Published in final edited form as:

Acad Radiol. 2013 July ; 20(7): 854–862. doi:10.1016/j.acra.2013.02.009.

An additive selection of markers to improve diagnostic accuracy based on a discriminatory measure

Liansheng Larry Tang¹, Le Kang², Chunling Liu³, Enrique F Schisterman⁴, and Aiyi Liu^{4,*}

¹Department of Statistics, George Mason University, Fairfax, VA 22030, USA

²Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD 20993, USA

³Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

⁴Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Blvd, Bethesda, MD 20892, USA

Abstract

Rationale and Objectives—The estimation of the area under the receiver operating characteristic (ROC) curve (AUC) often relies on the assumption that the truly positive population tends to have higher marker results than the truly negative population. The authors propose a discriminatory measure to relax such an assumption and apply the measure to identify the appropriate set of markers for combination.

Materials and Methods—The proposed measure is based on the maximum of the AUC and 1-AUC. The existing methods are applied to estimate the measure. The subset of markers are selected using a combination method which maximizes a function of the proposed discriminatory score with the number of markers as a penalty in the function.

Results—The properties of the estimators for the proposed measure were studied through large-scale simulation studies. The application was illustrated through a real example to identify the set of markers to combine.

Conclusion—Simulation results showed excellent small-sample performance of the estimators for the proposed measure. The application in the example yielded a reasonable subset of markers for combination.

Keywords

Receiver operating characteristic (ROC); Area under the ROC curves (AUC); Discriminatory score; Box-Cox transformation

*liua@mail.nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. INTRODUCTION

The receiver operating characteristic (ROC) curves have become an important tool for evaluating the accuracy of diagnostic markers [1, 2, 3, 4] and have been found useful in many areas of scientific research such as diagnostic imaging [5, 6], signal detection, and biometric identification [7, 3]. The ROC curve of a marker is obtained by plotting, over all possible values of the marker results, the true positive rate (or sensitivity), the probability that the marker correctly identifies a truly positive subject, versus the false positive rate (or 1-specificity), the probability that the marker incorrectly identifies a truly negative subject.

The area under the ROC curve (AUC) has been recognized as an important measure of the accuracy of the marker [8]. An AUC value close to 1 indicates high discriminatory power of a diagnostic marker, and a value of 0.5 suggests that a marker has a diagnostic ability no better than tossing a fair coin. Many authors developed methods to estimate the AUC, among which some are nonparametric [9, 10], some are parametric [11, 12, 13], and some are semiparametric [14, 15, 16, 17].

In general, a diagnostic marker is measured on a number of independent truly positive and truly negative subjects. A measurement (presumably on a continuous scale) from each subject is observed. These measurements are then used to construct a sample ROC curve and the area under it is used as an estimate of the true area under the population ROC curve. A number of parametric and nonparametric methods have been developed to estimate the AUC [8, 5, 14]. Most parametric methods are based on normal or gamma distributions, and the nonparametric methods are based on the empirical ROC curve or some smoothed version of it [14].

A major goal of microarray experiments in the study of cancer is to discover gene profiles that can distinguish subjects with different conditions (e.g. normal subjects versus cancer patients, cancer patients with different stages, responders and nonresponders to a newly developed drug, etc.). Because microarray experiments produce the expression of thousands of markers for each subject, one of the most challenging issues in these studies is inevitably reduction of data dimension and selection of individual informative markers that relate to these conditions. Since the AUC has been proved to be a useful index for evaluating discriminatory accuracy in diagnostic medicine, it is worth exploring its usage in searching for informative markers and reducing data dimensions.

In Section 2 we briefly review the contents of the ROC curves and the AUC. We define a useful index, *discriminatory score*, and explore its properties and applications in dimension reduction. In Section 3, we apply the method to the problem of classification of autism children from a study conducted by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. Some discussions are in Section 4.

2. MATERIALS AND METHODS

2.1. Brief Review of Receiver Operating Characteristic (ROC) Curves

The ROC curve is a composite function of the distribution functions of the truly positive and truly negative populations. The textbooks by Zhou et al., Pepe and Zou et al. [18, 19, 20] provide excellent overview of common statistical methods used for analyzing the ROC curve. The area under the ROC curve (AUC) is one of the most frequently used measures to summarize the ROC curve.

Consider a single marker, whose measurement is X from a truly positive subject and Y from a truly negative subject. Assume a larger measurement is more likely to indicate the positive status. Bamber [21] showed that the AUC is given by $A = Pr(X > Y)$, the probability of correct ordering between X and Y . An empirical estimate of A is equivalent to a Wilcoxon rank-sum statistic [5]. Parametric methods to estimate the AUC are commonly based on the normal distribution assumptions for marker results [11].

