

Methodological Review

# The use of receiver operating characteristic curves in biomedical informatics

Thomas A. Lasko<sup>a,b</sup>, Jui G. Bhagwat<sup>c</sup>, Kelly H. Zou<sup>c,d</sup>, Lucila Ohno-Machado<sup>a,c,\*</sup>

<sup>a</sup> Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

<sup>b</sup> Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA

<sup>c</sup> Department of Radiology, Brigham and Women's Hospital, USA

<sup>d</sup> Department of Health Care Policy, Harvard Medical School, USA

Received 22 November 2004

Available online 2 April 2005

## Abstract

Receiver operating characteristic (ROC) curves are frequently used in biomedical informatics research to evaluate classification and prediction models for decision support, diagnosis, and prognosis. ROC analysis investigates the accuracy of a model's ability to separate positive from negative cases (such as predicting the presence or absence of disease), and the results are independent of the prevalence of positive cases in the study population. It is especially useful in evaluating predictive models or other tests that produce output values over a continuous range, since it captures the trade-off between sensitivity and specificity over that range. There are many ways to conduct an ROC analysis. The best approach depends on the experiment; an inappropriate approach can easily lead to incorrect conclusions. In this article, we review the basic concepts of ROC analysis, illustrate their use with sample calculations, make recommendations drawn from the literature, and list readily available software.

© 2005 Elsevier Inc. All rights reserved.

**Keywords:** Receiver operating characteristic; Evaluation; Test accuracy

## 1. Introduction

Receiver operating characteristic (ROC) curves are frequently used in biomedical informatics research to evaluate computational models for decision support, diagnosis, and prognosis. There are a number of reviews describing the proper use of ROC curves in statistical and medical disciplines [1–4]. ROC analysis is complex, however, with no single analytic approach that is uniformly appropriate in all situations. Rather, the questions of which index to use, how to calculate it, how to estimate its variance and confidence bounds, and how to predict the sample size needed for a proposed experiment all depend on the specific application. In this

review, we briefly discuss the main concepts in ROC analysis relevant to biomedical informatics research. For those interested in this literature, the recent textbooks by Pepe [5] or Zhou et al. [6] cover many of these topics at a level of detail suitable for practicing biostatisticians.

## 2. Background

A basic classification tool in medicine is the binary test (also called a *discrete classifier*), which yields two discrete results (such as *positive* and *negative*), to infer an unknown, such as whether a disease is *present* or *absent*. We can also think of the test as a (possibly imperfect) means to separate a population into two subsets one where the disease is present, and one where it is

\* Corresponding author. Fax +1 617 739 3672.

E-mail address: [machado@dsg.harvard.edu](mailto:machado@dsg.harvard.edu) (L. Ohno-Machado).

absent. The accuracy of these tests is commonly assessed using measures of sensitivity  $SN$  and specificity  $SP$ , where

$$SN = \frac{TP}{TP + FN}, \quad (1)$$

$$SP = \frac{TN}{TN + FP}, \quad (2)$$

and  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  are the counts of true positives, true negatives, false positives, and false negatives, respectively, when the test is applied to a large population. Note that sensitivity depends only on measurements of diseased subjects ( $TP$  and  $FN$ ), and specificity only on healthy subjects ( $TN$  and  $FP$ ), so neither one depends on the prevalence of disease in the test population. For this reason, they are popular measures of test accuracy.

In contrast to a binary test, a continuous test or classifier does not produce discrete results of *positive* or *negative*. Instead, it produces a numeric value on a continuous scale. Without loss of generality, we assume higher values indicate a higher likelihood of disease. To convert an output value to a binary label, we must choose a threshold and compare the output value to that threshold, calling it *positive* if the value exceeds the threshold, and *negative* otherwise. Choosing a high threshold produces a low likelihood of a false positive result (with a consequent high specificity), and a high likelihood of a false negative result (with a consequent low sensitivity). Choosing a low threshold gives the opposite results.

Fig. 1 illustrates a hypothetical continuous test, showing the distributions of results from diseased and healthy populations. A threshold of 0.5 is selected, resulting in a sensitivity of 0.84 and a specificity of

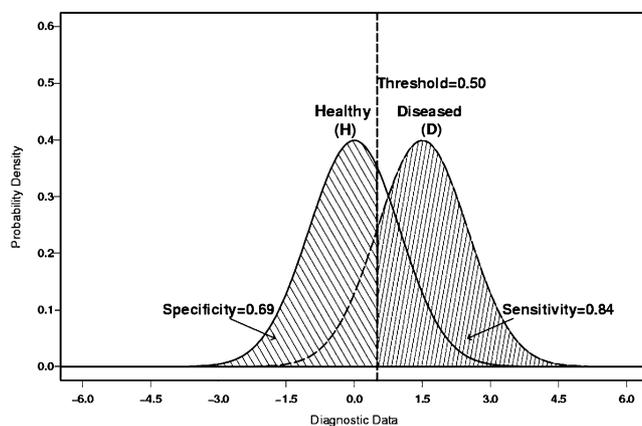


Fig. 1. Distributions of results for a hypothetical continuous diagnostic test on both healthy and diseased populations. A threshold must be chosen to convert the continuous test to a binary test, and the choice of threshold defines a particular sensitivity and specificity.

0.69. A different threshold would result in a different sensitivity and specificity. This figure also illustrates the notion that sensitivity depends only on the diseased population and specificity only on the healthy population.

Ordinal rating tests fall between continuous and binary. These tests produce results from a limited but ordered set of possible outcomes. A common example is a test in which a person is asked to classify items into a category chosen from a relatively small ordered set, such as 1, definitely normal; 2, probably normal; 3, equivocal; 4, probably abnormal; 5, definitely abnormal. On such a five-point rating scale, there are six possible thresholds to choose from if we wish to convert the output to a binary label, and the sensitivity/specificity trade-off is the same as with continuous tests.

For continuous and ordinal tests, therefore, there is no particular value of sensitivity or specificity that characterizes the overall accuracy of the test, but rather an entire range of values that vary depending on what we use as the threshold for discretizing the test result. The ROC curve captures in a single graph the trade-off between a test's sensitivity and specificity over its entire range [7].

The ROC curve plots  $SN$  vs.  $(1 - SP)$  of a test as the threshold varies over its entire range. Each data point on the plot represents a particular setting of the threshold, and each threshold setting defines a particular set of  $TP$ ,  $FP$ ,  $TN$  and  $FN$  counts, and consequently a particular pair of  $SN$  and  $(1 - SP)$  values. In Table 1, hypothetical data representing the results of a 2-h oral glucose tolerance test (OGTT) are presented. The measured plasma glucose values are represented in rows that are sorted in ascending order and subsequently separated into two columns that represent whether the patient was healthy or diseased (in this case, diabetic).

