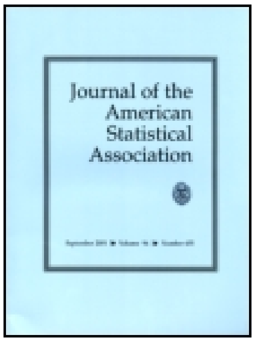

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Semiparametric Regression for the Area Under the Receiver Operating Characteristic Curve

Lori E Dodd & Margaret Sullivan Pepe

Semiparametric Regression for the Area Under the Receiver Operating Characteristic Curve

Lori E. DODD and Margaret Sullivan PEPE

Medical advances continue to provide new and potentially better means for detecting disease. Such is true in cancer, for example, where biomarkers are sought for early detection and where improvements in imaging methods may pick up the initial functional and molecular changes associated with cancer development. In other binary classification tasks, computational algorithms such as neural networks, support vector machines, and evolutionary algorithms have been applied to areas as diverse as credit scoring, object recognition, and peptide-binding prediction. Before a classifier becomes an accepted technology, it must undergo rigorous evaluation to determine its ability to discriminate between states. Characterization of factors influencing classifier performance is an important step in this process. Analysis of covariates may reveal subpopulations in which classifier performance is greatest or identify features of the classifier that improve accuracy. We develop regression methods for the nonparametric area under the receiver operating characteristic curve, a well-accepted summary measure of classifier accuracy. The estimating function generalizes standard approaches and, interestingly, is related to the two-sample Mann–Whitney U statistic. Implementation is straightforward, because it is an adaptation of binary regression methods. Asymptotic theory is nonstandard, because the regressor variables are cross-correlated. Nevertheless, simulation studies show that the method produces estimates with small bias and reasonable coverage probability. Application of the method to evaluate the covariate effects on a new device for diagnosing hearing impairment reveals that the device performs better in more severely impaired subjects and that certain test parameters, which are adjustable by the device operator, are key to test performance.

KEY WORDS: Classification; Classifier performance; Diagnostic test; Disease screening; Prediction.

1. INTRODUCTION

The performance of a binary classifier with continuous output is often evaluated with receiver operating characteristic (ROC) curve analysis (Zhu, Beling, and Overstreet 2002; Brusic, Rudy, Honeyman, Hammer, and Harrison 2002; Pepe 2000). For two states, D and $\overline{D}$, which are typically diseased and nondiseased states in medicine, and classifier output, Y, let $Y > c$ indicate classification into state D. The ROC curve plots $\{\Pr(Y > c \mid \overline{D}), \Pr(Y > c \mid D)\}$ for all possible thresholds c, and provides a visual description of the trade-offs between the true-positive rate (TPR) and the false-positive rate (FPR) as the threshold stringency (c) changes. For $t = FPR(c)$, we can write $ROC(t) = TPR\{FPR^{-1}(t)\}$. The curve lies in the unit-square, in which a useless classifier is represented by the diagonal line from vertices (0, 0) to (1, 1) and a curve pulled closer toward (0, 1) indicates better performance. When under development, a classifier's optimal threshold is not known. Because the relative importance of false-negative and false-positive misclassifications changes depending on the setting in which the technology is implemented, the optimal threshold varies. Hence a summary measure that aggregates performance information across possible thresholds is desirable. The area under the ROC curve (AUC) summarizes across all thresholds. The AUC has the interpretation $\Pr(Y^D > Y^{\overline{D}})$, where the superscripts indicate from which state the output arises (Bamber 1975). We prefer to interpret the AUC as an average true-positive rate across false-positive rates, because $AUC = \int_0^1 ROC(t)\,dt$. A perfect classifier has $AUC = 1$, whereas one that performs no better than chance has an AUC of 1/2. Although the AUC is by far the most commonly used summary index, other measures have been described (see Shapiro 1999 for a review) and are preferable in certain settings. In this article we focus on the AUC.

Classifier performance may depend on several factors, including characteristics of the population tested or operating parameters of the test. Consider the following study of an experimental hearing device developed to diagnose hearing impairment. The device under study, distortion product otoacoustic emission (DPOAE), measures the strength of the cochlear response from two sounds emitted into a single ear at different frequencies and intensities (Stover, Gorga, and Neely 1996). The strength of the DPOAE output, measured by DPOAE amplitude, indicates auditory function. Because the standard method for diagnosis of hearing impairment requires active subject participation, the DPOAE device might be useful for subjects who are too sick, too young, or not mentally capable to undergo the behavioral gold standard test.

One goal of the study was to determine whether DPOAE performance depends on the frequency and intensity of the two stimuli emitted into the ear to select optimal stimuli for further research. Additionally, the relationship between performance and severity of hearing impairment is of interest; for example, maybe DPOAE diagnoses the most severely impaired ears better than ears with mild impairment. Exploration of the relationship between severity of impairment and diagnostic accuracy yields information about the types of patients who will be diagnosed with the system. We refer to the severity covariate as "disease-specific" because it applies only to diseased (or hearing-impaired) subjects. The other covariates, frequency and intensity, are adjustable operating parameters of the device. Other applications may include covariates that characterize performance as a function of the population tested (e.g., age or gender) or of the testers (e.g., experience). Understanding the effects that such covariates have on the discrimination capacity of the classifier can suggest settings in which the classifier works best and motivate innovations in settings in which performance is inadequate.

We propose to evaluate covariate effects on classifier accuracy using a regression model for the AUC summary index of

Lori E. Dodd is Mathematical Statistician, Biometric Research Branch, National Cancer Institute, Rockville, MD 20892 (E-mail: *doddl@mail.nih.gov*). Margaret S. Pepe is Professor, Department of Biostatistics, University of Washington, Seattle, WA 98195 (E-mail: *mspepe@u.washington.edu*) and a Member of the Fred Hutchinson Cancer Research Center, Seattle, WA 98109. The authors thank the editor, the associate editor, and the reviewers for valuable comments. This research was supported by National Institutes of Health grants T32 HL07183 and R01 GM54438.

Journal of the American Statistical Association
June 2003, Vol. 98, No. 462, Theory and Methods
DOI 10.1198/016214503000198

the ROC curve. This is analogous to the evaluation of covariate effects on an outcome variable by using regression models for the mean, which is, after all, a summary statistic for the distribution of the variable. Alternative approaches to regression modeling of ROC curves have been proposed (see Pepe 1998 for a review), and we contrast them briefly with AUC regression in Section 7. First, we develop our approach.

