Statistical Methodology: III. Receiver Operating Characteristic (ROC) Curves

Mary Grzybowski, MPH, PhD, John G. Younger, MD

ABSTRACT

Measures including sensitivity, specificity, and positive and negative predictive values have been traditionally used to assess a diagnostic test's ability to detect the presence or absence of disease. Receiver operating characteristic (ROC) curve analysis allows visual evaluation of the trade-offs between sensitivity and specificity associated with different values of the test result, or different "cutpoints" for defining a positive result. The purpose of this article is to define, construct, and interpret a ROC curve using a hypothetical example applicable to emergency medicine practice.

Key words: receiver operating characteristic curves; ROC curves; diagnostic test; sensitivity; specificity; statistics; statistical methodology.

Acad. Emerg. Med. 1997; 4:818–826.

Receiver or relative operating characteristic (ROC) curve analysis originated from studies concerning the ability of human observers to distinguish between true signals and white noise in radar equipment.[1–3] A natural transition into the field of interpretation of radiologic data has occurred where ROC curve analysis has been applied toward interreader variability in the evaluation of radiologic imaging studies.[4] This method has been subsequently used in other medical disciplines such as experimental psychophysics and psychology, allergy, cardiology, and oncology research.[5] Despite its presence in the medical literature over the last decade, ROC curve analysis is still poorly understood by many clinicians.

ROC curves have increasingly been used to assess the ability of diagnostic tests to correctly classify disease status, or similarly, identify who will and will not develop a particular disease or health outcome. Whereas traditional dichotomous treatment of diagnostic test results yields only sensitivity and specificity data, ROC curve analysis objectively evaluates the performance of diagnostic tests at various cutpoints of positive and negative test results. This powerful tool allows the comparison of ≥2 diagnostic tests without specifying cutpoints for each test. ROC curve analysis permits the comparison of new tests against the truth or "criterion standard" test that would be impossible using only sensitivity and specificity.

From VAMC Health Services Research and Development, Ann Arbor, MI, Center for Practice Management and Outcomes Research (MG); Henry Ford Health Sciences Center, Detroit, MI, Department of Emergency Medicine (MG); and the University of Michigan, School of Medicine, Ann Arbor, MI, Section of Emergency Medicine (JGY).

Series editor: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor–UCLA Medical Center, Torrance, CA.

Received: September 25, 1995; revision received: September 10, 1996; accepted: January 24, 1997; updated: March 11, 1997.

Address: Mary Grzybowski, MPH, PhD, Center for Practice Management and Outcomes Research, Health Services Research and Development, VAMC, P.O. Box 130170, Ann Arbor, MI 48113-0170. Fax: 313-930-5159; e-mail: mgrzyb@umich.edu

USE OF SURROGATE DIAGNOSTIC TESTS

Diagnostic testing is a cornerstone of making clinical decisions and patient management. The amount of help a particular test result provides varies from patient to patient, depending on the patient's physiologic status. Numerous types of clinical data, not just a single test, are considered in the decision process. A test may include a laboratory value such as a peripheral white blood cell (WBC) count, an arterial blood pressure (BP), or the presence or absence of diminished breath sounds on chest auscultation. While all these tests are clinically meaningful, rarely are their results considered diagnostic of a disease process; i.e., they are not definitive. Yet the purpose of each clinical test is to provide some definable discrimination with the nondiseased state.

A definitive diagnostic test declares with absolute reliability the presence or absence of disease. Anything less is considered a surrogate and serves as an imperfect substitute for the definitive diagnostic test.[6] Diagnostic tests have either continuous or categorical outcomes. Peripheral WBC counts, arterial BPs, and weight are continuous and may assume a potentially infinite number of values

along a continuum. Data that can be classified into ≥2 naturally occurring or arbitrarily selected groups are termed "categorical data." Examples of categorical data include the presence or absence of breath sounds, a positive or negative history of heart disease, and A-B-O blood groups. Experience suggests that none of the above data, whether continuous or categorical, provide a definite answer in every clinical situation. Furthermore, because many diseases can be tested in a variety of methods (e.g., physical examinations, clinical suspicions, laboratory analyses), a means of determining the best test is essential.

Health care professionals are often concerned about how well a test performs clinically. There is always the possibility of replacing a current test with a newer test, adding a test to the already existing catalog of tests, or simply eliminating tests that do not appear useful. Therefore, it is important to know the accuracy of each diagnostic test or how one test performs compared with another test.

Why not opt for the definitive test every time? While diagnostic certitude may be comforting, it typically comes with a price. Definitive tests are often much more technically difficult, time-consuming, and expensive than their surrogates.[7] For example, the enzyme-linked immunosorbent assay is an inexpensive surrogate for Western Blot analysis for detection of HIV. In some circumstances, definite diagnostic tests may pose a bodily threat to patients, as in the case of pulmonary angiography for diagnosing pulmonary embolism or emergent pericardiocentesis for detecting cardiac tamponade. Thus, surrogates tests, while not as accurate as definitive tests, frequently must suffice.