Since the AUC can be written as a function of only two parameters as indicated in Metz et al. [22] and Dorfman and Alf [13], novel methods are proposed by these authors to estimate the parameters. Their methods are originally proposed for ordinal ratings which arise as radiologists rate medical images according to their belief of lesion existence in the images. For example, the rating may be on a five-point scale, with 1 indicating absolute normal, 2 indicating somehow normal, 3 indicating not sure, 4 indicating somehow abnormal, and 5 indicating absolute abnormal. Later, Metz et al. [12] extended the methods to account for continuous ratings which are yielded by applying computer algorithms on medical images. Additional examples of continuous ratings in radiology can be found in magnetic resonance imaging of brain tumors and spiral computed tomography of ureteral stone sizes [23, 24].

In practice, continuous marker results may not follow normal distributions, and a monotone transformation is needed to transform the original data so that the transformed data approximately follow normal distributions. The ROC curve estimated from the transformed data should remain the same as the one obtained from original data due to its transformation invariant property. The Box-Cox transformation has been recommended by Kramar et al. [25] and Schisterman [26] before applying the normality based approaches to estimate the AUC.

2.2. Discriminatory Measure of Markers

Since it is usually unknown whether the measurement is higher in the X s than in the Y s, we define an index, i , as $i = \max \{A, 1 - A\}$. We call hereafter i the *discriminatory score of the marker*. It measures how accurate it can be for the marker to distinguish the truly positive from the truly negative. Its estimator can be obtained by replacing A with some nonparametric or parametric estimator $\hat{A}$: $\hat{i} = \max \{\hat{A}, 1 - \hat{A}\}$. The detailed expressions of the estimators are provided in the appendix. The aforementioned semi-parametric methods [14, 15, 16, 17] can also be applied to estimate i , but they are not considered in this article.

2.3. Selection of a Group of Markers with High Discriminatory Power: An Additive Procedure

The objective here is to select a smaller set of markers that provide a satisfactory power in distinguishing truly positive subjects from truly negative subjects. Consider K markers. Let X_{ki} be the measurement of marker k in a truly positive subject i , $i = 1, \dots, m$, and Y_{kj} be that in a truly negative subject j , $j = 1, \dots, n$, for $k = 1, \dots, K$. Suppose that after the Box-Cox transformations, the measurements on the same truly positive subject or the same truly negative subject follow multivariate normal distributions whose expressions are given in the appendix. When measurements for each marker are obtained from the same set of subjects, the measurements across markers on the same subject are correlated. This correlation results in non-zero correlation coefficients in the covariance matrices of the multivariate normal distributions. Sometimes, when the ratings are obtained on separate sets of subjects for different markers, the correlation coefficients are zero and do not need to be considered in the matrices.

Let the linear combination of these K markers be $L(X) = \lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_K X_K$ for truly positive subjects, and $L(Y) = \lambda_1 Y_1 + \lambda_2 Y_2 + \dots + \lambda_K Y_K$ for truly negative subjects. Denote $\lambda = (\lambda_1, \dots, \lambda_K)^T$. Denote the AUC based on $L(X)$ and $L(Y)$ as $A_K = \Pr\{L(X) > L(Y)\}$. The discriminatory score of the combined marker is given by $i_K = \max\{A_K, 1 - A_K\}$. Following the results in Su and Liu [27], the coefficients for the optimal linear combination to maximize i_K are the same as those for maximizing the AUC. The explicit expression of the coefficients for the optimal combination is provided in the Appendix. The coefficients are a function of the covariance matrices and mean vectors from two groups of measurements. In practice, true means and covariance matrices are unknown. It is recommended in Su and Liu [27] to substitute sample means for the true means and substitute sample covariance matrices for the true covariance matrices in the calculation.

However, the more markers are added in the combination, the larger the discriminatory score is likely to be. If some marker contributes to little increment of the discriminatory score, it may seem unnecessary to researchers to add this marker to the marker combination given the additional cost and complexity of measuring the marker. We propose to use an objective function to reward a large discriminatory score and penalize the number of markers. The objective function is defined as $f(i_k, k)$, a joint function of the discriminatory score based on the optimal linear combination of k markers and the number of markers k . The objective function is an increasing function of the discriminatory score and a decreasing function of the number of markers. It is of no practical interest to add markers if the increment of the objection function by including additional markers is no longer larger than some threshold a_k . Usually the threshold a_k can be defined by medical experts. If no expertise is available, one can let $a_k = 0$ and define $f(i_k, k)$ accordingly. Some examples of the objective function are $f(i_k, k) = i_k + 1/k$, $f(i_k, k) = i_k + 1/k^2$, or $f(i_k, k) = i_k - \delta \log_{10} k$, where δ is a tuning parameter.

In our marker selection procedure, we add one marker at a time. At the first stage, the marker with the largest score out of K markers is chosen as the starting marker. We denote the marker as s_1 and its score for the objective function as $f(i_1, 1)$. At k th stage, $k = 2, \dots, K$,

we choose one marker from the rest of markers and combine it with markers $s_1, s_2, \dots, s_{k-1}$ using the optimal combination in Equation (3) in the appendix. The scores of the objective function are then calculated using Equation (4) in the appendix. The largest of these scores, denoted as $f(i_k, k)$, is chosen and the corresponding marker, denoted as s_k , is added to our marker pool. If $f(i_k, k) - f(i_{k-1}, k-1) \leq a_k$, we stop adding markers and conclude with the best marker selection; otherwise, we continue our combination procedure by using the same criteria. At the K th stage, we compare the value of the objective function by combining all markers, $f(i_K, K)$, with $f(i_{K-1}, K-1)$. If $f(i_K, K) - f(i_{K-1}, K-1) \leq a_K$, we stop without adding the last marker; otherwise, we include all markers in the model.