2. AREA UNDER THE CURVE BINARY REGRESSION

2.1 The Model

Although D and $\overline{D}$ may be any two states, we use terminology from diagnostic testing for them, so D is referred to as "disease" and $\overline{D}$ is referred to as "no disease." We use $\mathbf{X}$ to denote covariates and Y to denote classifier output. Let $(Y_i^D, \mathbf{X}_i^D)$ and $(Y_j^{\overline{D}}, \mathbf{X}_j^{\overline{D}})$ denote observations from D and $\overline{D}$, with $(i = 1, \ldots, n_D)$ and $(j = 1, \ldots, n_{\overline{D}})$. The result of Bamber (1975) suggest that we can write the covariate-specific AUC as $\Pr(Y_i^D > Y_j^{\overline{D}} \mid \mathbf{X}_i^D, \mathbf{X}_j^{\overline{D}}) \equiv \theta_{ij}$. The parameter θ_{ij} compares the results from the diseased population with covariates $\mathbf{X}_i^D$ with the results from the nondiseased population with covariates $\mathbf{X}_j^{\overline{D}}$. To simplify notation, let $\mathbf{X}_{ij}$ denote $(\mathbf{X}_i^D, \mathbf{X}_j^{\overline{D}})$ or a specified function of them. For a vector of parameters $\boldsymbol{\beta}$ and a monotone increasing link function g, we propose the following AUC regression model:

$$g(\theta_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}. \tag{1}$$

The probit and logit are natural link functions. When the logit link is used, exponentiated parameters have interpretations as AUC odds, where AUC odds are defined as $AUC/(1 - AUC) = \Pr(Y^D > Y^{\overline{D}})/\Pr(Y^D < Y^{\overline{D}})$. Because larger AUCs are associated with increasing accuracy, AUC odds greater than 1 indicate improved test accuracy.

Now consider a binary covariate such as gender with, say, $X = 0$ for males. In this case, the AUC is computed within each gender as an AUC comparing test results of diseased females to nondiseased males (or vice versa) is typically not of interest. Under the model $logit(\theta) = \beta_0 + \beta_1 X$, $exp(\beta_1)$ is the ratio of AUC odds for the test in women versus men. If $\beta_1 > 0$, then the test is better at distinguishing between diseased and nondiseased women than between diseased and nondiseased men.

When covariates are specific to the diseased group (e.g., stage of disease), the AUC is modeled as a function of the covariate X_i^D. That is, the covariate-specific AUC is defined as $\Pr(Y_i^D > Y_j^{\overline{D}} \mid X_i^D) \equiv \theta_i$. The model $logit(\theta_i) = \beta_0 + \beta_1 X_i^D$ describes the change in accuracy as a function of X_i^D on the logit scale. The number $exp(\beta_1)$ describes the ratio of AUC odds associated with a one-unit increase in stage of disease.

For a continuous covariate, the model of interest describes the change in accuracy as a covariate common to the diseased and nondiseased group changes. Consider, for example, the covariate *age*. Computation of an AUC for diseased subjects of age 80 and nondiseased subjects of age 50 is not scientifically relevant, whereas an AUC for diseased and nondiseased subjects both of age 80 (*or* of age 50) is of interest. The goal is to understand how the AUC *for diseased and nondiseased subjects of the same age* changes as age varies. The parameter β_1 in the model $logit(\theta_{ij}) = \beta_0 + \beta_1 X_i^D + \beta_2 (X_i^D - X_j^{\overline{D}})$ describes this relationship. If the covariate is age in years, then $exp(\beta_1)$ is the ratio of AUC odds associated with a 1-year increase in age for diseased and nondiseased subjects of the same age. If this value is greater than 1, then the AUC is an increasing function of age, and the test performs better in older subjects than in younger subjects.

2.2 Proposed Estimating Function

To estimate the regression parameters, we propose a binary regression. Define $U_{ij} = I(Y_i^D > Y_j^{\overline{D}})$, and let $N = n_D + n_{\overline{D}}$. Note that $E(U_{ij} \mid \mathbf{X}_{ij}) = \Pr(Y_i^D > Y_j^{\overline{D}} \mid \mathbf{X}_{ij}) = \theta_{ij}$. This suggests that our model, $g(\theta_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}$, is a generalized linear regression model for the binary variables U_{ij}. The estimating function

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_i^{n_D} \sum_j^{n_{\overline{D}}} \frac{\partial \theta_{ij}}{\partial \boldsymbol{\beta}} \nu(\theta_{ij})^{-1} (U_{ij} - \theta_{ij}) \equiv \sum_i^{n_D} \sum_j^{n_{\overline{D}}} \mathbf{S}_{ij}(\boldsymbol{\beta}) \tag{2}$$

is the classic estimating function for binary regression, except that the U_{ij}'s are not independent. The term $\partial \theta_{ij} / \partial \boldsymbol{\beta}$ is a $(p \times 1)$ vector of the partial derivatives of θ_{ij} with respect to the model parameters $\boldsymbol{\beta}$. The term $\nu(\theta_{ij})$ is the variance function, whereas the last term describes the mean model of U_{ij} conditional on X_{ij}.

The binary random variables U_{ij} in expression (2) are cross-correlated. For example, the indicator U_{ij} will be correlated with $U_{ij'}$ for all $j \neq j'$, because the ith diseased observation contributes to each indicator. Similarly, for each fixed j, the indicators are correlated across all i. As a result, asymptotic theory is not standard. The estimating function assumes that observations are independent, and, to borrow language from generalized estimating equations (GEEs), uses an independent working covariance matrix (WCM). Note that an WCM that accounted for the correlations might improve efficiency. However, the Pepe–Anderson condition that allows for a nondiagonal WCM often fails in diagnostic testing applications with repeated measures and would result in inconsistent estimates (Diggle, Heagerty, Liang, and Zeger 2002, p. 255). Furthermore, in applications in which the foregoing condition is met, the dimensionality of the nondiagonal WCM may be prohibitively large. For example, in the application here, the matrix would be of dimension 72,708 × 72,708.

2.3 Implementation

Data are observed as follows: $\{(Y_1^D, \mathbf{X}_1^D), \ldots, (Y_{n_D}^D, \mathbf{X}_{n_D}^D), (Y_1^{\overline{D}}, \mathbf{X}_1^{\overline{D}}), \ldots, (Y_{n_{\overline{D}}}^{\overline{D}}, \mathbf{X}_{n_{\overline{D}}}^{\overline{D}})\}$. Section 2.2 suggests that all pairs are included in (2), but one needs to include (and model) only subsets of pairs. First, note that if covariates are categorical and there are sufficient observations at each covariate level, then pairs are created only within strata, defined by distinct covariate values. However, when covariates are not categorical or data are too sparse within strata, pairs of $(Y_i^D, \mathbf{X}_i^D)$ and $(Y_j^{\overline{D}}, \mathbf{X}_j^{\overline{D}})$ must be created for subjects with different covariate values. It may not be appropriate to pair $(Y_i^D, \mathbf{X}_i^D)$ and $(Y_j^{\overline{D}}, \mathbf{X}_j^{\overline{D}})$ for all (i, j), because this allows covariate values far apart from one another to influence model fit. We propose to pair observations with covariate values that are within a neighborhood; for example, create a pair if $|\mathbf{X}_i^D - \mathbf{X}_j^{\overline{D}}| \le \zeta$. If covariates for the (i, j)th pair are farther than ζ apart, then that pair is not included in the estimating function. Observe that the estimating function is now a sum over only the (i, j) pairs satisfying $|\mathbf{X}_i^D - \mathbf{X}_j^{\overline{D}}| \le \zeta$. The number of pairs depends on ζ and on the distribution of covariates. For a given i, the number of observations from nondiseased subjects paired with Y_i^D is denoted by $n_{\overline{D}}(\zeta, i)$. Here the estimating function is the sum $\sum_i^{n_D} \sum_j^{n_{\overline{D}}(\zeta, i)} \mathbf{S}_{ij}(\boldsymbol{\beta})$. Choosing $\zeta = 0$ corresponds to pairing only observations with the same covariate value. At the other extreme, setting $\zeta = \infty$ corresponds to pairing all diseased and nondiseased results. There is a trade-off between bias and efficiency as ζ varies. For a small ζ, much of the data are excluded, and the method will be less efficient. On the other hand, for a large ζ, more structure is imposed on the data, and unless it is correct, this introduces bias. When fewer model restrictions are preferred, select ζ as small as possible, while including sufficient covariate pairs within a neighborhood to give estimates with adequate precision. There are obvious analogies here to the problem of smoothing in regression.