Diagnostic accuracy is the most fundamental component of any clinical tool. It assesses the test's ability to discriminate among alternative states of health. A diagnostic test's ability to distinguish between the population of diseased and nondiseased patients is typically assessed by measuring the test's sensitivity and specificity. "Sensitivity" is the probability of testing positive given the true presence of disease. "Specificity" is the probability of testing negative if the disease is truly absent. The ability to predict the presence or absence of disease from a diagnostic test is dependent on the prevalence of disease in the population under investigation as well as a test's sensitivity and specificity. The probability that any positive test is predictive of disease increases as the prevalence of disease in a population increases. The proportion of truly diseased individuals among all those with positive test results is referred to as the "positive predictive value" of the diagnostic test. Similarly, the proportion of non-

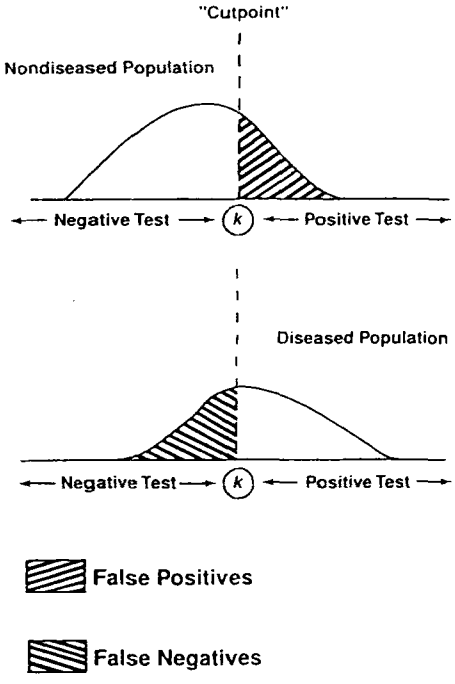

FIGURE 1. The cutoff value, *k*, discriminates between the diseased and nondiseased patients in a population. Patients with a measured value above *k* are defined as being positive for the disease. Patients with a measured value below *k* are defined as being negative for the disease.

diseased patients among all those with negative test results is the "negative predictive value." Table 1 summarizes the relationship of sensitivity, specificity, and predictive values between a surrogate and a definitive test according to a diagnostic criterion using a 2 × 2 contingency table.

Clinicians often attempt to predict the presence or absence of disease using a test to determine a patient's health status based on a decision threshold, or "cutpoint." For example, a rectal temperature higher than 38°C in a newborn is strongly suggestive of sepsis. Although the precise level of the cutpoint may be arbitrary, a decision threshold must be agreed upon to determine whether a test is positive or negative for an individual. As depicted in Figure 1, a patient presenting with a value above some threshold value, *k*, is defined as having an abnormal test result, and a patient presenting with a value below

TABLE 1 Sensitivity, Specificity, and Positive and Negative Predictive Values of Diagnostic Tests

	True Condition		
	Disease Positive	Disease Negative	Total
Test positive:	*a* (true positives)	*b* (false positives)	*a* + *b*
Test negative:	*c* (false negative)	*d* (true negative)	*c* + *d*
TOTAL	*a* + *c* (T_{pos})	*b* + *d* (T_{neg})	T_N*
True positive rate (TPR):	Sensitivity = a/(a + c)		
True negative rate (TNR):	Specificity = d/(b + d)		
False positive rate (FPR):	1 − *specificity*		
Positive predictive value	[a/a + b]: True positive/all who tested positive		
Negative predictive value	[d/c + d]: True negative/all who tested negative		

*T_N = total sample population.

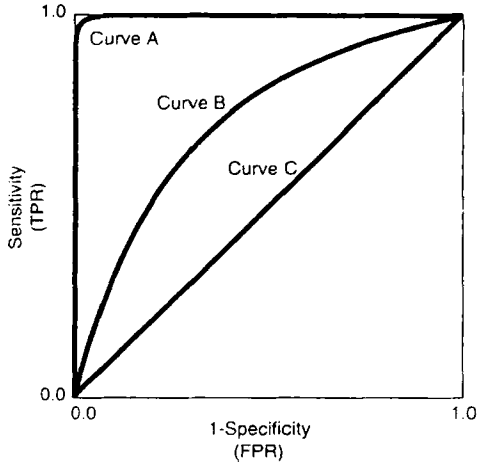

FIGURE 2. Receiver operating characteristic (ROC) curves. Curve A represents a perfect curve, and curves B and C represent nonperfect curves. Curve B is closer to the y-axis over a continuum of diagnostic threshold values than curve c. TPR = true-positive rate; FPR = false-positive rate.

k is defined as being negative for a disease or having a normal test result.

Based on diagnostic test results, the population with abnormal results is divided into a fraction that is truly ill (true-positive ratio, TPR) and a fraction incorrectly classified as ill (false-negative ratio, FNR). Normal or non-ill patients are similarly categorized into a fraction that is truly not ill, or normal, (true-negative ratio, TNR) and a fraction with erroneously positive test results (false-positive ratio, FPR). Sensitivity is equivalent to the TPR, and specificity is equivalent to the TNR.