The other way to select the best subset of markers is to choose all $2^K - 1$ possible subsets of markers. The best combinations are obtained for all subsets. The values of the objective function from these $2^K - 1$ combinations are then computed and ranked. The subset of markers with largest values is selected. This would be computationally intensive when the number of markers is large, which is common in gene expression studies. Compared to this full selection version, the computational task can be greatly simplified in our procedure. Suppose we consider the computation of best combinations. If there is no early stopping in our procedure, we merely need to compute $K(K+1)/2$ best combinations, much less than $2^K - 1$ combinations required by the full selection procedure. Clearly, when K gets larger or early stopping occurs, the computational advantage of our procedure becomes more obvious.

2.4. Description of the Simulation Studies

Simulation studies are conducted to evaluate the accuracy and consistency of our proposed *discriminatory score* measure in terms of bias and root mean square error (RMSE). Normal, lognormal, mixture normal and mixture lognormal data for truly negative and truly positive populations are generated with different parameters and sample sizes. For normal data, parametric and nonparametric AUC estimators are applied. For lognormal, mixture normal and mixture lognormal data, parametric AUC estimator is applied after the data is transformed by the Box-Cox type transformation, while nonparametric AUC estimator is applied to the original data. This is because the nonparametric estimator concerns the order of the data and remains the same regardless of whether data is transformed. True AUC values for various simulation configurations are calculated. Bias and RMSE are calculated with 10000 Monte Carlo samples for each simulation configuration. Here, bias is defined as the difference between the average of estimated AUCs across 10000 Monte Carlo samples and the true AUC, and RMSE is defined as square root of average squared difference between estimated AUCs and true AUC.

From Tables 2–5, it is showing that as mean separation between truly positive and truly negative subjects get larger, the bias and the RMSE get smaller. Meanwhile, the bias and RMSE tend to increase with the increased standard deviations. With normal, lognormal and mixture normal data, the parametric AUC estimator usually has smaller biases and RMSEs than the nonparametric AUC estimator. However, with mixture lognormal data that could not be transformed to normal, the nonparametric AUC estimator is better. The differences between the parametric AUC estimator and its nonparametric alternative seem to vanish as sample sizes get large. In other words, parametric and nonparametric AUC estimators are

asymptotically equivalent. Nevertheless, for normal data or data that can be transformed to normal, our recommendation is to use Equation (2) in the Appendix on the original data or the Box-Cox transformed data to estimate the proposed *discriminatory score* measure. For data that can not be transformed to normal, it is better to use the nonparametric estimator as in Equation (1). It is advised to explore data, e.g., visually diagnose via histogram, before deciding whether to apply our method.

2.5. Description of Real Data Analysis

The Autism/ASD Study was conducted by the Eunice Kennedy Shriver National Institute of Child Health and Human Development in 2002–2005. In the study, truly positive subjects are children with autism/ASD age 4–8 years and truly negative subjects are matched control children. The patient records of the Kelly O’Leary Center for Autism Spectrum Disorders at Cincinnati Children’s Hospital Medical Center were used to identify potential autism/ASD children. The study enrolled 81 subjects diagnosed with autism/ASD and age-matched 80 truly negative children. Their blood samples were obtained and markers including insulin-like growth factors (IGF-1, IGFII), insulin-like growth factor-binding protein (IGFBP-3), and dehydroepiandrosterone (DHEA) were measured. The detailed descriptions of subject enrollment and data collection are given by Mills et al. [28] and Liu et al. [29].

3. RESULTS

We use the Autism/ASD example to illustrate the application of the discriminatory score measure and the proposed marker selection method which may yield a desired subset of markers for combination. Our analysis excludes four true positive cases due to the lack of their blood samples.

The histograms of original measurements of these marker show that for some biomarkers, the results are right skewed. Thus, transformation is necessary for the data to be approximately normal. We use the Box-Cox transformation on the original data. We then consider the proposed additive procedure for marker selection. First, we use Equation (2) in the appendix for the estimation of discriminatory score $\hat{i}_k$ of each marker based on the transformed data. The estimated $\hat{i}_k$ ’s are included in Table 1. Since the Box-Cox type transformation is applied before the combination, the new marker is a linear combination of transformed marker results.

In this example, we consider an objective function, $f(i_k, k) = i_k - \delta \log_{10} k$, a joint function of the discriminatory score based on the optimal linear combination of k markers i_k and the number of markers k , wherein δ is a tuning parameter. The criteria to stop for the additive procedure is when $f(i_k, k) - f(i_{k-1}, k-1) \leq 0$.

