85), and a diagonal line corresponding to random chance (line C; AUC=0.5). As diagnostic test accuracy improves, the ROC curve moves toward A, and the AUC approaches 1.

Summary Measures

Confidence Intervals

A 95% confidence interval for the sensitivity at a given specificity, or vice versa, may be constructed using the bootstrap[29,30] or, for a bayesian model, using Markov-chain Monte Carlo simulation.[31] Alternatively, sample analytical approximations may be used instead of these computationally intensive numerical procedures.

Area Under the Curve

The AUC is an overall summary of diagnostic accuracy. AUC equals 0.5 when the ROC curve corresponds to random chance and 1.0 for perfect accuracy. On rare occasions, the estimated AUC is <0.5, indicating that the test does worse than chance.[31]

For continuous diagnostic data, the nonparametric estimate of AUC is the Wilcoxon rank-sum test, namely the proportion of all possible pairs of nondiseased and diseased test subjects for which the diseased result is higher than the nondiseased one plus half the proportion of ties. Under the binormal model, the AUC is a simple function of the mean and variance.[21,32]

Comparison of AUC Curves

An important problem concerns the comparison of 2 AUCs derived from 2 diagnostic tests administered on the same set of patients. Correlated *U* statistics may be compared.[33] Pearson correlation coefficients were used to estimate the correlation of the 2 AUCs.[34] A family of nonparametric comparisons based on a weighted average of sensitivities may be conducted.[35]

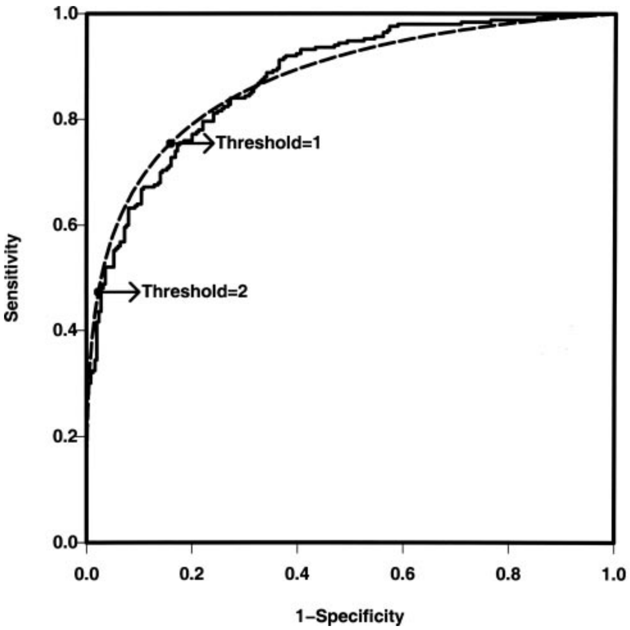

Figure 3. ROC curves derived from the example in Figure 1 using nonparametric and parametric estimation methods. The binormal model assumes a mixture distributions of N(0,1) and N(1.87,1.5^2), respectively. The nonparametric method yields a jagged curve; the parametric method yields a smoothed function. The AUC is 0.89. The (sensitivity, specificity) values correspond to the threshold values of 1 and 2, respectively. When the threshold value equals 1, (sensitivity, specificity)=(0.72, 0.84). In comparison, when the threshold value equals 2, (sensitivity, specificity)=(0.47, 0.98). At the optimal threshold of t=0.75, sensitivity=specificity=0.77.

Partial Area

The area under the ROC curve is a simple and convenient overall measure of diagnostic test accuracy. However, it gives equal weight to the full range of threshold values. When the ROC curves intersect, the AUC may obscure the fact that 1 test does better for 1 part of the scale (possibly for certain types of patients) whereas the other test does better over the remainder of the scale.[32,36] The partial area may be useful for the range of specificity (or sensitivity) of clinical importance (ie, between 90% and 100% specificity). However, partial area may be more difficult to estimate and compare on the basis of numerical integration methods; thus, full area is used more frequently in practice.[37]

Optimal Threshold

One criterion for evaluating the optimal threshold of a test is to maximize the sum of sensitivity and specificity. This is equivalent to maximizing the difference between the sensitivity of the test and the sensitivity that the test would have if it did no better than random chance.[9] For example, if both sensitivity and specificity are of importance in our example binormal model, the optimal threshold of t would be 0.75, where these 2 accuracy measures equal sensitivity and specificity equal 0.77 (Figure 3).

Discussion

ROC analysis is a valuable tool to evaluate diagnostic tests and predictive models. It may be used to assess accuracy quantitatively or to compare accuracy between tests or predictive models. In clinical practice, continuous measures are frequently converted to dichotomous tests. ROC analysis can be used to select the optimal threshold under a variety of clinical circumstances, balancing the inherent tradeoffs that exist between sensitivity and sensitivity. Several other specific applications of ROC analysis such as sample size determination[38–42] and meta-analysis[43,44] have been applied to clinical research. These can be derived from the fundamental principles discussed here.

Acknowledgments

We thank our colleagues, Daniel Goldberg-Zimring, PhD, and Marianna Jakab, MSc, of Brigham and Women's Hospital, Harvard Medical School, who assist in maintaining a comprehensive literature search website containing articles related to ROC methodology (http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html).

Sources of Funding

This research was made possible in part by grants R01LM007861, R01GM074068, U41RR019703, and P41RR13218 from the National Institutes of Health (NIH), Bethesda, Md. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Disclosures

None.