In practice, a single decision threshold will not accurately and reliably discriminate the true health status of all patients. The percentage of patients who are considered positive or negative for the disease varies as does k, such that a reciprocal relationship exists between sensitivity and specificity. As k increases, sensitivity decreases and specificity increases, resulting in an increased probability of misdiagnosing an ill patient. As k decreases, sensitivity increases and specificity decreases, resulting in an increased probability of diagnosing false positives. Figure 1 also shows the overlap of false positives and false negatives based on a specific cutpoint, k. Sensitivity and specificity vary with different decision thresholds. Accordingly, attempting to determine the value of a test result using the sensitivity and specificity at only one cutpoint level may influence a clinician to erroneously embrace a dubious diagnostic technique or abandon a potentially useful one.[8,9]

ROC CURVE ANALYSIS

If determining the sensitivity and specificity using one cutpoint does not accurately discriminate between health and disease or between ill and severely ill states of health, what other way is there to assess the accuracy of a test? ROC curve analysis is a comprehensive method of comparing diagnostic tests over the whole spectrum of test results without having to specify a decision threshold a priori. ROC curve analysis can also assist in deriving an optimal value for k.

ROC analysis allows the consideration of a test's performance across a range of cutpoint values. It shows the complete spectrum of sensitivity–specificity pairs that correspond to all possible threshold values. A ROC curve is plotted on a 2-dimensional unit square plot. The y-axis and x-axis represent the TPR (or sensitivity) and the FPR (1 − specificity), respectively (Fig. 2).

A ROC curve may be interpreted by simple visual assessment. Qualitative information about the diagnostic test's accuracy is reflected by the position of the curve on a plot. ROC curves must pass through the lower left hand corner and upper right hand corner of the plot. A moment's reflection reveals why: By definition, a test with 100% sensitivity (i.e., would under no circumstance misidentify an ill patient) unfortunately has no specificity. Conversely, a test that is 100% specific forfeits its sensitivity. While some useful clinical tests provide near-perfect sensitivity and specificity, none achieves 100% sensitivity or specificity without sacrificing all specificity or sensitivity, respectively. From the clinical perspective, achieving the highest sensitivity and specificity simultaneously is often desirable. Curve A in Figure 2 represents a definitive diagnostic test, in which the sensitivity and specificity are both 100% for all k levels. Such a curve begins at the origin in the lower left hand corner (0.0, 0.0), continues to the top left hand corner (0.0, 1.0), and then continues to the upper right hand corner (1.0, 1.0). A nonperfect curve is one that lies to the right of and below a perfect curve (curves B and C in Fig. 2). The closer a diagnostic method's curve lies to the diagonal passing from (0, 0) to (1, 1), the poorer its performance.

While determining a cutpoint level that could be used in deciding whether a test's result is positive or negative would be ideal, there is currently no definite way to determine an optimal cutpoint level (i.e., maximization of both sensitivity and specificity). Other practical factors in determining precise cutpoints are involved, including economic and patient health care concerns. For example, if sensitivity is the primary goal in emergency medicine, thereby sacrificing specificity, expensive evaluations or therapy in patients incorrectly diagnosed as ill (false positives) may occur. The cost of unnecessary treatment, hospitalization, and resource utilization could be significant. Thus, deciding at which level k to set a test's cutpoint can be quite daunting.

Graphic evaluation of ≥ 2 diagnostic tests is useful in that the relative positions of ≥ 2 ROC curves provide a qualitative comparison of the accuracies of the tests across a range of sensitivity–specificity pairs. However, graphic comparisons do not in themselves provide quantitative information concerning the accuracy of ≥ 1 ROC curve. The area under the

curve (AUC) is one way to quantitate the "goodness" or accuracy of a test in discriminating between the 2 states of health.[10,11] Conventionally, the AUC (often referred to as *theta*, θ), is expressed as a single number and ranges from 0.5, corresponding to no accuracy, to 1.0, corresponding to perfect accuracy. The AUC does not provide a translation of a test result. General guidelines have been suggested by Swets for interpreting the AUC: 0.5–0.7 represents no to low discriminatory power; 0.7–0.9 represents moderate discriminatory power; and >0.9 represents high discriminatory power.[12]

To assess the value of one diagnostic test or to compare 2 different tests, a more elegant strategy is necessary. Qualitatively, when assessing the performance between 2 tests, the curve with the highest AUC may appear to be more accurate. Strictly speaking, when comparing 2 curves, or comparing a surrogate test's AUC with a nondiscriminating test's AUC (i.e., 0.50), one cannot simply choose the test with the largest AUC. Statistical procedures are required to determine whether an AUC is statistically significant or to compare the AUCs between ≥2 tests. A nonparametric procedure calculates the AUC from the observed points using a trapezoidal method. A very close relationship of the trapezoidal method and the Mann-Whitney U statistic was first noted by Bamber.[11] Hanley and McNeil developed and popularized this statistical procedure[13] and later demonstrated a technique for statistically comparing ≥2 ROC curves.[14] They and others have also addressed such comparisons based on curves derived from dependent and independent data.[13–15] Thus, appropriate statistical comparisons are needed to determine whether the accuracy of one test is greater than some reference value or significantly different from the accuracy of another diagnostic technique.[13–16] Several other advanced statistical methods have been developed for estimating ROC curves, AUCs, and confidence intervals. However, such specific statistical methods are beyond the scope of this paper.

CLINICAL APPLICATION OF ROC CURVES

Clinical Scenario: Appendicitis is a frequently encountered illness in the ED in which clinical and laboratory evaluations are applied prior to a criterion standard diagnostic procedure (i.e., laparotomy). Experience suggests that some physiologic parameters, such as peripheral WBC count, are more useful than others, such as oral temperature (the presence or absence of fever), in correctly diagnosing appendicitis. The magnitude of this diagnostic advantage is easily approached using ROC curves.

