Receiver Operating Characteristic Curve in Diagnostic Test Assessment

Jayawant N. Mandrekar, PhD

Abstract: The performance of a diagnostic test in the case of a binary predictor can be evaluated using the measures of sensitivity and specificity. However, in many instances, we encounter predictors that are measured on a continuous or ordinal scale. In such cases, it is desirable to assess performance of a diagnostic test over the range of possible cutpoints for the predictor variable. This is achieved by a receiver operating characteristic (ROC) curve that includes all the possible decision thresholds from a diagnostic test result. In this brief report, we discuss the salient features of the ROC curve, as well as discuss and interpret the area under the ROC curve, and its utility in comparing two different tests or predictor variables of interest.

Key Words: Sensitivity, Specificity, ROC, AUC.

(*J Thorac Oncol.* 2010;5: 1315–1316)

In a previous article, we discussed the measures of sensitivity and specificity that rely on a single cutpoint to classify a test result as positive or negative.[1] In the event of a continuous or ordinal predictor, there are often multiple such cutpoints. Although sensitivity and specificity can be computed treating each value of the predictor as a possible cutpoint, a receiver operating characteristic (ROC) curve that includes all the possible decision thresholds from a diagnostic test result offers a more comprehensive assessment. In this review, we will introduce the salient features of an ROC curve, discuss the measure of area under the ROC curve (AUC), and introduce the methods for the comparison of ROC curves.

ROC CURVE

Simply defined, an ROC curve is a plot of the sensitivity versus 1 − specificity of a diagnostic test. The different points on the curve correspond to the different cutpoints used to determine whether the test results are positive. An ROC curve can be considered as the average value of the sensitivity for a test over all possible values of specificity or vice versa. A more general interpretation is that given the test results, the probability that for a randomly selected pair of patients with and without the disease/condition, the patient with the disease/condition has a result indicating greater suspicion.[2,3]

As a simple illustration, Table 1 gives the ratings of images obtained from 109 subjects by a radiologist.[2,3] Multiple cutpoints are possible for classifying a patient as normal or abnormal based on the image ratings. Suppose that ratings of 4 or above indicate, for instance, that the test is positive (abnormal), then the sensitivity and specificity would be 0.86 (44/51) and 0.78 (45/58), respectively. In contrast, if the ratings of 3 or above were to be considered as positive, then the sensitivity and specificity are 0.90 (46/51) and 0.67 (39/58), respectively. This illustrates that both sensitivity and specificity are specific to the selected decision threshold. Moreover, the designation of a cutpoint to classify the test results as positive or negative is relatively arbitrary.

An ROC curve, on the other hand, does not require the selection of a particular cutpoint. See Figure 1 for the ROC curve for the data presented in Table 1. An ROC curve essentially has two components, the empirical ROC curve that is obtained by joining the points represented by the sensitivity and 1 − specificity for the different cutpoints and the chance diagonal represented by the 45-degree line drawn through the coordinates (0,0) and (1,1). If the test results diagnosed patients as positive or negative for the disease/condition by pure chance, then the ROC curve will fall on the diagonal line. Sometimes a fitted (smooth) ROC curve based on a statistical model can also be plotted in addition to the empirical ROC curve.

An overall ROC curve is most useful in the early stages of evaluation of a new diagnostic test. Once the diagnostic ability of a test is established, only a portion of the ROC curve is usually of interest, for example, only regions with high specificity and not the average specificity over all sensitivity values. Similar to sensitivity and specificity, ROC curves are invariant to the prevalence of a disease but dependent on the patient characteristics and the disease spectrum. An ROC curve does not depend on the scale of the test results and can be used to provide a visual comparison of two or more test results on a common scale. The latter is not possible with sensitivity and specificity measures because a change in the cutpoint to classify the test results as positive or negative could affect the two tests differently.[4]

AREA UNDER THE ROC CURVE

AUC is an effective way to summarize the overall diagnostic accuracy of the test. It takes values from 0 to 1, where a

Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota.
Disclosure: The author declares no conflicts of interest.
Address for correspondence: Jayawant N. Mandrekar, PhD, Department of Health Sciences Research, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905. E-mail: mandrekar.jay@mayo.edu

ISSN: 1556-0864/10/0509-1315

TABLE 1. True Disease Status by Image Ratings

	Image Ratings					
True Disease Status	**1 = Definitely Normal**	**2 = Probably Normal**	**3 = Unsure**	**4 = Probably Abnormal**	**5 = Definitely Abnormal**	**Total**
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
Total	36	8	8	22	35	109