Once the pairing has been completed, estimation proceeds by setting the estimating function equal to 0. If the link function is chosen to be the identity, then closed-form expressions for $\hat{\boldsymbol{\beta}}$ are derived. Otherwise, estimation requires an iterative procedure, such as Newton–Raphson (McCullagh and Nelder 1997). Logistic or probit regression estimation routines in standard statistical packages can be used to calculate estimates, although standard errors require either the bootstrap or special programming for asymptotic variance forms.

3. ASYMPTOTIC DISTRIBUTION THEORY

The estimating function (2) is a sum of random variables that are cross-correlated. Hence standard theory developed for sums of independent random variables does not apply. To simplify notation, we assume that $\zeta = \infty$ here. The theory holds for $\zeta \in (0, \infty)$ but is notationally complex. As before, let $n_D(\zeta, j)$ denote the number of Y^D's paired with the jth result of the nondiseased population, and similarly for $n_{\overline{D}}(\zeta, i)$. Then, as long as $n_D(\zeta, j) = O(N)$ and $n_{\overline{D}}(\zeta, j) = O(N)$, the theory applies. If ζ is fixed and does not get smaller as N increases, then these conditions should be satisfied. In other words, as long as each diseased subject is paired with a proportion of the nondiseased subjects, the theory outlined here applies.

To derive theory, we assume the following conditions:

(C1) $\{(Y_i^D, \mathbf{X}_i^D) : i = 1, \ldots, n_D\}$ are iid, $\{(Y_j^{\overline{D}}, \mathbf{X}_j^{\overline{D}}) : j = 1, \ldots, n_{\overline{D}}\}$ are iid, and both vectors are mutually independent.

(C2) $\lim_{N\to\infty} n_D/N \to \lambda$, where $0 < \lambda < 1$ and $N = n_D + n_{\overline{D}}$.

(C3) $g(u)$ is monotone increasing and three-times differentiable with bounded derivatives.

(C4) There exists $\epsilon > 0$ such that $\nu(\theta_{ij}) > \epsilon$ for $\boldsymbol{\beta} \in N_\delta(\boldsymbol{\beta}_0) \equiv \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \delta\}$.

(C5) The covariate space is bounded.

(C6) The matrix $E(\partial \mathbf{S}_{ij}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta})$ is negative definite.

It follows from (C3)–(C6) that $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta})$, $\frac{1}{n_D n_{\overline{D}}}\frac{\partial^2}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T}\mathbf{S}_N(\boldsymbol{\beta})$, and $\frac{\partial}{\partial\boldsymbol{\beta}}E\{\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta})\}$ are bounded uniformly for $\boldsymbol{\beta} \in N_\delta(\boldsymbol{\beta}_0)$. To see this, one must show that each of the elements in $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta})$ and $\frac{1}{n_D n_{\overline{D}}}\frac{\partial^2}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T}\mathbf{S}_N(\boldsymbol{\beta})$ has a bound independent of $\boldsymbol{\beta}$. The boundedness condition of $\frac{\partial}{\partial\boldsymbol{\beta}}E\{\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta})\}$ is slightly more involved and requires demonstrating that its limit is equal to that of $E\{\frac{1}{n_D n_{\overline{D}}}\frac{\partial^2}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T}\mathbf{S}_N(\boldsymbol{\beta})\}$, whose bound does not depend on $\boldsymbol{\beta}$. We refer to this as property (B). Proofs of lemmas are given in the Appendix.

3.1 Consistency

Theorem 1. Under (C1)–(C6), as $N \to \infty$, solutions to $\mathbf{S}_N(\boldsymbol{\beta}) = 0$ are unique with probability converging to 1 and $\hat{\boldsymbol{\beta}} \xrightarrow[p]{} \boldsymbol{\beta}_0$.

Consistency is established by demonstrating the four conditions described by Foutz (1977), which are sufficient for the existence and uniqueness of consistent solutions to likelihood equations. Although the result was developed for likelihood equations, it can be applied to any estimating function satisfying the following four properties, which we refer to as "Foutz conditions":

(F1) $\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta})$ exists and is continuous for $\boldsymbol{\beta} \in N_\delta(\boldsymbol{\beta}_0)$.

(F2) $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta}) \xrightarrow[p]{} E\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{ij}(\boldsymbol{\beta})$ uniformly for $\boldsymbol{\beta} \in N_\delta(\boldsymbol{\beta}_0)$ as $N \to \infty$.

(F3) $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta}_0)$ is negative definite with probability converging to one as $N \to \infty$.

(F4) $ES_N(\boldsymbol{\beta}_0) = 0$.

These assumptions and the following two lemmas are sufficient for establishing the Foutz conditions.

Lemma 1. Under property (B), and if, for each fixed $\boldsymbol{\beta} \in N_\delta(\boldsymbol{\beta}_0)$, $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta})$ converges to $E\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{ij}(\boldsymbol{\beta})$ in probability as $N \to \infty$, then convergence of $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta})$ to $E\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{ij}(\boldsymbol{\beta})$ is uniform for $\boldsymbol{\beta} \in N_\delta(\boldsymbol{\beta}_0)$.

Lemma 2. $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta}) \xrightarrow[p]{} E\frac{\partial\mathbf{S}_{ij}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$ as $N \to \infty$.

Condition (F1) follows trivially from the foregoing assumptions by the existence of third derivatives of the elements of $\mathbf{S}_N(\boldsymbol{\beta})$. The sufficient conditions for uniform convergence required by (F2) are given by Lemma 1. Lemma 2 establishes the convergence results needed for Lemma 1. Hence Foutz's condition (F2) is satisfied. Condition (F3) follows because $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_N(\boldsymbol{\beta}_0) \xrightarrow[p]{} E\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{ij}(\boldsymbol{\beta}_0)$ by Lemma 2, which by assumption is a negative definite matrix. Finally, because by definition $E(U_{ij}) = \theta_{ij}$, condition (F4) is satisfied.