Imagine a prospective study in which patients suspected of having acute appendicitis have had peripheral WBC counts and oral temperatures taken. Once these data have been obtained, all patients are taken to laparotomy for a definitive diagnosis and undergo appendectomies as needed. In such a study, 2 surrogate tests (WBC count and oral temperature) can be compared with the truth (in this case, diagnosis at laparotomy, the criterion standard) and with each other. Certainly, if oral temperature were found to be as reliable as WBC count in determining appendicitis, phlebotomy could be avoided. If either test were as diagnostic as the criterion standard, unnecessary surgery could be avoided in patients with negative tests.

Assume 40 patients, 20 with appendicitis and 20 with other diagnoses at laparotomy, are studied. Table 2 contains the raw data from such a patient sample. Examination of these data reveal a wide range of WBC count test results and oral temperatures in both diseased and nondiseased patients. One approach to interpreting these results would be to compare the WBC count (assumed to be /μL [i.e., /mm^3]) and temperature (assumed to be °C) means for patients who had appendicitis with those of patients who had other diseases using an unpaired 2-sided Student's t-test (Table 3). From these calculations, it is clear that the mean WBC count among patients with appendicitis is significantly higher than that for patients with other diseases (12,819 vs

TABLE 2 Peripheral White Blood Cell (WBC) Counts (/μL) and Oral Temperatures (°C) for 40 Patients Presenting to the ED*

Appendicitis		Other Diagnosis	
WBC Count	Temperature	WBC Count	Temperature
4,600	38.0	1,043	38.4
6,790	36.3	1,296	39.5
8,234	39.7	4,300	38.0
9,300	37.9	4,500	36.4
10,000	37.3	4,600	36.9
10,000	37.7	5,640	37.8
10,010	37.7	5,900	38.0
10,030	38.2	6,920	36.5
11,400	37.3	6,940	38.0
11,890	39.4	6,974	37.6
11,940	38.0	7,630	38.0
12,200	38.9	8,245	36.0
15,000	37.9	8,500	38.7
15,400	39.9	8,900	39.0
16,000	39.2	9,130	37.8
16,700	38.8	9,360	39.5
18,000	39.6	9,970	39.3
18,790	39.5	11,000	38.5
19,600	40.0	14,000	40.0
20,500	38.5	15,970	39.6

*Disease classification (appendicitis or other diagnosis) is based on laparotomy.

TABLE 3 Mean White Blood Cell (WBC) Count (/μL) and Mean Temperature (°C) between Patients with Appendicitis and Patients with Other Diagnoses (Standard Errors)*

	Appendicitis (n = 20)	Other Illness (n = 20)	p-value
WBC count	12,819 (996)	7,540 (821)	≤0.001
Temperature	38.5 (0.2)	38.2 (0.3)	0.36

*The means were compared using Student's t-test for unequal variances.

3,675/μL, $p \leq 0.001$). Such a disparity is not apparent between the mean oral temperatures (38.5 vs 38.1°C, p = 0.36).

While useful, these conclusions leave unaddressed questions. For example, does the patient with a WBC count of 10,000/μL have appendicitis, or what temperature should prompt surgical exploration of a patient with abdominal pain? One method of analyzing the data is to calculate the sensitivity and specificity at several cutpoint values. Tables 4 and 5 include such calculations at 2 cutpoint values for WBC count and temperature, respectively. As noted in Table 4, when a WBC count *k* value of 8,500/μL is chosen, an 85% sensitivity is noted. Unfortunately, many people with WBC counts above this value will not have the disease (i.e., the specificity, in this case 55%, is low). When *k* is set at a higher value (15,000/μL), the sensitivity falls to 40%, while the specificity rises to 95%. These data reflect the reciprocity between sensitivity and specificity that is dictated by the cutpoint used to define a positive and negative test. Table 5 shows similar reciprocal behavior between sensitivity and specificity when the presence or absence of appendicitis is based on 2 specific oral temperature cutpoints. From these tables, it is evident that neither the sensitivity nor the specificity is particularly useful in discriminating between diseased states.

Analysis: ROC curve analysis is a more precise method to determine the accuracy of WBC count and oral temperature when determining which patients truly have appendicitis and which patients truly do not have appendicitis. ROC curve analysis also statistically tests which physiologic parameter—WBC count or temperature—is superior in its diagnostic ability. Thus, 3 distinct null hypotheses can be tested: 1) the peripheral WBC count is not accurate in distinguishing between ED patients with and without appendicitis ($\theta = 0.5$); 2) oral temperature is not accurate in discriminating between patients with and without appendicitis ($\theta = 0.5$); and 3) the accuracy of using WBC count as an indicator of appendicitis does not differ from the accuracy of using oral temperature as an indicator of appendicitis ($\theta_{WBC} = \theta_{OT}$). The ROC curves and statistical output must be analyzed to correctly interpret the results and to make decisions concerning the rejection or nonrejection of the 3 aforementioned hypoth-

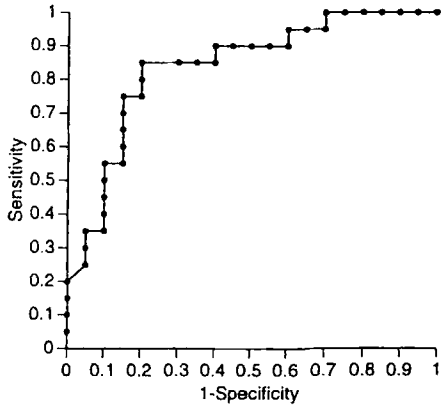

FIGURE 3. Receiver operating characteristic (ROC) curve obtained for peripheral white blood cell (WBC) count. The area under the curve is 0.82 ($p \leq 0.001$).