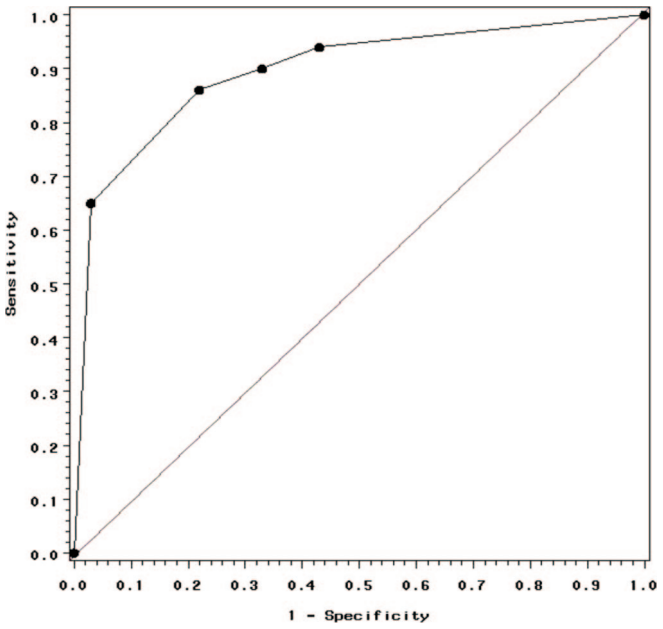

FIGURE 1. The receiver operating characteristic curve for the data in Table 1.

value of 0 indicates a perfectly inaccurate test and a value of 1 reflects a perfectly accurate test. AUC can be computed using the trapezoidal rule.[3] In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based on the test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.[5]

A value of 0.5 for AUC indicates that the ROC curve will fall on the diagonal (i.e., 45-degree line) and hence suggests that the diagnostic test has no discriminatory ability. ROC curves above this diagonal line are considered to have reasonable discriminating ability to diagnose patients with and without the disease/condition. It is therefore natural to do a hypothesis test to evaluate whether the AUC differs significantly from 0.5. Specifically, the null and alternate hypotheses are defined as H0: AUC = 0.5 versus H1: AUC ≠ 0.5. This test statistic given by $[\hat{AUC} - 0.5/SE(\hat{AUC})]$ is approximately normally distributed and has favorable statistical properties.[6]

For the data in Table 1, the AUC is 0.89. This suggests an 89% chance that the radiologist reading the image will correctly distinguish a normal from an abnormal patient based on the ordering of the image ratings. However, in the event of a tied rating, the assumption is that the radiologist will randomly assign one patient as normal and the other as abnormal. A formal hypothesis test of H0: AUC = 0.5 versus H1: AUC ≠ 0.5 for this example yields a test statistic of 12.2, with a p value <0.001, indicating that this test has excellent discriminating ability.[5]

COMPARING TWO OR MORE ROC CURVES

ROC curves are useful for comparing the diagnostic ability of two or more screening tests or for assessing the predictive ability of two or more biomarkers for the same disease. In general, the test with the higher AUC may be considered better. However, in cases where specific values of sensitivity and specificity are only clinically relevant for the comparison, then partial AUCs are compared.

ROC curves generated using data from patients where each patient is subjected to two (or more) different diagnostic tests of interest are considered as correlated ROC curves. ROC curves generated using data from different groups of patients where patients within each group is subjected to two different diagnostic tests are referred as uncorrelated ROC curves. The comparison of two uncorrelated ROC curves is relatively simple and is based on a form of a Z statistic that uses the difference in the area under the two curves and the SD of each AUC. In the case of correlated ROC curves, we refer the readers to a nonparametric approach proposed by DeLong et al.[7]

SUMMARY

Studies designed to measure the performance of diagnostic tests are important for patient care and health care costs. ROC curves are a useful tool in the assessment of the performance of a diagnostic test over the range of possible values of a predictor variable. The area under an ROC curve provides a measure of discrimination and allows investigators to compare the performance of two or more diagnostic tests.

REFERENCES

1. Mandrekar JN. Simple statistical measures for diagnostic accuracy assessment. *J Thorac Oncol* 2010;5:763–764.
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
3. Rosner B. Fundamentals of Biostatistics, 6th Ed. Chapter 3. Belmont, CA: Duxbury, 2005. Pp. 64–66.
4. Turner DA. An intuitive approach to receiver operating characteristic curve analysis. *J Nucl Med.* 1978;19:213–220.
5. Hosmer DW, Lemeshow S. Applied Logistic Regression, 2nd Ed. Chapter 5. New York, NY: John Wiley and Sons, 2000. Pp. 160–164.
6. Zhou XH, Obuchowski NA, Obuchowski DM. Statistical Methods in Diagnostic Medicine. Chapter 2. New York, NY: John Wiley and Sons, 2002. Pp. 27–33.
7. 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.