To generate the ROC curve, we consider a threshold (horizontal line) between every neighboring pair of measured values. We add an empty row to the top of the table to account for the case in which the threshold is lower than the smallest measured value. We then calculate the set of  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  counts defined by each threshold. Finally, we calculate the values of  $SN$  and  $1 - SP$  for each set of counts. In Table 1, each row displays the results calculated for a threshold placed between that row and the row below it. All measured values above that row in the table (which are less than the threshold) are considered negative test results, and all measured values below that row in the table (which are greater than or equal to the threshold) are considered positive results.

The count of true positives is the number of positive results that lie in the diseased column, and the count of false positives is the number of positive results that lie in the healthy column. Similarly, the count of true negatives is the number of negative results in the healthy col-

Table 1  
Hypothetical data for a 2-h OGTT and their ROC values

2-h plasma glucose (mmol/L)		Threshold-dependent values*					
Healthy	Diseased	<i>TP</i>	<i>TN</i>	<i>FP</i>	<i>FN</i>	$1 - SP$	<i>SN</i>
		10	0	10	0	1.00	1.00
4.86		10	1	9	0	0.90	1.00
5.69		10	2	8	0	0.80	1.00
6.01		10	3	7	0	0.70	1.00
6.06		10	4	6	0	0.60	1.00
6.27		10	5	5	0	0.50	1.00
6.37		10	6	4	0	0.40	1.00
6.55		10	7	3	0	0.30	1.00
7.29	7.29	9	8	2	1	0.20	0.90
7.82		9	9	1	1	0.10	0.90
	9.22	8	9	1	2	0.10	0.80
	9.79	7	9	1	3	0.10	0.70
	11.28	6	9	1	4	0.10	0.60
	11.83	5	9	1	5	0.10	0.50
12.06		5	10	0	5	0.00	0.50
	18.48	4	10	0	6	0.00	0.40
	18.50	3	10	0	7	0.00	0.30
	20.49	2	10	0	8	0.00	0.20
	22.66	1	10	0	9	0.00	0.10
	26.01	0	10	0	10	0.00	0.00

\* The values for *TP*, *TN*, *FP*, *FN*,  $1 - SP$ , and *SN* on a given line are calculated assuming a threshold that gives a negative prediction for all values in and above that line, and a positive prediction for all values below the line.

umn, and the count of false negatives is the number of negatives that lie in the diseased column.

The values of *SN* and  $(1 - SP)$  are plotted in Fig. 2, with each point on the plot representing a line in Table 1. Overlaid on these points is their estimated ROC curve. The true ROC curve for any test is a continuous function that can only be estimated from the data, in the same sense that a continuous probability distribution

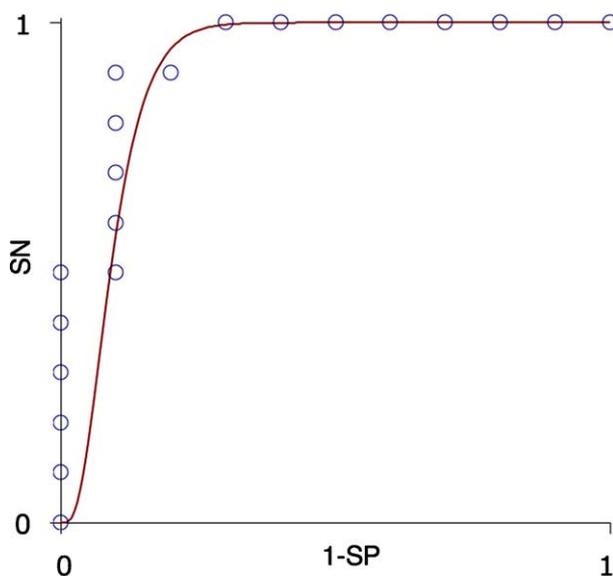


Fig. 2. The ROC curve plotted from the data in Table 1. The data are plotted as circles, with the solid line indicating an ROC curve suggested by the data.

can only be estimated from a finite number of points sampled from that distribution. The problem of estimating a test's ROC curve from a finite set of measured data will be addressed in Section 3.

A curve for a test with perfect accuracy would run vertically from the point (0,0) to the point (0,1) and then horizontally to (1,1) at the top right of the graph. A curve for a test that performed no better than random guessing would run diagonally from (0,0) to (1,1). Curves from real tests typically lie between these two extremes, in the upper left of the plot. If a test produces a curve that lies in the lower right, it means the test is incorrect more often than it is correct. The test could be improved by reversing its labels for positive and negative, which would reflect the ROC curve about the diagonal into the upper left of the plot. Any curve that lies completely above and to the left of another curve represents better test performance.

In this review, we will emphasize curves generated from continuous data since that is how most classifiers work in biomedical informatics. For a review focusing on ordinal data, see [3]. Beyond the simple diagnostic test, any system that attempts to separate two populations on the basis of an ordinal or continuous variable is amenable to ROC analysis. Recent examples of continuous classifiers evaluated using ROC curves include statistical models for retrieving highly useful clinical articles [8], radiology reports that suggest pneumonia [9] and radiology reports suggesting anthrax [10]. These models all produce a score on a continuous scale indicating how well a given document fits the desired category.

Other examples include a model that uses serum proteomic patterns to predict malignancy [11], and a cardiac diagnostic program [12]. To simplify the discussion in this review, we will continue to use our hypothetical simple diagnostic test as the paradigm for continuous classifiers.

### 3. ROC indices of accuracy

While ROC curves themselves are useful in assessing a test, we often desire a single *index* to summarize the accuracy of a test. Several different indices of ROC curves can be calculated, with the most common ones described below.

#### 3.1. Area under the curve

The full area under the ROC curve (AUC) is the most commonly used ROC index [13]. Conceptually, it has several interpretations: (1) the probability that the test will produce a value for a randomly chosen diseased subject that is greater than the value for a randomly chosen healthy subject [14], (2) the average sensitivity for all values of specificity [4], and (3) the average specificity for all values of sensitivity [4]. A perfect test has an AUC of 1.0, whereas random chance gives an AUC of 0.5.

In trying to calculate the AUC from test data, we encounter the problem of inferring the true ROC curve based on a finite sample of data. This problem is analogous to inferring a continuous statistical distribution based on a finite dataset. There are many approaches to making this inference, but we will focus on those more commonly used. A summary of recommendations on when to use each method is given in Table 3.