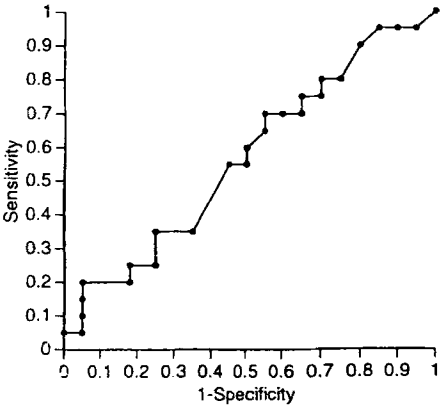

FIGURE 4. Receiver operating characteristic (ROC) curve obtained for oral temperature. The area under the curve is 0.566 (p = 0.23).

TABLE 4 Effect of Varied Cutpoints of White Blood Cell Counts (/μl) on the Presence and Absence of Appendicitis*

Cutpoint Values	2 × 2 Table			(TPR) Sensitivity	Specificity	(FPR) 1 − Specificity	PPV	NPV
		A+	A−					
≥8,500	Test +	17	8	85%	60%	40%	68%	80%
	Test −	3	12					
		A+	A−					
≥15,000	Test +	8	1	40%	95%	5%	89%	61%
	Test −	12	19					

*TPR = true-positive rate; FPR = false-positive rate; PPV = positive predictive value; NPV = negative predictive value; A+ = appendicitis; A− = other diagnosis.

TABLE 5 Effect of Varied Cutpoints of Temperature (°C) on the Presence and Absence of Appendicitis*

Cutpoint Values	2 × 2 Table			(TPR) Sensitivity	Specificity	(FPR) 1 − Specificity	PPV	NPV
		A+	A−					
≥38.0	Test +	13	13	65%	50%	50%	50%	50%
	Test −	7	7					
		A+	A−					
≥39.0	Test +	7	6	35%	70%	30%	54%	52%
	Test −	13	14					

*TPR = true-positive rate; FPR = false-positive rate; PPV = positive predictive value; NPV = negative predictive value; A+ = appendicitis; A− = other diagnosis.

eses. To conduct the analyses, a macro program written for the Statistical Analysis System (SAS 6.01, Cary, NC) package was used based on the statistical approach proposed by Hanley and McNeil.[13,14] The graphic and statistical output is described below.

Figures 3 and 4 show the ROC curves for WBC count and oral temperature regarding their discriminatory abilities to diagnose appendicitis, respectively, where the TPR and FPR are compared at various cutoff values. It is important to first assess the area under the ROC curve for each predictor of interest (Figs. 3 and 4, Table 6). Visually, it is evident that the AUC for WBC count is greater than the temperature's AUC from the ROC curve analysis. The ROC curve for WBC count is closer to the y-axis. By choosing to use the WBC count as a predictor of appendicitis, as opposed to the oral temperature, a clinician would diagnose fewer false negatives and false positives according to the graphic interpretation. But the hypotheses need to be statistically tested.

Is each AUC (WBC count for hypothesis 1 and oral temperature for hypothesis 2) different from 0.5? Table 6 shows, based on statistical evidence, that the null hypothesis can be rejected for the WBC count's ability to distinguish between patients with and without appendicitis. The AUC, when using WBC count as a predictor of appendicitis, is 0.819 and is statistically different from $\theta = 0.5$, $p \leq 0.001$. It can be concluded that peripheral WBC count is a useful parameter for discriminating appendicitis from other diseases. The AUC for oral temperature is 0.566. The null hypothesis is not contradicted by the data, and oral temperature is a poor predictor of appendicitis. Since this area is not significantly different from 0.5 ($p = 0.234$), the null hypothesis that oral temperature is a nondiscriminating index for patients with and without appendicitis cannot be rejected. According to Swet's general guidelines for interpreting the AUC, WBC counts and oral temperature demonstrate moderate and no discriminatory power, respectively.

To test for differences between the AUCs, a nonparametric test was used. Due to the linkage of the data (i.e., WBC counts and oral temperatures were obtained from the same patients), a method called the DeLong statistical procedure (using a χ distribution) was chosen to test for the dif-

TABLE 6 Statistical Output: Areas under the Curves for White Blood Cell (WBC) Count and Temperature by the Method of Hanley and McNeil[13,14]

Predictor	Area under the Curve (θ)	Standard Error of θ	p-value*
WBC count†	0.819	0.068	≤0.001
Temperature†	0.566	0.092	0.234

*Chi-square comparison of the areas under the curve of WBC count and oral temperature shows that the AUCs are significantly different ($\chi^2 = 6.99$ with 1 degree of freedom, $p \leq 0.01$).

†This is a 1-sided test of the null hypothesis: $\theta = 0.50$.