The penalizing part $-\delta \log_{10} k$ is 0 when there is only 1 marker (no penalization). For $\delta = 0.01$, we obtain a subset of markers {IGFII, IGFBP3, DHEA} with a combination of

$$0.0255 \times \text{IGFII} + 0.5196 \times \text{IGFBP3} - 0.2576 \times \text{DHEA}$$

and estimated $\hat{t}_k = 0.7768$. To put more penalty on the number of markers, we set $\delta = 0.05$ and only the first marker IGFII is selected with estimated $\hat{t}_k = 0.7659$. Alternatively, less penalty can be considered so that $\delta = 0.005$, and we include all markers in the selection. The resulting combination of all markers {IGFII, IGFBP3, DHEA, IGF1} is

$$0.0248 \times \text{IGFII} + 0.4514 \times \text{IGFBP3} - 0.2681 \times \text{DHEA} + 0.0069 \times \text{IGF1}$$

with estimated $\hat{t}_k = 0.7777$.

We plot the fitted binormal ROC curves and empirical ROC curves for both individual marker and the combined marker with $\delta = 0.01$ and $\delta = 0.005$ in Figures 1 and 2, respectively. We observe that from the estimated linear combination coefficients associated with $\hat{t}_k$ that IGF1 contributes little to the combined marker. Using our proposed discriminatory score in the additive selection procedure, we may arrive at a subset of markers that can be linearly combined to achieve satisfactory performance in distinguishing truly positive subjects from truly negative subjects. Compared to the full set of markers, we essentially lose little discriminating power but save the cost of measuring additional markers, e.g., IGF1. Also it is worthwhile to note that our additive selection procedure provides the flexibility to include more or less markers for cost-performance tradeoff.

4. DISCUSSION

In ROC studies, measurements from truly positive subjects are assumed to be higher than those from truly negative subjects so that the AUC is guaranteed to be greater than 0.5. We have proposed a discriminatory score to relax such an assumption. We have illustrated the consistency of the proposed score through large-scale simulation studies and theoretical proofs. More importantly, we have applied the proposed discriminatory score to select a subset of markers by penalizing the number of markers and to combine the selected markers to yield optimal diagnostic accuracy.

Computer algorithms in medical imaging studies often yield continuous measurements [23]. These algorithms can also be referred to as continuous markers. Following the proposed selection procedure, a subset of algorithms can be selected and measurements from these algorithms can be combined to generate a composite marker which tends to have better accuracy than an individual algorithm.

As one referee points out, our proposed algorithm can be easily extended to marker selection in terms of the partial AUC (pAUC) [30] and/or multi-reader multi-case (MRMC) ROC. In the first case, we can provide clinically relevant values of test sensitivity or specificity and replace full AUC in discriminatory score of the marker by the pAUC. However, it is important to realize that various markers might have different preset values of test sensitivity or specificity for clinical interest and therefore it is more complicated for marker selection based on pAUC. In the second scenario with MRMC ROC, we could substitute mean AUC across multiple readers into the discriminatory score of a marker and proceed the marker selection methods outlined in this article. More researches are needed in these fields.

The proposed criteria bear some resemblance to the Akaike information criterion (AIC) [31] and the Bayesian information criterion (BIC) [32] which are commonly used statistical criteria for model selection. However, there is a distinct difference among them as the proposed method is mainly based on the ROC summary measure, and the latter methods rely on the goodness of fit. The difference is elaborated as follows. The AIC and BIC procedures are based on the likelihood function which is derived from appropriate distribution assumptions. Suppose in a one sample case with n i.i.d. observations from only one population which is either the truly positive population or the truly negative population. They are described as

$$AIC = -2\ln(L(\theta_1, \dots, \theta_p)) + 2p, BIC = -2\ln(L(\theta_1, \dots, \theta_p)) + p\ln(n),$$

where $L(\theta_1, \dots, \theta_p)$ is the likelihood function of p parameters $\theta_1, \dots, \theta_p$, and n is the number of observations. Suppose that observations from multiple markers follow a multivariate normal distribution. In the AIC and BIC criteria, the parameters $\theta_1, \dots, \theta_p$ are means, variances and correlation coefficients to be estimated, and the likelihood function $L(\theta_1, \dots, \theta_p)$ are derived by multiplying the density distributions. For the two-sample case which is considered in the ROC literature, we can consider the AIC and BIC in the following fashion:

$$AIC^* = -2\ln(L_1(\theta_1, \dots, \theta_p)L_2(\theta_1^*, \dots, \theta_p^*)) + 2 \times 2p, BIC^* = -2\ln(L_1(\theta_1, \dots, \theta_p)L_2(\theta_1^*, \dots, \theta_p^*)) + 2p\ln(m+n).$$

Here $L_1(\theta_1, \dots, \theta_p)$ and $L_2(\theta_1^*, \dots, \theta_p^*)$ are the likelihood functions of the truly positive sample and the truly negative sample, respectively. The best model can be selected by minimizing the AIC or BIC. It can be seen that these AIC and BIC criteria reward the goodness of fit of the normal distributions to the observations, and penalize the number of parameters in the model. However, the goodness of fit does not imply large diagnostic accuracy. It is possible to have perfect model fitting to two identical distributions for the two populations. In this case, the likelihood function is maximized, but the markers have no distinguishing power. On the other hand, our criteria is based on minimizing the objective function $f(i_k, k)$ which rewards the ROC summary measure, i_k , and penalizes the number of marker, k , and should be more appropriate for selecting and combining markers. Several colleagues [33, 34] have found that sometimes although added marker is statistically significantly associated with the outcome, it fails to produce significant improvement of AUC. Therefore, it is necessary to present marker selection method based on AUC as an alternative to likelihood based model selection methods, e.g., AIC.