3.2 Asymptotic Normality

To derive the limiting distribution, we find a sum that closely approximates $\mathbf{S}_N(\boldsymbol{\beta})$ to which a central limit theorem for triangular arrays can be applied. First, we take the conditional expectation of U_{ij} at a fixed test result for a diseased subject. Consider the following:

$$
\begin{aligned}
E\{U_{ij} | Y_i^D = y_i^D, \mathbf{X}_i^D, \mathbf{X}_j^{\overline{D}}\} &= E\{I(y_i^D > Y_j^{\overline{D}}) \mid \mathbf{X}_i^D, \mathbf{X}_j^{\overline{D}}\} \\
&= P_{\mathbf{X}_j^{\overline{D}}}(y_i^D > Y^{\overline{D}}) \equiv F^{\overline{D}}_{\mathbf{X}_j^{\overline{D}}}(y_i^D).
\end{aligned}
$$

This notation denotes the probability of observing a value of y_i^D or lower in the distribution of test results of nondiseased subjects that have covariate pattern $\mathbf{X}_j^{\overline{D}}$. We refer to $1 - F^{\overline{D}}_{\mathbf{X}_j^{\overline{D}}}(y_i^D)$

as a *placement value.* It indicates the "place" that the diseased observation has in the distribution of nondiseased subjects' test results with covariate pattern $\mathbf{X}_j^{\overline{D}}$. For a given y_i^D, a value of $F^{\overline{D}}_{\mathbf{X}_j^{\overline{D}}}(y_i^D)$ closer to 1 indicates that most of the nondiseased subjects' test results fall below it. Note that $E\{F^{\overline{D}}_{\mathbf{X}_j^{\overline{D}}}(Y_i^D) \mid \mathbf{X}_i^D\} = \Pr(Y_i^D > Y_j^{\overline{D}} \mid \mathbf{X}_i^D, \mathbf{X}_j^{\overline{D}}) = \theta_{ij}$. If the Y^D's on average fall in the upper tail of the distribution of $Y^{\overline{D}}$, then the AUC will be larger.

An analogous entity is defined by conditioning on a nondiseased observation as follows: $E(U_{ij} \mid Y_j^{\overline{D}} = y_j^{\overline{D}}, \mathbf{X}_i^D, \mathbf{X}_j^{\overline{D}}) = 1 - F^D_{\mathbf{X}_i^D}(y_j^{\overline{D}}) \equiv \overline{F}^D_{\mathbf{X}_i^D}(y_j^{\overline{D}})$. The interpretation is similar to the placement value concept for y_i^D. We define the following sum:

$$\mathbf{S}_{N,P}(\boldsymbol{\beta}) = \sum_i^{n_D} \sum_j^{n_{\overline{D}}} \boldsymbol{\omega}_{ij} \Big[\Big\{F_{\mathbf{X}_j^{\overline{D}}}(Y_i^D) - \theta_{ij}\Big\} + \Big\{F_{\mathbf{X}_j^{\overline{D}}}(Y_i^D) - \theta_{ij}\Big\}\Big], \qquad (3)$$

where $\boldsymbol{\omega}_{ij} = (\partial\theta_{ij}/\partial\boldsymbol{\beta})\nu^{-1}(\theta_{ij})$. Arguments from U-statistic theory can be used to show that $N^{-3/2}\{\mathbf{S}_{N,P}(\boldsymbol{\beta}) - \mathbf{S}_N(\boldsymbol{\beta})\} \xrightarrow[p]{} 0$. Because $\mathbf{S}_N(\beta)$ and $\mathbf{S}_{N,P}(\boldsymbol{\beta})$ are asymptotically equivalent, the asymptotic normality claimed in Theorem 2 is proven by applying a central limit theorem for triangular arrays to $\mathbf{S}_{N,P}(\boldsymbol{\beta})$, which is a sum of independent random variables.

Theorem 2. Under (C1)–(C6), $\sqrt{\frac{n_D n_{\overline{D}}}{N}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow[d]{} Z \sim N(0, I(\boldsymbol{\beta}_0)^{-1} \boldsymbol{\Sigma} I(\boldsymbol{\beta}_0)^{-1})$ as $N \to \infty$, where $I(\boldsymbol{\beta}_0) \equiv -E\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{S}_{ij}(\boldsymbol{\beta}_0)$ and

$$\begin{aligned}\boldsymbol{\Sigma} &= \lim_{N\to\infty}\Bigg[\frac{n_D}{N}\Bigg\{\frac{1}{n_{\overline{D}}}\sum_j^{n_{\overline{D}}}\frac{1}{n_D^2}\sum_i^{n_D}\sum_k^{n_D}\boldsymbol{\omega}_{ij}\boldsymbol{\omega}_{kj}^T \\ &\qquad \times \operatorname{cov}(\overline{F}^D_{\mathbf{X}_i^D}(Y_j^{\overline{D}}), \overline{F}^D_{\mathbf{X}_k^D}(Y_j^{\overline{D}}))\Bigg\}\Bigg] \\ &\quad + \lim_{N\to\infty}\Bigg[\frac{n_{\overline{D}}}{N}\Bigg\{\frac{1}{n_D}\sum_i^{n_D}\frac{1}{n_{\overline{D}}^2}\sum_j^{n_{\overline{D}}}\sum_l^{n_{\overline{D}}}\boldsymbol{\omega}_{ij}\boldsymbol{\omega}_{il}^T \\ &\qquad \times \operatorname{cov}(F^{\overline{D}}_{\mathbf{X}_j^{\overline{D}}}(Y_i^D), F^{\overline{D}}_{\mathbf{X}_l^{\overline{D}}}(Y_i^D))\Bigg\}\Bigg] \\ &\equiv \lambda\, \boldsymbol{\Sigma}_{\overline{D}} + (1-\lambda)\, \boldsymbol{\Sigma}_D. \end{aligned} \qquad (4)$$

Observe that the asymptotic variance is comprised of one component that depends on variability in Y^D and another that depends on $Y^{\overline{D}}$, with each weighted by its relative contribution to the overall sample size. To obtain variance estimates, models for $F^D_{\mathbf{X}^D}$ and $F^{\overline{D}}_{\mathbf{X}^{\overline{D}}}$ must be specified. Bootstrapped standard errors are recommended when covariate data are continuous or sparse, because making such assumptions is undesirable in practice. When covariates are discrete and there are sufficient observations at each level to estimate $F^D_{\mathbf{X}^D}$ and $F^{\overline{D}}_{\mathbf{X}^{\overline{D}}}$, this formula could be applied. The theory is extended to repeated-measures data when the number of diseased and nondiseased subjects gets large. To show this, we identify all of the ij pairs in the score equation and call the sum of these U_{ij}^*. Similar theory can then be applied to the U_{ij}^*s, although the variance has a different form. When there are repeated measures, we recommend using the bootstrap to obtain appropriate standard errors.