TABLE 7 Statistical Output: Sensitivities and Specificities of White Blood Cell Count (/μL) as a Predictor Variable for Appendicitis

Cutpoint	True Positives	True Negatives	Sensitivity	Specificity	1 − Specificity
1,043	20	0	1.00	0.00	1.00
1,296	20	1	1.00	0.05	0.95
4,300	20	2	1.00	0.10	0.90
4,500	20	3	1.00	0.15	0.85
4,600	20	4	1.00	0.20	0.80
5,640	19	5	0.95	0.25	0.75
5,900	19	6	0.95	0.30	0.70
6,790	19	7	0.95	0.35	0.65
6,920	18	7	0.90	0.35	0.65
6,940	18	8	0.90	0.40	0.60
6,974	18	9	0.90	0.45	0.55
7,630	18	10	0.90	0.50	0.50
8,234	18	11	0.90	0.55	0.45
8,245	17	11	0.85	0.55	0.45
8,500	*17*	*12*	*0.85*	*0.60*	*0.40*
8,900	17	13	0.85	0.65	0.35
9,130	17	14	0.85	0.70	0.30
9,300	17	15	0.85	0.75	0.25
9,360	16	15	0.80	0.75	0.25
9,970	16	16	0.80	0.80	0.20
10,000	16	16	0.80	0.80	0.20
10,010	14	16	0.70	0.80	0.20
10,030	13	16	0.65	0.80	0.20
11,000	12	17	0.60	0.85	0.15
11,400	12	18	0.60	0.90	0.10
11,890	11	18	0.55	0.90	0.10
11,940	10	18	0.50	0.90	0.10
12,200	9	18	0.45	0.90	0.10
14,000	8	18	0.40	0.90	0.10
15,000	8	19	0.40	0.95	0.05
15,400	7	19	0.35	0.95	0.05
15,970	6	19	0.30	0.95	0.05
16,000	6	20	0.30	1.00	0.00
16,700	5	20	0.25	1.00	0.00
18,000	4	20	0.20	1.00	0.00
18,790	3	20	0.15	1.00	0.00
19,600	2	20	0.10	1.00	0.00
20,500	1	20	0.05	1.00	0.00

TABLE 8 Statistical Output: Sensitivities and Specificities of Temperature (°C) as a Predictor Variable for Appendicitis

Cutpoint	True Positives	True Negatives	Sensitivity	Specificity	1 − Specificity
36.0	20	0	1.00	0.00	1.00
36.3	20	0	1.00	0.05	0.95
36.4	19	1	0.95	0.05	0.95
36.5	19	2	0.95	0.10	0.90
36.9	19	3	0.95	0.15	0.85
37.3	19	4	0.95	0.20	0.80
37.6	17	4	0.85	0.20	0.80
37.7	17	5	0.85	0.25	0.75
37.8	15	5	0.75	0.25	0.75
37.9	15	7	0.75	0.35	0.65
38.0	13	7	0.65	0.35	0.65
38.2	11	11	0.55	0.55	0.45
38.4	10	11	0.50	0.55	0.45
38.5	10	12	0.50	0.60	0.40
38.7	9	13	0.45	0.65	0.35
38.8	9	14	0.45	0.70	0.30
38.9	8	14	0.40	0.70	0.30
39.0	7	14	0.35	0.70	0.30
39.2	7	15	0.35	0.75	0.25
39.3	6	15	0.30	0.75	0.25
39.4	6	16	0.30	0.80	0.20
39.5	5	16	0.25	0.80	0.20
39.6	4	18	0.20	0.90	0.10
39.7	3	19	0.15	0.95	0.05
39.9	2	19	0.10	0.95	0.05
40.0	1	19	0.05	0.95	0.05

ferences in the AUCs[16] (asterisk footnote in Table 6). The AUCs for WBC count and oral temperature are significantly different from each other (0.819 vs 0.566, $p \leq 0.01$). Therefore, the third null hypothesis can be rejected, and it can be concluded that peripheral WBC count is superior to oral temperature in its ability to discriminate between patients with and without appendicitis ($\chi^2 = 6.99$, 1 degree of freedom, $p \leq 0.01$). In conclusion, these results, albeit contrived, showed that obtaining and measuring WBC counts may provide value for emergency physicians in the diagnosis of appendicitis, whereas oral temperatures do not help discriminate disease (appendicitis) from nondisease (no appendicitis).

Tables 7 and 8 represent sensitivity–specificity pair calculations for unique cutpoints in the data set generated from the ROC curve analysis for WBC count and temperature, respectively, using the macro program written in SAS (this macro can be obtained through correspondence with the author). It is actually through the use of these data that the ROC curve is generated.

ROC Data Details: To demonstrate the usefulness of the output and to interpret the results in these tables, one row of output serves as the example in this paper. The output for that row is explained column by column. For simplicity's sake, the italicized row in Table 7 represents one of the WBC cutpoints shown in Table 4, i.e., a cutpoint of 8,500/μL. It will also become apparent how 2 × 2 tables (such as Table 1) are formulated.