Acknowledgments

This article is dedicated to and in memory of Professor Charles E. Metz of the University of Chicago, Professor Harry S. Wieand of the University of Pittsburgh, Professor Donald D. Dorfman of the University of Iowa, and Dr Robert F. Wagner of the U.S. Food and Drug Administration.

References

1. Swets J. Form of empirical rocs in discrimination and diagnostic tasks: implications for theory and measurement of performance.

2. Metz CE. Basic principles of roc analysis. *Seminars in Nuclear Medicine*. 1978; 8(4):283–298. URL <http://www.sciencedirect.com/science/article/pii/S0001299878800142>. [PubMed: 112681]
3. Alemayehu D, Zou KH. Applications of roc analysis in medical research: Recent developments and future directions. *Academic radiology*. 2012; 19:1457–1464. [PubMed: 23122565]
4. Eng J. Sampling the latest work in receiver operating characteristic analysis: What does it mean? *Academic radiology*. 2012; 19:1449–1451. [PubMed: 23122564]
5. Hanley J, McNeil B. A method of comparing the areas under (ROC) curves derived from same cases. *Radiology*. 1983; 148:839–843. [PubMed: 6878708]
6. Pesce L, Metz C, Berbaum K. On the convexity of roc curves estimated from radiological test results. *Academic radiology*. 2010; 17(8):960–968. [PubMed: 20599155]
7. Bolle, R.; Connell, J.; Pankanti, S.; Ratha, N.; Senior, A. *Guide to Biometrics*. Springer; 2004.
8. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
9. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989; 76:585–592.
10. DeLong ER, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
11. Reiser B, Guttman I. STATISTICAL-INFERENCE FOR PR(Y-LESS-THAN-X) - THE NORMAL CASE. *TECHNOMETRICS*. 1986; 28(3):253–257.
12. Metz CE, Herman BA, Shen J-H. Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in Medicine*. 1998; 17(9): 1033–1053. URL [http://dx.doi.org/10.1002/\(SICI\)1097-0258\(19980515\)17:9<1033::AID-SIM784>3.0.CO;2-Z](http://dx.doi.org/10.1002/(SICI)1097-0258(19980515)17:9<1033::AID-SIM784>3.0.CO;2-Z). [PubMed: 9612889]
13. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology*. 1969; 6:487–496.
14. Zou K, Hall W, Shapiro D. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *STATISTICS IN MEDICINE*. 1997; 16(19):2143–2156. [PubMed: 9330425]
15. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*. 1998; 54:124–135. [PubMed: 9544511]
16. Pesce L, Horsch K, Drukker K, Metz C. Semiparametric estimation of the relationship between roc operating points and the test-result scale: Application to the proper binormal model. *Academic Radiology*. 2011; 18(12):1537–1548. [PubMed: 22055797]
17. Tang LL, Zhou X-H. A semiparametric separation curve approach for comparing correlated roc data from multiple markers. *Journal of Computational and Graphical Statistics*. 2012; 21(3):662–676. [arXiv:http://www.tandfonline.com/doi/pdf/10.1080/10618600.2012.663303](http://www.tandfonline.com/doi/pdf/10.1080/10618600.2012.663303), URL <http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.663303>. [PubMed: 23074360]
18. Zhou, XH.; Obuchowski, NA.; McClish, DK. *Statistical Methods in Diagnostic Medicine*. New York: Wiley; 2002.
19. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press; 2003.
20. Zou, KH.; Liu, A.; Bandos, AI.; Ohno-Machado, L.; Rockette, HE. *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. Chapman & Hall; 2011.
21. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975; 12:387–415.
22. Metz, CE.; Wang, P-L.; Kronman, H. A new approach for testing the significance of differences between roc curves measured from correlated data. In: F, D., editor. *Information Processing in Medical Imaging: Proceedings of the Eighth Conference*; The Hague; Martinus Nijhoff. 1984.
23. Zou K, Warfield S, Fielding J, Tempany C, Wells W, Kaus M, Jolesz F, Kikinis R. Statistical validation based on parametric receiver operating characteristic analysis of continuous classification data. *Academic radiology*. 2003; 10(12):1359–1368. [PubMed: 14697004]

24. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, W WM III, Jolesz FA, Kikinis R. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic Radiology*. 2004; 11(2):178–189. URL <http://www.sciencedirect.com/science/article/pii/S1076633203006718>. [PubMed: 14974593]
25. Kramar A, Faraggi D, Ychou M, Reiser B, Grenier J. The generalized ROC criterion for the evaluation of several tumor markers. *Revue d'Epidemiologie et de Sante Publique*. 1999; 47:217–226.
26. Schisterman, E. Ph.D. thesis. State University of New York at Buffalo; 1999. Lipid peroxidation and antioxidant biomarkers and biomarker disease.
27. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*. 1993; 88(424):1350–1355. URL <http://www.jstor.org/stable/2291276>.
28. Mills JL, Hediger ML, Molloy CA, Chrousos GP, Manning-Courtney P, Yu KF, Brasington M, England LJ. Elevated levels of growth-related hormones in autism and autism spectrum disorder. *Clinical Endocrinology*. 2007; 67(2):230–237. URL <http://dx.doi.org/10.1111/j.1365-2265.2007.02868.x>. [PubMed: 17547689]
29. Liu A, Li Q, Liu C, Yu K, Yu KF. A rank-based test for comparison of multidimensional outcomes. *Journal of the American Statistical Association*. 2010; 105(490):578–587. [arXiv:http://arxiv.org/abs/1011.1111](http://arxiv.org/abs/1011.1111), URL <http://amstat.tandfonline.com/doi/pdf/10.1198/jasa.2010.ap09114>, URL <http://amstat.tandfonline.com/doi/abs/10.1198/jasa.2010.ap09114>. [PubMed: 21625372]
30. Hillis S, Metz C. An analytic expression for the binormal partial area under the roc curve. *Academic Radiology*. 2012; 19(12):1491–1498. [PubMed: 23122568]
31. Bozdogan H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*. 1987; 52(3):345–370.
32. Schwarz G. Estimating the dimension of a model. *The annals of statistics*. 1978; 6(2):461–464.
33. Demler O, Pencina M, et al. Misuse of delong test to compare auCs for nested models. *Statistics in Medicine*.
34. Vickers A, Cronin A, Begg C. One statistical test is sufficient for assessing new predictive markers. *BMC medical research methodology*. 2011; 11(1):13. [PubMed: 21276237]

Appendix

Property of $\hat{i}_k$ and some proof

Let $X_{k1}, \dots, X_{km}$ and $Y_{k1}, \dots, Y_{kn}$ be the expression levels of marker k from m truly positive subjects and n truly negative subjects, respectively. Then from Bamber [1975], a nonparametric estimate of i_k is give by

$$\hat{i}_k = \max \left\{ \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(X_{ki} > Y_{kj}), \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(X_{ki} < Y_{kj}) \right\}. \quad (1)$$

and $\beta = \sigma_N / \sigma_D$.

If we assume that after a Box-Cox type transformation, the ratings of marker k follow a normal distribution $N(\mu_{kD}, \sigma_{kD}^2)$ for a truly positive subject and $N(\mu_{kN}, \sigma_{kN}^2)$ for a truly negative subject, respectively. Then the index becomes

$$i_k = \Phi \left(\frac{|\mu_{kD} - \mu_{kN}|}{\sqrt{\sigma_{kN}^2 + \sigma_{kD}^2}} \right). \quad (2)$$

and its estimator $\hat{i}_k$ is obtained by substituting the mean parameters by the sample means, and the variance parameters by the sample variances. We can show that the estimator is approximately normally distributed with mean zero. The proof is available upon request from the authors.

Multivariate normal distributions

The multivariate normal distributions of measurements for truly positive and truly negative subjects are given by $\mathbf{X}_i = (X_{1i}, X_{2i}, \dots, X_{Ki})^T$ and $\mathbf{Y}_j = (Y_{1j}, Y_{2j}, \dots, Y_{Kj})^T$ follow multivariate normal distributions, respectively. That is, $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$, and $\mathbf{Y}_j \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$, where $\boldsymbol{\mu}_X = (\mu_{1D}, \mu_{2D}, \dots, \mu_{KD})^T$ and $\boldsymbol{\mu}_Y = (\mu_{1N}, \mu_{2N}, \dots, \mu_{KN})^T$ are mean vectors, and $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ are variance-covariance matrices.

The discriminatory score based on the optimal linear combination

The coefficients for the optimal linear combination is given by

$$\boldsymbol{\lambda}^* = (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_Y)^{-1}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y). \quad (3)$$

The resulting discriminatory score is

$$i_K = \Phi \left(\sqrt{(\boldsymbol{\mu}_X + \boldsymbol{\mu}_Y)(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_Y)^{-1}}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y) \right). \quad (4)$$

Figure 1.
ROC Curve ($\delta = 0.01$)

Figure 2.
ROC Curve ($\delta = 0.005$)

Table 1

Additive procedure for combining markers: an example

	IGFII	IGFBP3	IGF1	DHEA
$\hat{t}_k$	0.7659	0.7372	0.6779	0.5189

Table 2

Monte Carlo bias and root mean square error (normal data, nsim = 10000)