##### 3.1.1. Nonparametric methods

*The empirical method.* We can approximate the ROC curve by simply connecting the data points ( $SN, 1 - SP$ ) the straight lines, and then calculating the estimated area  $A\hat{U}C$  using the trapezoidal rule. This is referred to as the *empirical* or *nonparametric* method, and the estimated  $A\hat{U}C$  calculated in this way has been shown [14] to be equivalent to the Mann–Whitney  $U$  statistic normalized by the number of possible pairings of diseased and healthy values. It is also known as the two-sample Wilcoxon rank-sum statistic and the c-index [15].

If we let  $d_1, d_2, \dots, d_{n_D}$  be the test values for  $n_D$  diseased subjects, and  $h_1, h_2, \dots, h_{n_H}$  be the test values for  $n_H$  healthy subjects, and define a comparison function  $C(d_i, h_j)$  where

$$C(d_i, h_j) = \begin{cases} 1 & \text{if } d_i > h_j, \\ 0.5 & \text{if } d_i = h_j, \\ 0 & \text{if } d_i < h_j, \end{cases} \quad (3)$$

then the estimated area is the average value of the comparison function for all possible pairs of diseased vs. non-diseased subjects, or

$$A\hat{U}C = \frac{1}{n_D n_H} \sum_{i=1}^{n_D} \sum_{j=1}^{n_H} C(d_i, h_j). \quad (4)$$

Its value can be computed efficiently if the  $d$ 's and  $h$ 's are separated and sorted as shown in Table 2. This table presents the data from Table 1 and lists the partial sums required for the calculation of the  $A\hat{U}C$  via Eq. (4). The numbers in the third column are the results of the comparison function  $C(d_i, h_j)$  for a particular healthy value summed over all diseased values, and the numbers in the fourth column are the results of  $C(d_i, h_j)$  for a particular diseased value summed over all healthy values. The value of 7.5 in the fourth column, for example, comes from the fact that the diseased value of 7.29 is greater than seven healthy values, tied with one healthy value, and less than two healthy values. This gives a sum for that cell of  $7(1) + 1(0.5) + 2(0) = 7.5$ . The normalized total of the values in the third or fourth column is equal to the  $A\hat{U}C$ . It does not matter which column total we use as the area estimate since the two should be equal.

Compared to the other methods presented below, the empirical method has both strengths and weaknesses. It has the advantage of imposing no structural assumptions on the data, and is therefore widely applicable. Additionally, its equivalence to other statistics allows us to apply knowledge about those statistics to ROC analysis (such as when calculating confidence intervals,

Table 2  
Hypothetical data for a 2-h OGTT and the calculation of their  $A\hat{U}C$

2-h plasma glucose (mmol/L)		Partial sums	
$h_j$	$d_i$	$\sum_{i=1}^{n_D} C(d_i, h_j)$	$\sum_{j=1}^{n_H} C(d_i, h_j)$
4.86		10	
5.69		10	
6.01		10	
6.06		10	
6.27		10	
6.37		10	
6.55		10	
7.29	7.29	9.5	7.5
7.82		9	
	9.22		9
	9.79		9
	11.28		9
	11.83		9
12.06		5	
	18.48		10
	18.50		10
	20.49		10
	22.66		10
	26.01		10
$A\hat{U}C = \text{column total}/n_D n_H$		0.935	0.935

Table 3

Summary of recommendations on methods for calculating the  $A\hat{U}C$  for continuous data

If the two distributions are poorly separated ( $A\hat{U}C$  expected to be  $<0.80$ ) and at least one of the two distributions is suspected to be strongly bimodal or of greater complexity, then use either the *empirical method* (simpler) or the *kernel density method* (slightly more accurate).

If the two distributions are well separated ( $A\hat{U}C$  expected to be  $>0.80$ ) or neither distribution is suspected to be strongly bimodal, then use either the *empirical method* or the *binormal method*. If, in addition, the  $n_D$  and  $n_H$  are both moderate to small ( $<100$ ), then the *binormal method* is likely to give tighter asymptotic confidence bounds. If  $n_D$  and  $n_H$  are large ( $>100$ ), then the decision between empirical and binormal methods can be made on the basis of convenience.

discussed below). The main disadvantage of the empirical method is that its  $A\hat{U}C$  is biased downward if there are only a few points on the curve [2,14]. Since all methods come with similar trade-offs, it is not always clear which method will give the best estimate in a given situation.

To address this problem, Faraggi and Reiser [16] conducted Monte Carlo simulations to determine which methods reported the most accurate AUC estimate with a given combination of distribution shape, sample size ( $n_D = n_H = 20$  vs.  $n_D = n_H = 100$ ), and separation of populations (i.e., whether the  $A\hat{U}C$  was low, moderate, or high). They found that no single method produced the best estimate under all conditions. But they found that whichever method produced the best estimate in a particular case, the empirical method usually came in a close second. Specifically, they found that the root mean standard error (RMSE) of the empirical estimate usually differed from the RMSE of the best estimate in only the third decimal place of the AUC, regardless of distribution shape or sample size. This finding supports the use of the empirical method as a robust method for continuous data when diseased and healthy population sizes are at least 20. In many situations, another method may produce a better estimate, but the improvement is likely to be very small.

Confidence intervals can be calculated non-parametrically for the  $A\hat{U}C$  and for the difference of  $A\hat{U}C$ s when comparing two curves [17]. They can be constructed in a number of ways. DeLong et al. [17] have proposed the *nonparametric asymptotic method* which, like the empirical method for calculating the  $A\hat{U}C$ , makes no parametric assumptions of the data (Hanley and Hajian-Tilaki [18] have given an implementation-friendly description of it). The method works by transforming each test value to a *placement value*. The placement value for a measurement from the diseased population is its percentile among the values of the healthy population. Thus, if a diseased value  $d_i$  is greater than 80% of the healthy values, then its placement value  $V_{D_i}$  is 0.80. Similarly, each measurement from the healthy population gives rise to a placement value which corresponds to 1 minus its percentile in the diseased population. So if a healthy value  $h_j$  is greater than 80% of the diseased values, its placement value  $V_{H_j}$  would be 0.20. The variance of the AUC estimate is then the weighted average of the variance of healthy and diseased placements:

$$\text{var}[A\hat{U}C] = \frac{\text{var}[V_H]}{n_H} + \frac{\text{var}[V_D]}{n_D}. \quad (5)$$