4. RELATIONSHIPS WITH EXISTING METHODS

4.1 Comparing Two Areas Under the Curve

Consider the following model to compare two tests administered to each subject: $\theta_k = g^{-1}(\beta_0 + \beta_1 X_k)$, where $(k = 1, 2)$ and X_k is an indicator variable for test type with value 0 when $k = 1$. For this simple case, the proposed method recovers an existing approach in the literature. The model parameterizes the AUCs for the two tests as $g^{-1}(\beta_0)$ and $g^{-1}(\beta_0 + \beta_1)$. To compare the AUCs for the two tests, we test the null hypothesis $H_0 : \beta_1 = 0$. Denote $U_{ijk} = I(Y_{ik}^D > Y_{jk}^{\overline{D}})$ and let $\nu(\theta_{ij}) = 1$. The estimating function is simply

$$\sum_{k=1}^{2}\sum_{i=1}^{n_{Dk}}\sum_{j=1}^{n_{\overline{D}k}}\begin{pmatrix}1\\X_k\end{pmatrix}\{U_{ijk} - g^{-1}(\beta_0 + \beta_1 X_k)\}. \qquad (5)$$

The estimator of $g^{-1}(\beta_0)$ under the null hypothesis is

$$g^{-1}(\hat{\beta}_0^0) = \left(\prod_{k=1}^{2} n_{Dk} n_{\overline{D}k}\right)^{-1} \sum_{k=1}^{2}\sum_i^{n_{Dk}}\sum_j^{n_{\overline{D}k}} U_{ijk}.$$

We obtain a score-like statistic by evaluating the second element of (5) at $\hat{\beta}_0^0$,

$$Score_{H_0} = N^\star\left(\frac{\sum_i\sum_j U_{ij2}}{n_{D2}n_{\overline{D}2}} - \frac{\sum_i\sum_j U_{ij1}}{n_{D1}n_{\overline{D}1}}\right),$$

where the term $N^\star = (n_{D1}n_{\overline{D}1}n_{D2}n_{\overline{D}2})/(n_{D1}n_{\overline{D}1} + n_{D2}n_{\overline{D}2})$.

Recall that the standard empirical estimate of the AUC is the Mann–Whitney U statistic, and recognize the terms $\sum_i\sum_j U_{ijk}/n_{Dk}n_{\overline{D}k}$ as such. Hence we can write $Score_{H_0} = N^\star(\hat{\theta}_2 - \hat{\theta}_1)$, which is the standardized difference in empirical AUCs, the standard nonparametric statistic for comparing two or more diagnostic tests as described by DeLong, DeLong, and Clarke-Pearson (1988). Our arguments show, therefore, that our regression approach yields the standard nonparametric procedure for comparing two tests as a special case.

4.2 Comparison With Existing Area Under the Curve Regression Methods

4.2.1 Derived Variables Approach. Thompson and Zucchini (1989) proposed AUC regression methods for diagnostic tests based on derived variables. Consider a covariate X_k that takes K distinct values. Denote an AUC estimate at the kth covariate level as $\hat{\theta}_k$. The derived variables AUC regression model is given by

$$E(\hat{\theta}_k) = \beta_0^d + \beta_1^d X_k.$$

Because the AUC takes values in the interval (0, 1), a model of a transformation of $\hat{\theta}$, such as $E\{g(\hat{\theta}_k)\} = \beta_0^d + \beta_1^d X_k$, where $\{g : (0, 1) \mapsto R^1\}$ so that it takes on less-restricted values, may be preferred. Note that this model prohibits transformation back to the original AUC scale. A major weakness of this method is that continuous covariates cannot be modeled. Further, because different numbers of subjects often contribute to AUC estimates across covariate levels, the regression assumption of equal variances will frequently fail.

4.2.2 Jackknifed Area Under the Curve Approach. Dorfman, Berbaum, and Metz (1992) proposed a method based on computing jackknifed AUC values for each subject to estimate random-effects models. We consider a simple extension of their approach to a linear regression model to make their method more comparable with ours. Let $\hat{\theta}_k$ denote the AUC estimate and N_k denote the total number of observations at the kth covariate level. Jackknifed AUC values for the ith subject are computed as $\theta^*_{ik} = N_k\hat{\theta}_k - (N_k - 1)\hat{\theta}_{k(i)}$, where $\hat{\theta}_{k(i)}$ is an estimate of θ_k with the ith subject deleted. Jackknifed AUC values are treated as independent variables, and linear regression methods are used to obtain parameter estimates. In some sense, each θ^*_{ik} represents the contribution of the ith subject to the AUC estimate at covariate level k. The regression model is given by $E(\theta^*_{ik}) = \beta^J_0 + \beta^J_1 X_k$. Because $E(\theta^*_{ik}) \in (0, 1)$, we again consider using nonlinear regression methods to estimate models of the form

$$g\{E(\theta^*_{ik})\} = \beta^J_0 + \beta^J_1 X_k,$$

where the function g is defined as before. Like the derived-variables AUC regression method, this approach also has the major limitation of not allowing continuous covariates.

4.2.3 Analytical Comparisons.

Theorem 3. When $n_{Dk} = n_D$ and $n_{\bar{D}k} = n_{\bar{D}}$ for all k, $\hat{\theta}_k = \frac{1}{n_D n_{\bar{D}}}\sum_{ij} U_{ijk}$, and a linear regression model with g the identity link function is assumed, the parameter estimators of the proposed, derived-variable, and jackknifed-AUC methods are identical.

Refer to the Appendix for a proof. Under less-restrictive conditions, such as unequal numbers of observations across covariate levels or a nonidentity link function, the estimators differ. In the following section we compare the three methods under more general conditions via simulation studies.

5. FINITE-SAMPLE PERFORMANCE

We conduct several simulation studies to compare, under a more general setting than assumed in Theorem 3, the methods described in Section 4.2. Next we evaluate the small-sample performance of the proposed method under a model for continuous covariates. We generate data such that $Y^D_i \sim N(\mu_{D,X}, \sigma^2_D)$ and $Y^{\bar{D}}_j \sim N(\mu_{\bar{D},X}, \sigma^2_{\bar{D}})$, where we let $\mu_{\bar{D},X} = \gamma_0 + \gamma_1 X$ and $\mu_{D,X} = \gamma_0 + (\gamma_1 + \gamma_2)X$. Under this parameterization,

$$\theta_X = \Phi\left(\frac{\mu_{D,X} - \mu_{\bar{D},X}}{\sqrt{\sigma^2_{\bar{D}} + \sigma^2_D}}\right) = \Phi\left(\frac{\gamma_2 X}{\sqrt{\sigma^2_{\bar{D}} + \sigma^2_D}}\right) = \Phi(\beta X), \quad (6)$$

where $\beta = \frac{\gamma_2}{\sqrt{\sigma^2_{\bar{D}} + \sigma^2_D}}$ and $\Phi(\cdot)$ is the cumulative normal distribution function. (See Pepe 1998 for a derivation of this model.)