Config ($\mu_N, \mu_D, \sigma_N, \sigma_D$)	Estimator	Sample Sizes (n_1, n_2)					
		(10, 10)	(20, 30)	(50, 50)	(50, 100)	(50, 100)	(50, 100)
(0, 0, 1, 1)	P	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)
		0.100 (0.126)	0.065 (0.081)	0.045 (0.057)	0.039 (0.049)	0.039 (0.049)	0.039 (0.049)
(0, 0, 2, 5)	NP	0.106 (0.132)	0.067 (0.084)	0.047 (0.058)	0.040 (0.050)	0.040 (0.050)	0.040 (0.050)
	P	0.103 (0.130)	0.060 (0.076)	0.046 (0.057)	0.034 (0.043)	0.034 (0.043)	0.034 (0.043)
(0, 0.5, 1, 1)	NP	0.111 (0.139)	0.065 (0.082)	0.050 (0.062)	0.036 (0.046)	0.036 (0.046)	0.036 (0.046)
	P	0.016 (0.099)	0.004 (0.073)	-0.001 (0.053)	0.000 (0.047)	0.000 (0.047)	0.000 (0.047)
(0, 0.5, 2, 5)	NP	0.020 (0.103)	0.005 (0.075)	-0.001 (0.054)	0.001 (0.048)	0.001 (0.048)	0.001 (0.048)
	P	0.071 (0.107)	0.031 (0.059)	0.017 (0.043)	0.010 (0.034)	0.010 (0.034)	0.010 (0.034)
(0, 1, 1, 1)	NP	0.080 (0.116)	0.036 (0.064)	0.020 (0.046)	0.012 (0.036)	0.012 (0.036)	0.012 (0.036)
	P	-0.002 (0.100)	-0.002 (0.067)	-0.001 (0.047)	0.000 (0.040)	0.000 (0.040)	0.000 (0.040)
(0, 1, 2, 5)	NP	0.002 (0.105)	0.002 (0.069)	-0.001 (0.047)	0.001 (0.041)	0.001 (0.041)	0.001 (0.041)
	P	0.044 (0.098)	0.013 (0.061)	0.005 (0.049)	0.002 (0.040)	0.002 (0.040)	0.002 (0.040)
(0, 1, 2, 5)	NP	0.052 (0.105)	0.016 (0.064)	0.007 (0.052)	0.003 (0.042)	0.003 (0.042)	0.003 (0.042)

P: parametric (normal-based) AUC estimator

NP: nonparametric AUC estimator

Table 3

Monte Carlo bias and root mean square error (lognormal data, nsim = 10000)

Config ($\mu_N, \mu_D, \sigma_N, \sigma_D$) [†]	Estimator	Sample Sizes (n_1, n_2)					
		(10, 10)	(20, 30)	(50, 50)	(50, 100)		
(0, 0, 1, 1)	P	0.104 (0.130)	0.066 (0.083)	0.046 (0.057)	0.040 (0.049)		
	NP	0.106 (0.132)	0.067 (0.084)	0.047 (0.058)	0.040 (0.050)		
(0, 0, 2, 5)	P	0.110 (0.138)	0.065 (0.081)	0.047 (0.059)	0.035 (0.044)		
	NP	0.111 (0.139)	0.066 (0.083)	0.048 (0.060)	0.037 (0.046)		
(0, 0.5, 1, 1)	P	0.018 (0.101)	0.003 (0.073)	0.001 (0.054)	0.001 (0.047)		
	NP	0.019 (0.102)	0.003 (0.074)	0.002 (0.055)	0.001 (0.047)		
(0, 0.5, 2, 5)	P	0.078 (0.115)	0.034 (0.063)	0.019 (0.045)	0.010 (0.035)		
	NP	0.078 (0.116)	0.035 (0.064)	0.020 (0.046)	0.011 (0.036)		
(0, 1, 1, 1)	P	0.001 (0.100)	-0.001 (0.067)	0.000 (0.047)	0.000 (0.041)		
	NP	0.002 (0.105)	-0.001 (0.069)	0.001 (0.048)	0.000 (0.041)		
(0, 1, 2, 5)	P	0.051 (0.103)	0.015 (0.063)	0.006 (0.051)	0.002 (0.041)		
	NP	0.052 (0.105)	0.016 (0.064)	0.007 (0.052)	0.002 (0.042)		

P: parametric (normal-based) AUC estimator

NP: nonparametric AUC estimator

[†] Mean and standard deviation of the distribution on the log scale

Table 4

Monte Carlo bias and root mean square error (normal mixture data, nsim = 10000)