References

1. Lusted LB. Signal detectability and medical decision making. *Science*. 1971;171:1217–1219.
2. Lloyd CJ. Using smooth receiver operating characteristic curves to summarize and compare diagnostic systems. *J Am Stat Assoc*. 1998;93:1356–1364.
3. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. New York, NY: Wiley & Sons; 2002.
4. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press; 2003.
5. Campbell G. General methodology I: advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Stat Med*. 1994;13:499–508.
6. Shapiro DE. The interpretation of diagnostic tests. *Stat Methods Med Res*. 1999;8:113–134.
7. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology*. 2003;229:3–8.
8. Eng J. Receiver operating characteristic analysis: a primer. *Acad Radiol*. 2005;12:909–916.
9. O'Malley AJ, Zou KH, Fielding JR, Tempany CMC. Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: prostate biopsy and spiral CT of ureteral stone. *Acad Radiol*. 2001;8:713–725.
10. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform*. 2005;38:404–415.
11. Maisel A, Hollander JE, Guss D, McCollouph P, Nowak R, Green G, Saltzberg M, Ellison SR, Bhalla MA, Bhalla V, Clopton P, Jesse R, for the REDHOT Investigators. A multicenter study of B-type natriuretic peptide levels, emergency department decision making, and outcomes in patients presenting with shortness of breath. *J Am Coll Cardiol*. 2004;44:1328–1333.
12. Mauri L, Orav J, O'Malley AJ, Moses JW, Leon MZB, Holmes DR, Teirstein PS, Schofer J, Breithardt G, Cutlip DE, Kereiakes DJ, Shi C, Firth BG, Donohoe DJ, Kuntz R. Relationship of late loss in lumen diameter to coronary restenosis in sirolimus-eluting stents. *Circulation*. 2005;111:321–327.
13. O'Neill J, Young JB, Pothier CE, Lauer MS. Peak oxygen consumption as a predictor of death in patient with heart failure receiving β-blockers. *Circulation*. 2005;111:2313–2318.
14. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39:207–215.
15. Zhou XH, Higgs RE. Assessing the relative accuracies of two screening tests in the presence of verification bias. *Stat Med*. 2000;19:1697–1705.
16. Toledano AY, Gatsonis C. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics*. 1999;55:488–496.
17. Johnson WO, Gastwirth JL, Pearson LM. Screening without a "gold standard": the Hui-Walter paradigm revisited. *Am J Epidemiol*. 2001;153:921–924.
18. Phelps CE, Hutson A. Estimating diagnostic test accuracy using a "fuzzy gold standard." *Med Decis Making*. 1995;15:44–57.
19. Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann Stat*. 1996;24:24–40.
20. Zou KH, Hall WJ, Shapiro DE. Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Stat Med*. 1997;16:2143–2156.
21. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory: a direct solution. *Psychometrika*. 1968;33:117–124.
22. Metz CE, Herman BA, Shen J. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuous distributed data. *Stat Med*. 1998;17:1033–1053.
23. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. *J Appl Stat*. 2000;27:621–631.
24. Cai T, Moskowitz CS. Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics*. 2004;5:573–586.
25. Zou KH, Wells WM 3rd, Kikinis R, Warfield K. Three validation metrics for automated probabilistic image segmentation of brain tumours. *Stat Med*. 2004;23:1259–1282.
26. Hanley JA. The robustness of the "binormal" assumptions used in fitting ROC curves. *Med Decis Making*. 1988;8:197–203.
27. Walsh SJ. Goodness-of-fit issues in ROC curve estimation. *Med Decis Making*. 1999;19:193–201.
28. Zou KH, Resnic FS, Talos IF, Goldberg-Zimring D, Bhagwat JG, Haker SJ, Kikinis R, Jolesz FA, Ohno-Machado L. A global goodness-of-fit test for receiver operating characteristic curve analysis via the bootstrap method. *J Biomed Inform*. 2005;38:395–403.
29. Platt RW, Hanley JA, Yang H. Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test. *Stat Med*. 2000;19:313–322.
30. Zhou XH, Qin G. Improved confidence intervals for the sensitivity at a fixed level of specificity of a continuous-scale diagnostic test. *Stat Med*. 2005;24:465–477.
31. Hanley JA, McNeil BJ. The meaning and use of the area under a ROC curve. *Radiology*. 1982;143:27–36.
32. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making*. 1989;9:190–195.
33. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
34. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148:839–843.
35. Weiand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic makers with paired or unpaired data. *Biometrika*. 1989;76:585–592.
36. Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics*. 2003;59:614–623.
37. Walter SD. The partial area under the summary ROC curve. *Stat Med*. 2005;24:2025–2040.
38. O'Malley AJ, Zou KH. Bayesian multivariate hierarchical transformation models for ROC analysis. *Stat Med*. 2006;25:459–479.
39. Linnett K. Comparison of quantitative diagnostic tests: type I error, power and sample size. *Stat Med*. 1987;6:147–158.
40. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Stat Med*. 1997;16:1529–1542.
41. Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res*. 1998;7:371–392.
42. Eng J. Sample size estimation: a glimpse beyond simple formulas. *Radiology*. 2004;230:606–612.
43. Moses LE, Shapiro DE, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*. 1993;12:1293–1316.
44. Rutter CM, Gatsonis C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20:2865–2884.

KEY WORDS: diagnosis ■ ROC curve ■ sensitivity and specificity ■ statistics ■ tests