To compare two ROC curves generated by two different tests administered to the same subjects, we calculate the difference between the areas, and the variance of that difference. These curves will be partially correlated, and the variance of the difference in area must take into account that correlation. This is done using a covariance term that is the weighted average of the covariance of healthy and diseased placements, or

$$\begin{aligned} \text{cov}[A\hat{U}C_A, A\hat{U}C_B] &= \frac{\text{cov}[V_{HA}, V_{HB}]}{n_H} + \frac{\text{cov}[V_{DA}, V_{DB}]}{n_D} \\ &= \frac{1}{n_H} \left( \frac{1}{n_H - 1} \sum_{j=1}^{n_H} V_{HA_j} V_{HB_j} - \bar{V}_{HA} \bar{V}_{HB} \right) \\ &\quad + \frac{1}{n_D} \left( \frac{1}{n_D - 1} \sum_{i=1}^{n_D} V_{DA_i} V_{DB_i} - \bar{V}_{DA} \bar{V}_{DB} \right), \end{aligned} \quad (6)$$

where  $V_{HA_j}$  is the placement of healthy subject  $j$  using test A. Using this covariance term, the variance of the difference in areas is then

$$\begin{aligned} \text{var}[A\hat{U}C_A - A\hat{U}C_B] &= \text{var}[A\hat{U}C_A] + \text{var}[A\hat{U}C_B] \\ &\quad - 2\text{cov}[A\hat{U}C_A, A\hat{U}C_B]. \end{aligned} \quad (7)$$

Accumetric's AccuROC software [19,37] implements the empirical method for estimating the AUC and comparing correlated curves.

Alternative methods for calculating confidence intervals of empirical estimates use resampling techniques, which we describe in Section 3.1.3.

*Smoothed-curve methods.* A true ROC curve is smooth, but because the empirical method connects the ROC points with straight line segments, the estimated curve is jagged. Therefore, we may improve our estimate of the true curve by simply smoothing the empirical curve without imposing any parametric assumptions on the data.

One approach is to smooth the histograms that give rise to the curves [20]. We smooth a histogram by placing a *kernel density function* at the location of each data point along the horizontal axis and summing the functions. The horizontal scaling of the kernel controls the degree of smoothing, and is specified by a *bandwidth* parameter  $h$ . The area under each kernel function is  $1/n$ , so the summed area is always unity. The estimated smooth histogram is called the kernel estimate  $\hat{f}(\cdot)$ . If

we use the standard normal probability density function  $\phi$  as the kernel function and bandwidth parameter  $h_D$  to smooth a distribution of diseased data, the kernel estimate  $\hat{f}(d)$  for that distribution becomes

$$\hat{f}(d) = \frac{1}{n_D h_D} \sum_{i=1}^{n_D} \phi\left(\frac{d - d_i}{h_D}\right). \tag{8}$$

The smoothed-kernel estimate for the healthy data  $\hat{f}(h)$  would be calculated the same way, using its own  $n_H$  and bandwidth parameter  $h_H$ . The choice of kernel function and bandwidth parameter has been studied in depth in [20] and [21].

The advantage of this method is that it produces a smooth curve, free from parametric assumptions, that can fit arbitrarily complex distributions. Nonparametric confidence intervals can also be calculated with this method although they are usually wider than those calculated parametrically. The main disadvantage of the kernel smoothing method is that it is unreliable near the extremes of the ROC curve or when the histograms are close to zero.

Faraggi and Reiser [16] found that smoothed-curve methods outperformed the competing methods discussed here when the diseased and/or healthy distribution was a bimodal mixture and the two were poorly separated. The empirical method comes in a close second in performance.

### 3.1.2. Parametric methods

For ordinal rating tests, there are generally a small number of points along the ROC curve. In the example above with a set of five possible output values, there would be only six points on the ROC curve regardless of the size of the data set. The downward bias in the non-parametric estimate of the AUC in these instances is likely to be quite large. Using parametric methods to estimate the true ROC curve can reduce the estimated error and increase the statistical power of the study, as long as the modeling assumptions are valid. Since these methods can be sensitive to the accuracy of the assumptions, we should carefully consider the trade-off between the extra power and the validity of the assumptions.

The *binormal method* is the most common parametric method. It assumes that there is a monotonic function that will simultaneously transform both the diseased and healthy data into normal distributions. An ROC curve generated by a distribution that meets this criterion can be completely specified by two parameters  $a$  and  $b$ , where

$$a = \frac{\mu_D - \mu_H}{\sigma_H} \quad \text{and} \quad b = \frac{\sigma_D}{\sigma_H}. \tag{9}$$

The curve itself takes the form

$$SN = \Phi(a + b\Phi^{-1}(1 - SP)), \tag{10}$$

and the AUC is calculated by

$$AUC = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right), \tag{11}$$

where  $\Phi$  is the standard normal cumulative distribution function.

The estimated area  $\hat{AUC}$  is found by substituting the sample means and standard deviations into Eq. (9) and the resulting parameters into Eq. (11). If the data are not transformable to binormal, then the estimate will contain some error, which can be minimized by calculating  $a$  and  $b$  using maximum likelihood estimation.

Fortunately, the binormal model is forgiving of departures from its distributional assumptions. Faraggi and Reiser's simulations [16] demonstrated that the binormal model performs similarly to the nonparametric method for most combinations of sample size, distribution shape, and population separation. It performs significantly worse short, however, when the diseased and/or healthy distribution is a bimodal mixture and the two distributions are poorly separated. Hajian-Tilaki et al. [22] also found this method to perform comparably to the nonparametric method in all cases studied, although they did not consider a case as strongly bimodal as Faraggi and Reiser. A rule of thumb might be that the binormal method is appropriate for all but poorly separated, complex distributions.

Parametric confidence intervals can be found using the method developed by Wieand et al. [23], where

$$\begin{aligned} \text{var}[AUC] = & \frac{1}{n_D + n_H} \left[ \frac{1}{\sigma_D^2 + \sigma_H^2} \left( \frac{\sigma_H^2}{n_H} + \frac{\sigma_D^2}{n_D} \right) \right. \\ & \left. + \frac{AUC^2}{2(\sigma_D^2 + \sigma_H^2)^2} \left( \frac{\sigma_H^4}{n_H - 1} + \frac{\sigma_D^4}{n_D - 1} \right) \right]. \tag{12} \end{aligned}$$

To estimate the variance of  $\hat{AUC}$ , we substitute  $\hat{AUC}$  for AUC in Eq. (12) with the sample means and standard deviations as appropriate.