5.1 Comparison With Existing Area Under the Curve Methods

Observations are generated from the model in (6) across five covariate levels ($X = 1, 2, 3, 4, 5$) with balanced and unbalanced distributions across categories. We chose $\mu_{D,X} = .5X$, $\sigma_D = 1.2$, $\mu_{\bar{D},X} = 0$, and $\sigma_{\bar{D}} = 1$ so that the model is $\Phi^{-1}(\theta_k) = .32X_k$. Sample sizes of 50, 100, and 200 are studied. We fit the three models described in Section 4.2. Results for a sample size of 100 are presented in Table 1. Results for other sample sizes have been given by Dodd (2001). Our method produces estimates that are both the least biased and the most efficient for all scenarios studied. As expected, when the balance in the number of observations across covariates is distorted, the proposed method provides a more natural weighting and results in an even greater increase in efficiency. Efficiencies relative to our method, computed from the ratios of variances across the 1,000 realizations of the model, are as low as 14% for the jackknifed-AUC and 76% for the derived-variables method.

5.2 Model With Continuous Covariates

To evaluate the method in a setting with continuous covariates, we generate data from the model in (6), except $X \sim \textit{Uniform}(0, 10)$. Parameter estimates are obtained from generating U_{ij}'s for all pairs of test results from diseased and non-diseased subjects. Let $Z_1 = X^D$, where X^D is the covariate value from a diseased subject, and $Z_2 = X^{\bar{D}} - X^D$. In the notation of Section 2, $X_{ij} = (Z_1, Z_2)$. We fit the model $\Phi^{-1}(\theta_{Z_1,Z_2}) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2$. When $X^{\bar{D}} = X^D$, $Z_2 = 0$, and thus the parameter β_1 quantifies the effect of a common value of X on the AUC. Across sample sizes ranging from 30 to 200 per group, estimation is reasonable (Table 2). The largest amount of bias, $\hat{\beta}_1$, is 6% for a sample size of 30 per group, and bias diminished with increasing sample size. The bootstrapped standard error estimates tended to slightly overestimate the truth, except for a sample size of 30 per group. Coverage probability for confidence intervals using bootstrap standard errors is near the nominal level, although it is anticonservative for $n = 30$.

6. ASSESSMENT OF HEARING LOSS DEVICE

We apply our methodology to a study designed to evaluate the hearing device described in Section 1. The other AUC methods are not applicable, because one of the covariates is con-

Table 1. Bias and Efficiency Comparison of Three AUC Regression Methods for Balanced and Unbalanced Covariates With 100 Samples Each from States D and $\bar{D}$ Under the Model Described in Section 5.1, with $g = \Phi^{-1}$

	Balanced design			*Unbalanced design*		
Method	*Proposed*	*Derived*	*Jackknife*	*Proposed*	*Derived*	*Jackknife*
$\hat{\beta}_1$	.326	.338	.341	.329	.332	.360
% Bias	2.0	5.5	6.6	2.7	3.7	12.6
Relative efficiency	1	.88	.43	1	.76	.14

NOTE: True $\beta_1 = .320$, the balanced design sampled equal numbers at each covariate level. The unbalanced design sampled 50%, 10%, 10%, 10%, and 20% within covariate levels X = 1, 2, 3, 4, 5. Results represent 1,000 realizations from the model.

Table 2. Bias in Parameter and Bootstrapped Standard Error Estimates, and Coverage Probability for Confidence Intervals Under the Model Described in Section 5.2

Sample size (per group)	*Mean* $\hat{\beta}_1$	*Percent* bias	*Bootstrap* SE	*True* SE	*Percent* bias	*Coverage* 95% CI
30	.442	6.3%	.166	.180	−8.0%	.930
50	.433	4.1%	.140	.133	5.2%	.950
100	.427	2.5%	.090	.086	5.1%	.955
200	.417	.2%	.062	.060	3.0%	.953

NOTE: CIs were computed assuming normality with bootstrapped standard error (SE) estimates. Results represent 1,000 realizations of the model and 200 bootstrap samples each.

tinuous. The data presented are from a study of 105 hearing-impaired and 103 normally hearing subjects who were examined at three frequency and three intensity settings of the DPOAE device, resulting in a total of nine combinations of settings. The effect of severity of hearing impairment is also of interest. Data are analyzed from measurements taken on one ear per subject, although the method could be used if results were provided on both ears. The gold standard method for diagnosing impairment is a behavioral test in which subjects indicate whether a sound is audible for a range of frequencies until a hearing threshold is determined, and was conducted on each ear.

For estimation, pairing of covariates has been accomplished by design, because the frequency and intensity covariates were stratified, and the severity covariate applies to the impaired group only. The model of interest is $\log(AUC/1 - AUC) = \beta_0 + \beta_1 int + \beta_2 freq + \beta_3 sev$, where *int* is stimulus intensity (per 10 dB SPL), *freq* is stimulus frequency (per 100 Hz), and *sev* is severity of impairment, so that positive values indicate impairment in units of 10 dB SPL. Confidence interval estimates assume a normal distribution. We use the bootstrap, resampling by subject because of the repeated measures, to obtain standard error estimates. The model estimates indicate that the AUC odds decrease by 42% for every 10 dB increase in stimulus intensity (AUC odds, .58; 95% confidence interval, CI, .43, .79) and that the AUC odds increase 85% for every 10 dB worsening in impairment (AUC odds, 1.85; 95% CI, 1.49, 2.50), indicating that DPOAE discriminates severely impaired ears from normal ears better than mildly impaired ears from normal ears. Lastly, increasing the frequency setting appears to increase the AUC odds by 7% for every 100 Hz increase (AUC odds, 1.07; 95% CI, .99, 1.16), but this result is not statistically significant.

Graphical methods, such as plots of fitted versus empirical AUCs, were used to evaluate model fit (Fig. 1). Severity was categorized into four categories. Note the cloud of points in the upper right quadrant [Fig. 1(a)]. Plots of frequency for fixed severity and intensity suggested a lack of fit (not shown). Hence the model was refit with frequency as dummy variables and the fit is somewhat better [Fig. 1(b)]. Finally, jackknife procedures were used to identify influential points. Removal of one subject's observations was found to decrease the frequency coefficient substantially, further increasing our wariness about interpreting the relationship between this covariate and accuracy.