Config ($\mu_N, \mu_D, \sigma_N, \sigma_D$)	Estimator	Sample Sizes (n_1, n_2)					
		(10, 10)	(20, 30)	(50, 50)	(50, 100)	(50, 100)	(50, 100)
(0, 0, 1, 1)	P	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)
	P	0.104 (0.130)	0.066 (0.083)	0.046 (0.057)	0.039 (0.049)	0.039 (0.049)	0.039 (0.049)
(0, 0, 2, 5)	NP	0.105 (0.132)	0.067 (0.084)	0.047 (0.058)	0.040 (0.050)	0.040 (0.050)	0.040 (0.050)
	P	0.107 (0.133)	0.063 (0.078)	0.047 (0.058)	0.035 (0.044)	0.035 (0.044)	0.035 (0.044)
(0, 0.5, 1, 1)	NP	0.111 (0.139)	0.066 (0.082)	0.049 (0.062)	0.037 (0.046)	0.037 (0.046)	0.037 (0.046)
	P	0.016 (0.098)	0.005 (0.073)	0.002 (0.054)	0.000 (0.047)	0.000 (0.047)	0.000 (0.047)
(0, 0.5, 2, 5)	NP	0.018 (0.101)	0.006 (0.074)	0.004 (0.055)	0.000 (0.047)	0.000 (0.047)	0.000 (0.047)
	P	0.068 (0.106)	0.026 (0.056)	0.015 (0.044)	0.006 (0.032)	0.006 (0.032)	0.006 (0.032)
(0, 1, 1, 1)	NP	0.078 (0.116)	0.034 (0.063)	0.022 (0.049)	0.010 (0.036)	0.010 (0.036)	0.010 (0.036)
	P	-0.003 (0.099)	-0.001 (0.067)	-0.001 (0.047)	0.000 (0.040)	0.000 (0.040)	0.000 (0.040)
(0, 1, 2, 5)	NP	-0.002 (0.105)	0.002 (0.069)	0.001 (0.048)	0.001 (0.041)	0.001 (0.041)	0.001 (0.041)
	P	0.038 (0.094)	0.003 (0.058)	0.001 (0.044)	0.000 (0.039)	0.000 (0.039)	0.000 (0.039)
(0, 1, 2, 5)	NP	0.048 (0.100)	0.015 (0.064)	0.008 (0.053)	0.002 (0.043)	0.002 (0.043)	0.002 (0.043)

P: parametric (normal-based) AUC estimator

NP: nonparametric AUC estimator

$$0.8N(\mu_i, \frac{3\sqrt{2}}{4}\sigma_i) + 0.2N(\mu_i, \frac{\sqrt{2}}{2}\sigma_i), i = N, D.$$

Table 5

Monte Carlo bias and root mean square error (lognormal mixture data, nsim = 10000)

Config ($\mu_N, \mu_D, \sigma_N, \sigma_D$) [†]	Estimator	Sample Sizes (n_1, n_2)					
		(10, 10)	(20, 30)	(50, 50)	(50, 100)	(50, 100)	(50, 100)
(0, 0, 1, 1)	P	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)	Bias (RMSE)
	P	0.103 (0.129)	0.068 (0.083)	0.045 (0.057)	0.039 (0.049)	0.039 (0.049)	0.039 (0.049)
(0, 0, 2, 5)	NP	0.104 (0.131)	0.069 (0.085)	0.047 (0.059)	0.040 (0.050)	0.040 (0.050)	0.040 (0.050)
	P	0.114 (0.141)	0.074 (0.091)	0.058 (0.072)	0.047 (0.058)	0.047 (0.058)	0.047 (0.058)
(0, 0.5, 1, 1)	NP	0.106 (0.133)	0.065 (0.082)	0.051 (0.063)	0.037 (0.046)	0.037 (0.046)	0.037 (0.046)
	P	0.019 (0.100)	0.003 (0.075)	0.002 (0.053)	-0.001 (0.047)	-0.001 (0.047)	-0.001 (0.047)
(0, 0.5, 2, 5)	NP	0.018 (0.100)	0.003 (0.076)	0.003 (0.054)	-0.000 (0.047)	-0.000 (0.047)	-0.000 (0.047)
	P	0.085 (0.121)	0.051 (0.080)	0.041 (0.066)	0.035 (0.056)	0.035 (0.056)	0.035 (0.056)
(0, 1, 1, 1)	NP	0.073 (0.110)	0.034 (0.064)	0.022 (0.050)	0.010 (0.037)	0.010 (0.037)	0.010 (0.037)
	P	0.002 (0.101)	0.002 (0.068)	0.001 (0.047)	-0.001 (0.041)	-0.001 (0.041)	-0.001 (0.041)
(0, 1, 2, 5)	NP	0.003 (0.105)	0.002 (0.069)	0.002 (0.047)	-0.001 (0.041)	-0.001 (0.041)	-0.001 (0.041)
	P	0.064 (0.112)	0.040 (0.080)	0.035 (0.067)	0.033 (0.057)	0.033 (0.057)	0.033 (0.057)
(0, 1, 2, 5)	NP	0.048 (0.100)	0.016 (0.065)	0.009 (0.053)	0.002 (0.043)	0.002 (0.043)	0.002 (0.043)
	P	0.048 (0.100)	0.016 (0.065)	0.009 (0.053)	0.002 (0.043)	0.002 (0.043)	0.002 (0.043)

P: parametric (normal-based) AUC estimator

NP: nonparametric AUC estimator

[†] Mean and standard deviation of the distribution on the log scale

$$0.8LN(\mu_i, \frac{3\sqrt{2}}{4}\sigma_i) + 0.2LN(\mu_i, \frac{\sqrt{2}}{2}\sigma_i), i = N, D.$$