As in the nonparametric case, additional terms must be calculated to compensate for covariance when comparing areas. If we let  $c_H$  and  $c_D$  be the covariances between tests A and B of the healthy and diseased values, respectively, or

$$c_H = \text{cov}[H_A, H_B] \quad \text{and} \quad c_D = \text{cov}[D_A, D_B] \tag{13}$$

then the variance of the difference is

$$\begin{aligned} \text{var}[AUC_A - AUC_B] = & \frac{1}{n_D + n_H} \left[ \frac{1}{\sigma_A \sigma_B} \left( \frac{c_H}{n_H} + \frac{c_D}{n_D} \right) \right. \\ & \left. + \frac{1}{2} \frac{AUC_A}{\sigma_A} \frac{AUC_B}{\sigma_B} \left( \frac{c_H^2}{n_H - 1} + \frac{c_D^2}{n_D - 1} \right) \right], \tag{14} \end{aligned}$$

where  $\sigma_A^2 = \sigma_{DA}^2 + \sigma_{HA}^2$ , and  $\sigma_{DA}$  and  $\sigma_{HA}$  are the standard deviation of the diseased and healthy values produced by test A. As above, to estimate variance of the difference in area, we use the values of  $\hat{AUC}$  with sample means, standard deviations, and covariances. The ROC-KIT software [24,47] has long been in use for applying

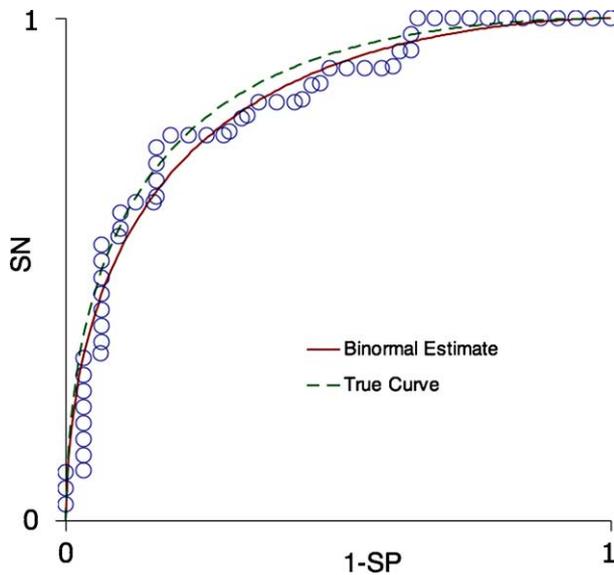


Fig. 3. Estimated ROC curve fitted by the binormal method overlaid with the true binormal curve.

this method to ordinal data, and has now adapted it for use with continuous data.

Both the empirical and binormal methods are illustrated in Fig. 3. The data shown in the figure were sampled from continuous independent distributions, with 30 healthy data points sampled from  $N(0,1)$  and 30 diseased data points from  $N(1.5,1)$ . An empirical curve would directly connect the data points with straight lines, whereas the binormal method estimates the solid curve fit to the data. The true theoretical curve is indicated in dashed lines. The empirical method estimates an  $AUC = 0.834$  (95% CI [0.695, 0.973]), whereas the binormal method estimates  $AUC = 0.831$  (95% CI [0.783, 0.885]). The true theoretical AUC is 0.856. As expected, the binormal and empirical methods estimate nearly identical areas, but the binormal method estimates much tighter confidence bounds, increasing the power of the analysis. The parametric estimate is preferred in this case because the data are from unimodal, moderately separated distributions, and there are a relatively low number of data points. If the distributions were more complex or poorly separated, the parametric bounds may give an inappropriately narrow range, which could lead us to inappropriately reject the null hypothesis of our experiment.

Of course, with only 30 data points from each population, it can be difficult to infer from the data alone how complex the underlying distribution is. These cases require us to apply domain knowledge and good judgement about the nature of the likely distribution. If the disease has several subtypes, for example, then we may suspect a multimodal distribution of the diseased population. If there are other diseases that have some similar features to what we're testing for, then we may suspect the healthy population to have a more complex distribution.

With a larger sample size, the difference between the two confidence bound estimates decreases. Hajian-Tilaki and Hanley [25] compared the parametric and nonparametric confidence interval estimates, and found that for  $n_D = n_H = 100$ , there is no significant difference between the two, even for nonbinormal data.

### 3.1.3. Resampling and other methods

The *resampling methods* are useful for calculating confidence intervals in extreme situations such as with a small sample size or with highly complex or extreme distributions. These methods have the advantage of being widely applicable, and can be used with both parametric and nonparametric AUC estimates, as well as with indices other than the AUC, when direct methods of estimating confidence bounds are unavailable or undesired.

The *jackknife method* [26] uses *pseudovalues*, which are values that can be thought of as representing the “contribution” of a particular data point to the average. An implementation-friendly description is available in [18]. The pseudovalues have the same average as the original data, but they have the advantage of behaving like independent identically distributed samples, so that the standard methods of calculating variances and confidence bounds can be used with them. A pseudovalue  $P_j$  for a particular data point  $x_j$  is calculated by taking the weighted difference of the  $AUC$  using all the data and the  $AUC_{-j}$  generated with all but the point  $x_j$ . More precisely,

$$P_j = (n_D + n_H)AUC - (n_D + n_H - 1)AUC_{-j}. \quad (15)$$

The variance of the  $AUC$  is then simply

$$\text{var}[AUC] = \frac{\text{var}[P]}{n_D + n_H}, \quad (16)$$

and the covariance between two correlated curves is given by

$$\text{cov}[AUC_a, AUC_b] = \frac{\text{cov}[P_a, P_b]}{n_D + n_H}. \quad (17)$$

Comparison of correlated curves can then be done using Eq. (7).

The LABMRMC software [26,47] uses the Jackknife method with ANOVA techniques to compare multiple correlated ROC curves.

Unlike other resampling methods described below, the jackknife is of little use when the sample size is too small for direct methods because it does not increase the number of data points. It only transforms them to values that are easier to work with.

The *bias-corrected and accelerated interval bootstrap method* (BCa) [28] can be used to calculate confidence intervals with small sample sizes or extreme distributions. It is an extension of the *percentile bootstrap method*. In the percentile bootstrap method, the data are resampled with replacement many times, with the index

of interest calculated each time. Confidence intervals are then estimated by simply sorting the data and taking the middle section corresponding to our desired confidence interval bounds. If we used 1000 resamplings, for example, we would sort the resulting  $\hat{AUC}$ s and take the 25th and 975th values as the bounds of the 95% CI. The BCa method resamples in the same fashion, but calculates more accurate confidence intervals using a formula that is a function of the calculated  $\hat{AUC}$  and asymptotic estimates of its variance.