The first column (*Cutpoint*) represents a specific cutoff value for WBC counts. The cutpoints generated from the statistical procedure are arbitrary. As shown in the 14th row in Table 7, the cutpoint for classifying patients as having appendicitis is 8,500/μL. Using this cutpoint, patients presenting to the ED with WBC counts <8,500 were classified as nondiseased or negative for appendicitis, and any patients with levels ≥8,500 were classified as diseased or positive for appendicitis. The number of patients who were positive according to a cutpoint of 8,500 is noted in the second column of the output (*True Positives*). For example, 17 patients were classified as having appendicitis, according to the 8,500 WBC cutpoint, who were truly diagnosed at laparotomy with appendicitis. The third column (*True Negatives*) denotes the number of patients who were negative according to the 8,500 cutpoint. For example, 11 patients were classified as having a medical condition other than appendicitis at this cutpoint. Column 2 is the cumulative total moving from the bottom up and column 3 is the cumulative total moving from the top down. In column 2, for example, in the last row in Table 7 where $k = 20{,}500/\mu L$, only 1 patient is classified as a true positive, whereas in the first row where $k = 1{,}296$, all 20 patients are classified as true positives. In column 3, in the first row in Table 7 where $k = 1{,}296$, no patients are classified as true negatives, whereas in the last row where $k = 20{,}500$, all 20 patients are classified as true negatives.

The sensitivity of the test at a specific cutpoint is calculated in column 4 (*Sensitivity*). For each cutpoint, sensitivity is defined as the number of patients with appendicitis who have WBC counts equal to or above the cutpoint (true positives) divided by the total number of patients who truly have appendicitis at laparotomy (true positives plus false negatives). For example, when 8,500 is used as the cutpoint, the sensitivity of the test is 85% ($n = 17/20$).

The specificity of the test at a specific cutoff value is noted in column 5 (*Specificity*). For example, when 8,500 is used as the WBC count cutpoint, the specificity is 55% ($n = 11/20$). The specificity is defined as the number of patients without appendi-

citis with values beneath the cutpoint (true negatives) divided by the total number of truly nonappendicitis patients (true negatives plus false positives). The last column is merely 1 − specificity. The values in column 6 (1 − specificity) are plotted against the values in column 4 (sensitivity) for each cutpoint noted in Table 7. Hence, the ROC curve is created.

The sensitivity and specificity can also be calculated by hand using the output in Table 7. Using Table 1 as a calculation template, one can simply insert the numbers of true positives and true negatives in the appropriate cells of the 2 × 2 table and calculate the number of false negatives and false positives. For example, it is known that 20 patients truly have appendicitis and 20 patients have another disease at laparotomy. Therefore, the total sum of the columns, T_{pos} and T_{neg}, in Table 1 are fixed at 20 patients each. By referring to Table 1 and knowing that there are 17 true positives (cell *a*) from the second column in Table 7, the number of false negatives is 3 (cell *c*). Similarly, there are 11 patients who truly do not have appendicitis (cell *d*); therefore, there are 9 false positives (cell *b*).

CONCLUSION

Table 9 is an overview of ROC curve analysis. ROC curves are frequently used to evaluate diagnostic tests to differentiate "healthy" individuals from "diseased" individuals. A diagnostic test's sensitivity and sensitivity are not the most useful methods in determining a test's ability to distinguish between diseased and nondiseased patients. The advantage of ROC curve analysis is that a graphic display of the sensitivity–specificity interdependence of the diagnostic test(s) with varying cutpoints allocated for the decision is produced.

Selecting the best ROC curve cutpoint is difficult and presents a risk–benefit enigma. Cost analysis can be performed to determine the level of tolerance for missed diseased and missed nondiseased patients. There may be clinical situations where proper categorization of all diseased patients may be more important than the risk of mislabeling nondiseased patients, and vice versa. Generally, moving up and to the right on the ROC curve is associated with more

TABLE 9 Overview of ROC Curve Analysis

Alternate names: Also referred to as receiver or relative operating characteristic curves.

Data type:

Continuous or categorical data: If there are >2 categories, the categories must be ordered in a clinically meaningful way (ordinal data).

Dependent or independent data: Dependent data are data derived from different diagnostic tests from a single sample of patients. Independent data are data derived from different diagnostic tests from different samples of patients.

Appropriate tests and assumptions: There is a binormal assumption used in many of the ROC-fitting algorithms. However, simulation studies have shown that even gross violations of the binormal assumption tend not to produce substantively different results. A parametric test is used when the data are binormal in their distribution. The maximum-likelihood estimation method is used to generate the model.

A nonparametric test such as the Mann-Whitney U test has an underlying continuity assumption. For this test, the AUC is theoretically $P(X < Y)$, where X denotes the values of the test in the nondiseased group and Y denotes the values of the test in the diseased group, respectively.

Principal results: The slope of the curve at a point is equal to the likelihood ratio associated with that result.

AUC: If the AUC (θ) = 0.5, there is no discriminatory power of the diagnostic test. If θ is significantly >0.5, the diagnostic test could be considered useful in its ability to discern the diseased from the nondiseased.

Testing the difference in the AUC for 2 different diagnostic tests provides evidence whether the 2 AUCs differ in their discriminatory powers.

Strengths:

1. Provides information concerning the discriminatory power of a diagnostic test.
2. Provides a visual representation of the trade between sensitivity and specificity.
3. Results can be used to choose the best clinical strategy in clinical practice.
4. Ordinal regression models in diagnostic test assessment can be used to control for covariates.
5. Existing computer programs for ROC curve analyses are becoming more available. Available ROC curve programs are listed in the suggested reading list (suggested readings 2, 7, 12, and 13).