In conclusion, this analysis suggests that to achieve greater accuracy, stimuli with lower intensities should be used. Severity of impairment is an important determinate of accuracy and should be incorporated into decisions regarding the use of this device. The results are by no means conclusive about the association between the AUC and stimulus frequency. These data suggest that the relationship likely is not linear, but more data are necessary for its characterization. Finally, note that although

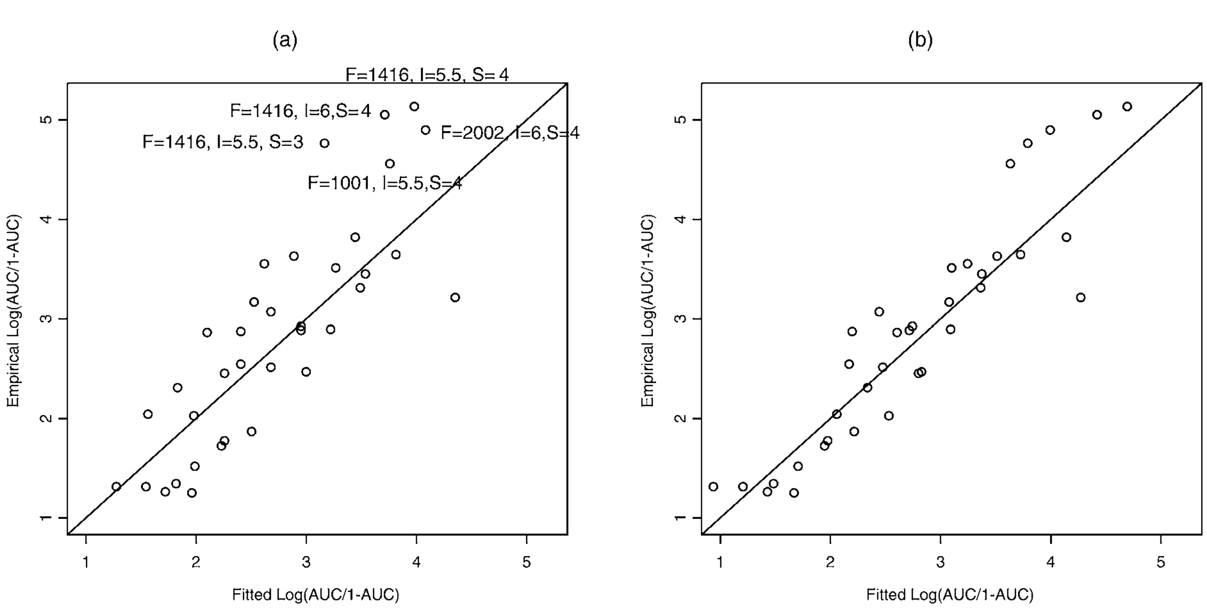

Figure 1. (a) Empirical Versus Fitted AUCs on the Log AUC-Odds Scale With Frequency as Continuous (F, Frequency; I, Intensity; S, Severity Category) and (b) Empirical Versus Fitted Log AUC-Odds With Frequency as Dummy Variables.

the AUC odds interpretation is succinct, ascribing value to a parameter requires a more general, decision-theoretic framework that establishes a clinically meaningful change in odds.

7. DISCUSSION

We have proposed a method for evaluating covariate effects on the AUC. The AUC is a measure of separation between the distributions of two random variables that is well established in diagnostic testing. It was recently proposed with different nomenclature by Fine and Bosch (2000) for use in toxicology and by Foulkes and De Gruttola (2002) for predicting human immunodeficiency virus resistance to antiretroviral therapy. Because the AUC is the Mann–Whitney U statistic, it is recognized as a monotone function of the Wilcoxon two-sample test statistic. In this sense, the AUC is already often used in clinical trials for comparing study arms when the outcome measure is continuous. We believe that the regression methods proposed herein may also find application outside of diagnostic testing. For example, AUC regression could be used to explore interactions between covariates and treatment effect in clinical trials. Other applications may extend more broadly to the optimization of classifiers such as evolutionary algorithms, support vector machines, and neural networks.

Measures other than the AUC can also be used to summarize the separation between random variables Y^D and $Y^{\bar{D}}$. However, we have shown that regression methods for the AUC is particularly simple, because it is based on binary regression algorithms for indicator variables of the form $I(Y^D > Y^{\bar{D}})$. A related method is under development for modeling the partial AUC $= \int_0^t ROC(t)\,dt$, a summary index that is gaining popularity particularly in disease screening applications. Binary regression methods can also be adapted for this purpose (Dodd 2001). Regression methods for other ROC summary indices have not been proposed.

Alternative approaches to ROC regression include that of Pepe (1997), which stipulates a regression model for the ROC curve, and that stemming from work by Tosteson and Begg (1987), which models the probability distributions for the test results Y^D and $Y^{\bar{D}}$. The latter approach, modeling probability distributions, requires the strongest assumptions, whereas Pepe's approach, which models the *relationship* between those distributions as characterized by the ROC curve, requires fewer assumptions. Our approach requires fewer assumptions still, because covariate effects on a summary index need only be specified. In future work we will investigate whether this leads to robustness for our approach over others. (See Pepe 1998 for discussion of the attributes of different approaches to ROC regression methods.)

In conclusion, we have proposed a new method for making inference about covariate effects on the performance of a classifier. Advantages of this approach are that it can be simply applied by adapting standard binary regression methods, it requires fewer assumptions than existing ROC regression methods, it is the only AUC regression method that can deal with continuous covariates, asymptotic distribution theory is established and, as a special case, it reduces to standard methods for comparing two ROC curves. Simulation studies show good small-sample performance for inferential procedures, and in an example we found that the methods leads to important insights into the performance of a hearing test. Further applications of the method to real data will eludicate the value of the method in practice.

APPENDIX: PROOFS

Here we provide proofs of the lemmas and Theorem 3. Lemmas 1 and 2 help establish a method of inference for the proposed method. However, because parametric assumptions are necessary to obtain variance estimates, in practice we recommend bootstrapping. Theorem 3 analytically demonstrates an equivalence with existing approaches in a restricted setting.

A.1 Proof of Lemma 1

We show that under (C1)–(C6), if the sum $\frac{1}{n_D n_{\bar{D}}} \frac{\partial}{\partial \beta} \mathbf{S}_N(\beta) \underset{p}{\rightarrow} E\frac{\partial}{\partial \beta}\mathbf{S}_{ij}(\beta)$ as $N \to \infty$, then convergence of $(n_D n_{\bar{D}})^{-1}\partial \mathbf{S}_N(\beta)/\partial\beta$ to its expectation is uniform for $\beta \in N_\delta(\beta_0)$. First, we find a finite union of intervals with known length that cover $N_\delta(\beta_0)$. For $\psi > 0$, define intervals $C_k = (\beta_k, \beta_{k+1})$ such that $|\beta_{k+1} - \beta_k| < \psi$, and a finite union of these intervals, $\bigcup_{k=1}^K C_k$ covers $N_\delta(\beta_0)$. The triangle inequality gives the following:

$$
\begin{aligned}
&\sup_{\beta \in N_\delta(\beta_0)} \left| \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta)}{\partial \beta} - E\left\{ \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta)}{\partial \beta} \right\} \right| \\
&\quad = \max_k \sup_{\beta \in C_k} \left| \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta)}{\partial \beta} - \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} \right. \\
&\qquad\qquad + E\left\{ \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} \right\} - E\left\{ \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta)}{\partial \beta} \right\} \\
&\qquad\qquad \left. + \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} - E\left\{ \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} \right\} \right| \\
&\quad \le \max_k \sup_{\beta \in C_k} \left| \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta)}{\partial \beta} - \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} \right| \\
&\qquad + \max_k \sup_{\beta \in C_k} \left| E\left\{ \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} \right\} - E\left\{ \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta)}{\partial \beta} \right\} \right| \\
&\qquad + \max_k \sup_{\beta \in C_k} \left| \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} - E\left\{ \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} \right\} \right| \\
&\quad = A_{1,N} + A_{2,N} + A_{3,N}.
\end{aligned}
\tag{A.1}
$$