To assess a difference in areas, under ROC curves resampling is done in parallel from both sets of test results, the difference in areas is computed, and this difference is used as the basic element for the rest of the process, which gives an estimate of the variance of area differences.

The *bootstrap t method* [28] functions similarly to the BCa method, except it uses a “studentized pivot statistic” in calculating the CI.

In the unusual event that the  $\hat{AUC}$  equals 1.0, none of the above methods will generate a lower confidence bound less than 1.0, so we must use a different method to estimate the precision of the result. Although we may hope that our test is perfect when its  $\hat{AUC} = 1.0$ , we must realize that if our sample size is small, we may simply have a biased sample. Obuchowski and Lieber [27] performed a Monte Carlo simulation with binormal data to generate recommendations on confidence intervals for this case. They found that for continuous data, with  $n_D$  and  $n_H$  both over 30 and the variance of the populations within a factor of three of each other, the lower 95% confidence bound was near 0.99. For cases where  $n_D$  or  $n_H$  are under 30, they produced tables listing the lower confidence bound, depending on the size of  $n_D$  and  $n_H$  and the ratio of variances.

Unfortunately, no single method has been shown to estimate reliably confidence intervals for all situations. Obuchowski and Lieber [28] have investigated in some detail various methods for estimating  $\hat{AUC}$  variance for both continuous and ordinal data, and have made recommendations for their use. Their investigation assumed strictly normal data with equal variances, but their guidelines for the use of each method depending

on the data’s sample size may give a lower bound for the use of those methods with other distributions.

Table 4 summarizes their recommendations.

### 3.2. Partial area under the curve

While AUC is the most popular ROC index, it is not always the most appropriate to use for the evaluation of a test. If the ROC curves of two different tests cross at some point, then the full area under the curve may not be the best performance indicator. When deciding between two medical tests, for example, where the test must produce a specificity above 0.8 to be useful, we want to select the test that has superior performance in this useful range. This may or may not be the curve with the largest AUC if the two curves cross.

In such cases, the partial area under the curve (pAUC) may be a more meaningful index. It can be calculated both parametrically (using binormal assumptions) [29] and nonparametrically [30] with advantages and disadvantages similar to their full-area counterparts.

To aid in interpretation of the pAUC, we can divide by the width of the interval, giving an index with a maximum value of unity. Other transformations can give a partial area index that behaves more like the AUC, with a minimum value of 0.5 and a maximum of 1.0 [29].

The confidence bounds for pAUC (or any transformation thereof) can be easily calculated using one of the resampling methods. Alternatively, Zhang et al. [30] developed an analytic expression for nonparametric variance that is analogous to DeLong’s nonparametric expression for full areas. Similarly, McClish [29] described a method to calculate the variance of a partial area estimate using the binormal parameters. Both of these articles also describe methods to compare partial areas of correlated curves.

A limiting case of the pAUC is where the width of the interval is zero, and the normalized pAUC becomes the sensitivity of a test at a given specificity. If we choose a high value of specificity (say 95%) and call that the upper range of normal for the healthy patients, then the test’s sensitivity at that point is the fraction of diseased patients whose test result is higher than the normal range.

Table 4

Recommendations on methods for calculating confidence intervals of an AUC estimate of continuous normal distributions of identical variance (from [28])

---

*For non-parametric estimates*

If the  $\hat{AUC}$  is high ( $\geq 0.95$ ) and  $n_D$  and  $n_H$  are both large ( $>120$ ), then use *the asymptotic method*; otherwise use *the BCa method*.

If the  $\hat{AUC}$  is moderate (0.80–0.95) and  $n_D$  and  $n_H$  are both moderate (30), then use *the asymptotic method*.

Otherwise, if  $n_D$  and  $n_H$  are similar, then use *the BCa method*; if only one is small ( $<20$ ), use *the bootstrap t method*.

*For parametric estimates*

If the  $\hat{AUC}$  is high ( $\geq 0.95$ ), use *the asymptotic method* if  $n_D$  and  $n_H$  are both large ( $>150$ ); otherwise use *the bootstrap t method*.

If the  $\hat{AUC}$  is moderate (0.80–0.95) and  $n_D$  and  $n_H$  are both moderate ( $>30$ ), use *the asymptotic method*.

Otherwise, if  $n_D$  and  $n_H$  are similar, use *the BCa method*; if only one is small ( $<20$ ), use *the bootstrap t method*.

---

The variance of the sensitivity at a given specificity must take into account uncertainty in both the sensitivity and specificity, which makes it dependent on both the diseased and non-diseased populations. A parametric estimate of this variance is given by Obuchowski et al. [31], and a nonparametric estimate by Pepe [32].

#### 4. Sample size calculations

The calculation of sample size for a given experiment that will use ROC analysis is a complex question. The sample size estimate depends on the index of interest in the evaluation and whether we are evaluating a single test against a hypothesized fixed standard (such as random guessing or a current standard test) or against another test in the same experiment. Common methods invoke binormal and/or asymptotic assumptions that are generally reasonable given the approximate nature of these calculations. We will give here an estimate for the most common scenario, and reference other work for the remainder.

From Pepe [33], the number of diseased subjects  $n_D$  needed to test the hypothesis that  $A\hat{U}C_a$  is different from  $A\hat{U}C_b$  is

$$n_D = (\kappa \text{var}_D + \text{var}_H) \left[ \frac{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)}{A\hat{U}C_a - A\hat{U}C_b} \right]^2, \quad (18)$$

where  $\kappa = n_D/n_H$ ,  $\alpha$  is the desired minimum type 1 error rate,  $\beta$  the desired minimum type 2 error rate,  $\text{var}_D$  is the anticipated variance of the test on diseased subjects,  $\text{var}_H$  that of the healthy subjects, and  $\Phi^{-1}$  is the inverse standard normal cumulative distribution function (which we must look up in a table). Obuchowski [34] notes that when choosing an expected area difference in the absence of pilot data, investigators often err by assuming a difference that is too large. This is tempting because a much larger dataset is needed to resolve the smaller difference. The consequence of such an error is likely to be an underpowered study. A good guideline to use when uncertain about expected area differences is to consider the difference in terms of sensitivity gain at a given specificity. An area difference of 0.1, for example, translates to a gain in sensitivity of 0.2 to 0.33, depending on the specificity at which the test operates. This is a much larger difference than would be expected in most studies. Sample size estimates for the other indices and comparisons mentioned are given in [5] and [35].

Reporting the estimated optimum sample size is important, even if the calculated size is not achieved in the actual study. If the optimum sample size is not achieved, report the power calculation for the *actual* sample size to detect the desired difference in indices. These estimates are essential in interpreting a negative result.