Limitations:

1. Complex computations are involved, especially when evaluating multivariate characteristics.
2. The results must be weighed against the diagnostic test's risks and costs.
3. ROC curves may not equivocally identify the superior test because it is possible for the curves to cross.
4. Despite the results of ROC curve analysis, clinicians are frequently constrained by external forces to operate in a particular area of the ROC plane (e.g., they may have to maintain a sensitivity above 95% due to patients' expectations or institutional liability issues). In such an environment, a curve with a lower AUC may still perform better in the clinical arena, given that its application is restricted.
5. It is difficult to use ROC methods to evaluate combinations of test, whether sequential or simultaneous. Since patterns of test results are clinically more meaningful than single results, this is a serious drawback to the practical application of ROC methods.

and more lenient standards for diagnosis and, conversely, moving down and to the left is associated with more and more stringency for diagnosis. Therefore, the more common the disease of interest, the more lenient should be the standards of diagnosis, and vice versa. Further, the greater the cost of false negatives compared with false positives, the more lenient the standards for diagnoses should be, and vice versa.

As previously noted, diagnostic test outcomes can either be continuous or categorical. The appendicitis application of ROC curve analysis application dealt with continuous clinical data (WBC counts and oral temperatures) and a dichotomous outcome (appendicitis or nonappendicitis). There are situations, nevertheless, in which the categorical test result consists of >2 categories. For tests with >2 categorical variables, a fundamental requirement for ROC analysis is that the test results must be ordered in a clinically meaningful way. In other words, there is an inherently implied order of outcome of interest. For example, in diagnostic imaging, a scaling or rating system may be used: normal, equivocal, and abnormal.

One last relevant topic concerning ROC curve analysis is that of resampling methodology. There are occasions in which the underlying distribution of the data is unknown and the characteristics of a statistical model or calculated parameter are needed. Resampling refers to the statistical methods used to validate a statistical model without using an independent data set or to obtain an estimate of a parameter of interest and a parameter of uncertainty from a single set of observations. Resampling methods, such as the bootstrap and jackknife methods, involve repeating the statistical analysis a number of times, and each time it is repeated, a different simulated data set is used. Each simulated data set is obtained from the observed set of data using some plan or algorithm. The multiple statistical analyses used in resampling determine the statistical model's or calculated estimate's accuracy. For researchers who wish to understand ROC curves and resampling methods in more depth, a suggested reading list is provided.

The authors recognize the help of Drs. Nowak and Rivers, Henry Ford Hospital Emergency Department, and the Ann Arbor HSRD for supporting the development of the manuscript.

REFERENCES

1. McNeil BJ, Hanley J. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making. 1984; 4:137–50.
2. Egan JP. Signal Detection Theory and ROC Analysis. New York: Academic Press, 1975, pp 1–277.
3. Swets JA, Pickett RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York: Academic Press, 1982, pp 1–253.
4. Lusted LB. Signal detectability and medical decision-making. Science. 1971; 171:1217–9.
5. Metz CE. ROC methodology in radiological imaging. Invest Radiol. 1986; 21:720–33.
6. Feinstein AR. Clinical Epidemiology: The Architecture of Clinical Research. Philadelphia: W. B. Saunders, 1985, pp 1–812.
7. Henderson AR. Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. Ann Clin Biochem. 1993; 30:521–39.
8. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1987; 299: 926–30.
9. Diamond GA. Clinical epidemiology of sensitivity and specificity. J Clin Epidemiol. 1992; 45:9–12.
10. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging. 1989; 29:307–35.
11. 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.
12. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988; 240:1285–93.
13. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36.
14. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148:839–43.
15. Metz CE, Wang P-L, Kronman HB. A new approach for testing the significance of differences between ROC curves from correlated data. In: Deconink F (ed). Information Processing in Medical Imaging. The Hague: Nijhoff, 1984, pp 432–45.
16. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44:837–45.

SUGGESTED READINGS

1. Centor RM. Signal detectability: the use of ROC curves and their analyses. Med Decis Making. 1991; 11:102–6.
2. Görög G. An excel program for calculating and plotting receiver-operator characteristic (ROC) curves, histograms and descriptive statistics. Comput Biol Med. 1994; 24:167–9.
3. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36.
4. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148:839–43.
5. Hnatkova K, Poloniecki JD, Camm AJ, Malik M. Computation of multifactorial receiver operator and predictive accuracy characteristics. Comput Methods Programs Biomed. 1994; 42:147–56.
6. Katz D, Foxman B. How well do prediction equations predict? Using receiver operating characteristic curves and accuracy curves to compare validity and generalizability. Epidemiology. 193; 4:319–26.
7. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8:283–98.
8. Nettleman MD. Receiver operating characteristic curves. Infect Control Hosp Epidemiol. 1988; 9:374–7.
9. Schäfer H. Efficient confidence bounds for ROC curves. Stat Med. 1994; 13:1551–61.
10. Sukhatme S, Beam CA. Stratification in nonparametric ROC studies. Biometrics. 1994; 50:149–63.
11. Tosteson AN, Weinstein MC, Wittenber J, Begg CB. ROC curve regression analysis: the use of ordinal regression models for diagnostic test assessment. Environ Health Perspect. 1994; 102(suppl 8):73–8.
12. Vida S. A computer program for nonparametric receiver operating characteristic analysis. Comput Methods Programs Biomed. 1993; 40:95–101.
13. Zweig MH, Campbell G. Receiver-operating characteristic ROC plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993; 39:561–77.