#### 5. Software

A number of software packages are available for ROC analysis, as discussed above. We now summarize them in Table 5, along with the methods they use for calculating ROC area, confidence intervals and curve comparisons. Stephan et al. [36] rated them in terms of correctness, completeness, and ease of use. They found that no single package was adequate for all needs, but all performed correct calculations given their assumptions and methods. The last entry in the table, ROC-KIT, is the product of a research laboratory that has been contributing free ROC analysis software to the statistical community for decades. Using pre-coded software takes much of the work out of ROC analysis, but it comes with pitfalls if the software is misused. Before relying on any software package, its methods, limitations, assumptions should be noted.

#### 6. An example

There are many examples of the use of ROC curves in biomedical informatics. We present here an example by Lu et al. [49] on the accuracy of several different machine learning models on preoperative prediction of ovarian tumor malignancy. Data from the preoperative evaluation of 265 patients with a persistent extrauterine pelvic mass were used to train six different models to predict whether the mass was malignant or benign. The models were built using logistic regression (LR) and least-squares support vector machines using the linear ( $\text{SVM}_{\text{Lin}}$ ) and radial basis function ( $\text{SVM}_{\text{RBF}}$ ) kernels. These were each trained using two different sets of input variables taken from the set of cases. The reference standard was postoperative histological examination.

Each of the six models was tested on a common data set of 160 subsequently collected cases. Additionally, the simpler Risk of Malignancy Index (RMI) was calculated for each patient and reported for comparison. An ROC curve was produced for each model and for the RMI, with the  $A\hat{U}C$  calculated using the empirical method. Standard errors for each  $A\hat{U}C$  and the difference between  $A\hat{U}C$ s of models were calculated using the Delong method [17] for correlated curves (though only the  $p$  values, rather than the standard errors of the differences, were reported). The best performing model was an  $\text{SVM}_{\text{RBF}}$ , which produced an  $A\hat{U}C$  of 0.922 (SE 0.021). Pairwise comparisons between all seven methods were not performed, but the SVM models were compared to the LR models and the RMI. The difference in  $A\hat{U}C$  between the best  $\text{SVM}_{\text{RBF}}$  and the two LR models was 0.006 ( $p = 0.4$ ) and 0.011 ( $p = 0.3$ ). The difference between  $\text{SVM}_{\text{RBF}}$  and RMI was 0.049 ( $p = 0.05$ ). Thus, the investigators did not detect a difference between the SVM and LR models, but they did find  $\text{SVM}_{\text{RBF}}$  to be more accurate than the

Table 5  
Selected ROC software and their self-identified methods for calculating the  $\hat{AUC}$ , confidence intervals, and correlated curve comparison

Name (type)	Methods			Notes
	$\hat{AUC}$	Confidence intervals	Curve comparison	
AccuROC [19,37] (commercial)	Empirical	Nonparametric asymptotic, BCa, other bootstrap	Nonparametric calculated correlation correction [17]	For Microsoft Windows. Includes evaluation measures other than ROC curves.
Analyse-It [38] (commercial)	Empirical	Parametric [13]	Correlation correction Table [43]	For Microsoft Windows. Microsoft Excel add-in. Includes statistics other than ROCs.
CMDT [39] (free)	Parametric and nonparametric	Bootstrap and other	DFPT* [40]	For Microsoft Windows. Microsoft Excel add-in. Developmental.
GraphROC [41,42] (free <sup>a</sup> /commercial)	Empirical	Nonparametric asymptotic	Correlation correction Table [43]	For Microsoft Windows. Allows comparison of pAUC or sensitivity at given specificity.
MedCalc [44,45] (commercial)	Empirical	Parametric [13]	Correlation correction Table [43]	For Microsoft Windows. Includes statistics other than ROCs.
LABMRMC [26,47] (free)	Binned binormal, with empirical for degenerate data [46]	Jackknife	Jackknife	For Microsoft Windows or Macintosh. Developed to analyze multi-reader multi-case ratings data.
ROCKIT [24,47,48] (free)	Binned binormal	Binned binormal asymptotic	Binormal parametric correlation correction	For Microsoft Windows or Macintosh. Supercedes previous versions of ROCFIT, LABROC, CORROC2, CLABROC, and INDROC.

<sup>a</sup> GraphROC 1996 version planned to be released as freeware with new version commercial.

\* DFPT, distribution-free permutation test.

RMI by a clinically relevant margin. In this experiment, they did not adjust for multiple comparisons.

The investigators then performed a second experiment because of suspected bias towards more difficult cases in the test set of the first experiment, given that the test set was constructed from cases that came later in time. They combined all 425 cases, and randomly divided them into 265 training cases and 160 test cases, fixing the ratio of malignant to benign cases in each division. They trained each model on the resulting training set and calculated the AUC for each model on the test set. They repeated this 30 times and reported the mean and standard deviation of the AUCs from the 30 repetitions. They used one-way ANOVA followed by Tukey multiple comparison to determine significant differences. As with the first experiment they found differences significant at the 95% confidence level between RMI and the six machine learning models, but found no difference at this level between the six models.

This example illustrates many of the points we have raised in this review. The investigators appropriately chose ROC analysis for evaluation of continuous classifiers, and selected an appropriate method for calculating the  $\hat{AUC}$ , its confidence intervals, and comparison between curves. They reported  $p$  values for those comparisons, although reporting the confidence intervals

would have been more informative. A sample size or power calculation also would have been helpful in interpreting the lack of difference detected between models.

## 7. Summary of recommendations

Listed below is the summary of our recommendations on using ROC analysis in the evaluation of continuous classifiers in biomedical informatics.

1. Select the ROC index for evaluation (such as full AUC, partial AUC, sensitivity at a given specificity, other), depending on the range of interest of the classifier being evaluated, as outlined in Section 3.2.
2. Estimate the smallest relevant difference in the selected index, and calculate the optimum sample size to detect that difference using methods outlined in Section 4. Report this sample size, even if it is not achieved, and the power of the sample size that was achieved (if different from the optimum).
3. Select and report the method of estimating the index (empirical, binormal, smoothed-curve, etc.), as outlined in Sections 3.1 and 3.2, and summarized in Table 3.

4. Report confidence intervals of the index estimates and the method of calculating them, as outlined in Sections 3.1 and 3.2 and summarized in Table 4, with special attention for the case of an  $AUC = 1.0$ .
5. If comparing correlated curves, report the confidence interval of the difference in indices, as outlined in Sections 3.1 and 3.2 and summarized in Table 4.  $P$  values derived from these intervals may also be reported, if desired.
6. Cite any software used to perform any part of the analysis.

## 8. Conclusion

We have presented here some basic concepts of ROC analysis, and have made recommendations drawn from the literature on properly performing such an analysis. An appropriate ROC analysis, however, is only one element of a valid and meaningful study. The careful construction of the reference standard, for example, is crucial in any evaluation of test accuracy. Spectrum, verification, selection, and incorporation biases, and random errors in the reference standard and in the test's input data can affect both the correctness and the wider applicability of our conclusions. Whiting et al. [50] have performed a systematic review of these and many other sources of error, and they discuss the general effect that each has been found to have on test accuracy evaluation.

The concepts we have reviewed are only the basics of ROC evaluation. There are many more topics for the interested reader to explore, such as the methods of combining multiple ROC curves for a meta-analysis [51], evaluating classifiers that predict more than two alternatives [52–54], and applying ROC analysis to tests in a clustered environment [55], or for tests repeated over time to monitor for occurrence of an event [56].

## Acknowledgments

This work was supported in part by NIH Grant NIHR01LM007861 and NLM Training Grant T15LM07092.

## References

- [1] Shapiro DE. The interpretation of diagnostic tests. *Stat Methods Med Res* 1999;8(2):113–34.
- [2] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39(4):561–77.
- [3] Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* 2003;229(1):3–8.
- [4] Metz CE. ROC methodology in radiologic imaging. *Invest Radiol* 1986;21(9):720–33.
- [5] Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford; 2003.
- [6] Zhou XH, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. New York: Wiley; 2002.
- [7] Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240(4857):1285–93.
- [8] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high quality article retrieval in internal medicine. *J Am Med Inform Assoc* 2004.
- [9] Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. *J Biomed Inform* 2001;34(1):4–14.
- [10] Chapman WW, Cooper GF, Hanbury P, Chapman BE, Harrison LH, Wagner MM. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. *J Am Med Inform Assoc* 2003;10(5):494–503.
- [11] Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA. Data mining techniques for cancer detection using serum proteomic profiling. *Artif Intell Med* 2004;32(2):71–83.
- [12] Fraser HS, Long WJ, Naimi S. Evaluation of a cardiac diagnostic program in a typical clinical setting. *J Am Med Inform Assoc* 2003;10(4):373–81.
- [13] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29–36.
- [14] 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.
- [15] Harrell Jr FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;247(18):2543–6.
- [16] Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Stat Med* 2002;21(20):3093–106.
- [17] 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(3):837–45.
- [18] Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Acad Radiol* 1997;4(1):49–58.
- [19] <http://www.accumetric.com>. Accessed October 19, 2004.
- [20] Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Stat Med* 1997;16(19):2143–56.
- [21] Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *J Am Stat Assoc* 1998;93(444):1356–64.
- [22] Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Med Decis Making* 1997;17(1):94–102.
- [23] Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989;76(3):585–92.
- [24] Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Stat Med* 1998;17(9):1033–53.
- [25] Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. *Acad Radiol* 2002;9(11):1278–85.
- [26] Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Invest Radiol* 1992;27(9):723–31.
- [27] Obuchowski NA, Lieber ML. Confidence bounds when the estimated ROC area is 1.0. *Acad Radiol* 2002;9(5):526–30.

- [28] Obuchowski NA, Lieber ML. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Acad Radiol* 1998;5(8):561–71.
- [29] McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989;9(3):190–5.
- [30] Zhang DD, Zhou XH, Freeman Jr DH, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Stat Med* 2002;21(5):701–15.
- [31] Obuchowski NA, Lieber ML, Wians Jr FH. ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem* 2004;50(7):1118–25.
- [32] Pepe MS. In: The statistical evaluation of medical tests for classification and prediction. New York: Oxford; 2003. p. 100.
- [33] Pepe MS. In: The statistical evaluation of medical tests for classification and prediction. New York: Oxford; 2003. p. 226.
- [34] Obuchowski NA. Determining sample size for ROC studies: what is reasonable for the expected difference in tests' ROC areas? *Acad Radiol* 2003;10(11):1327–8.
- [35] Obuchowski NA. Computing sample size for receiver operating characteristic studies. *Invest Radiol* 1994;29(2):238–43.
- [36] Stephan C, Wesseling S, Schink T, Jung K. Comparison of eight computer programs for receiver-operating characteristic analysis. *Clin Chem* 2003;49(3):433–9.
- [37] Vida S. A computer program for non-parametric receiver operating characteristic analysis. *Comput Methods Programs Biomed* 1993;40:95–101.
- [38] <http://www.analyse-it.com/products/clinical/diagnostic-testing.htm>. Accessed October 19, 2004.
- [39] <http://city.vetmed.fu-berlin.de/~mgreiner/CMDT/cmdt.htm>. Accessed October 19, 2004.
- [40] Venkatraman ES, Begg CB. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* 1996;83:835–48.
- [41] <http://www.netti.fi/~maxiw/>. Accessed January 20, 2005.
- [42] Kairisto V, Poola A. Software for illustrative presentation of basic clinical characteristics of laboratory tests—GraphROC for Windows. *Scand J Clin Lab Invest Suppl* 1995;222:43–60.
- [43] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148(3):839–43.
- [44] <http://www.medcalc.be>. Accessed October 19, 2004.
- [45] Schoonjans F, Zalata A, Depuydt CE, Comhaire FH. MedCalc: a new computer program for medical statistics. *Comput Methods Programs Biomed* 1995;48:257–62.
- [46] Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest. Radiol* 1989;24:234–45.
- [47] [http://www-radiology.uchicago.edu/krl/KRL\\_ROC/software\\_index.htm](http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index.htm). Accessed October 19, 2004.
- [48] Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially paired datasets. *Med Decis Making* 1998;18:110–21.
- [49] Lu C, Van Gestel T, Suykens JA, Van Huffel S, Vergote I, Timmerman D. Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines. *Artif Intell Med* 2003;28(3):281–306.
- [50] Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140(3):189–202.
- [51] Kester AD, Buntinx F. Meta-analysis of ROC curves. *Med Decis Making* 2000;20(4):430–9.
- [52] Hand DJ, Till RJ. A simple generalization of the area under the ROC curve for multiple class classification problems. *Mach Learn* 2001;45:171–86.
- [53] Mossman D. Three-way ROCs. *Med Decis Making* 1999;19(1):78–89.
- [54] Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. *Med Decis Making* 2000;20(3):323–31.
- [55] Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics* 1997;53(2):567–78.
- [56] Parker CB, DeLong ER. ROC methodology within a monitoring framework. *Stat Med* 2003;22(22):3473–88.