The mean value theorem gives the following result for the first term in (A.1):

$$
\begin{aligned}
A_{1,N} &= \max_k \sup_{\beta \in C_k} \left| \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta)}{\partial \beta} - \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} \right| \\
&= \frac{1}{n_D n_{\bar{D}}} \max_k \sup_{\beta \in C_k} (\beta - \beta_k) \frac{\partial}{\partial \beta} \frac{\partial \mathbf{S}_N(\beta^\star)}{\partial \beta}, \\
&\qquad \text{for } \beta^* \in (\beta, \beta_k) \\
&< \psi M_1, \qquad \text{where } M_1 < \infty,
\end{aligned}
$$

because the largest interval length is ψ and the derivative is assumed to be uniformly bounded by M_1 for $\beta \in N_\delta(\beta_0)$. The mean value theorem and the uniform boundedness of $\frac{\partial}{\partial \beta} E \partial \mathbf{S}_N(\beta^\star)/\partial\beta$ similarly imply $A_{2,N} < \psi M_2$, where $M_2 < \infty$. Finally, because $(n_D n_{\bar{D}})^{-1}\partial \mathbf{S}_N(\beta_k)/\partial\beta$ converges in probability to its expectation, for a given k, we can find an N_ϵ such that when $N > N_\epsilon$,

$$
\Pr\left[\frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} - E\left\{ \frac{1}{n_D n_{\bar{D}}} \frac{\partial \mathbf{S}_N(\beta_k)}{\partial \beta} \right\} > \epsilon/2 \right] < \gamma/K.
$$

That is, for $\epsilon > 0$ and $\gamma > 0$,

$$\begin{aligned}
\Pr\left[\max_k \sup_{\beta\in C_k}\left|\frac{1}{n_D n_{\bar D}}\frac{\partial \mathbf{S}_N(\beta_k)}{\partial\beta} - E\left\{\frac{1}{n_D n_{\bar D}}\frac{\partial \mathbf{S}_N(\beta_k)}{\partial\beta}\right\}\right| > \epsilon/2\right]
&= \Pr\left[\max_k \left|\frac{1}{n_D n_{\bar D}}\frac{\partial \mathbf{S}_N(\beta_k)}{\partial\beta} - E\left\{\frac{1}{n_D n_{\bar D}}\frac{\partial \mathbf{S}_N(\beta_k)}{\partial\beta}\right\}\right| > \epsilon/2\right]\\
&< \sum_k \Pr\left[\left|\frac{1}{n_D n_{\bar D}}\frac{\partial \mathbf{S}_N(\beta_k)}{\partial\beta} - E\left\{\frac{1}{n_D n_{\bar D}}\frac{\partial \mathbf{S}_N(\beta_k)}{\partial\beta}\right\}\right| > \epsilon/2\right]\\
&< \sum_k \gamma/K = \gamma \text{ eventually.}
\end{aligned}$$

Choose ψ such that $(M_1 + M_2)\psi < \epsilon/2$, it follows that $\Pr(A_{1,N} + A_{2,N} + A_{3,N} > \epsilon/2 + \epsilon/2) < \gamma$, for large N.

A.2 Proof of Lemma 2

To establish convergence in probability, consider the term $E(\partial\mathbf{S}_{ij}(\beta)/\partial\beta | Y_i^D)$, which is random with respect to Y_i^D and independent across all i. By the triangle inequality,

$$\begin{aligned}
\Pr\left[\left|\frac{1}{n_D n_{\bar D}}\frac{\partial}{\partial\beta}\mathbf{S}_N(\beta) - E\frac{\partial \mathbf{S}_{ij}(\beta)}{\partial\beta}\right| > \epsilon\right]
&= \Pr\left[\left|\frac{1}{n_D n_{\bar D}}\frac{\partial}{\partial\beta}\mathbf{S}_N(\beta) - \frac{1}{n_D}\sum_i E\left\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)|Y_i^D\right\}\right.\right.\\
&\qquad \left.\left. + \frac{1}{n_D}\sum_i E\left\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)|Y_i^D\right\} - E\frac{\partial \mathbf{S}_{ij}(\beta)}{\partial\beta}\right| > \epsilon\right]\\
&\le \Pr\left[\left|\frac{1}{n_D n_{\bar D}}\frac{\partial}{\partial\beta}\mathbf{S}_N(\beta) - \frac{1}{n_D}\sum_i E\left\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)|Y_i^D\right\}\right| > \epsilon/2\right]\\
&\qquad + \Pr\left[\left|\frac{1}{n_D}\sum_i E\left\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)|Y_i^D\right\} - E\frac{\partial \mathbf{S}_{ij}(\beta)}{\partial\beta}\right| > \epsilon/2\right].
\end{aligned} \tag{A.2}$$

Consider the first term on the right side of the inequality in (A.2):

$$\begin{aligned}
E\left|\frac{1}{n_D n_{\bar D}}\frac{\partial}{\partial\beta}\mathbf{S}_N(\beta) - \frac{1}{n_D}\sum_i E\left\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)|Y_i^D\right\}\right|
&= E\left|\frac{1}{n_D}\sum_i \frac{1}{n_{\bar D}}\sum_j \frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta) - E\left\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)|Y_i^D\right\}\right|\\
&\le \frac{1}{n_D}\sum_i E\left|\frac{1}{n_{\bar D}}\sum_j \frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta) - E\left\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)|Y_i^D\right\}\right|.
\end{aligned} \tag{A.3}$$

The terms inside the expectation in (A.3) are iid across j for fixed i. Hence, by the weak law of large numbers (WLLN), (A.3) $\overrightarrow{p}$ 0. Because convergence in mean implies convergence in probability, it follows that for all $\epsilon > 0$, $\Pr[|\frac{1}{n_D n_{\bar D}}\frac{\partial}{\partial\beta}\mathbf{S}_N(\beta) - \frac{1}{n_D}\sum_i E\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)| Y_i^D\}| > \epsilon/2] \to 0$ as $N \to \infty$.

Now consider the second term on the right side in (A.2). The terms $E\{\partial\mathbf{S}_{ij}(\beta)/\partial\beta \mid Y_i^D\}$ are independent and have finite expectation. From the weak law of large numbers,

$$\frac{1}{n_D}\sum_i E\{\partial\mathbf{S}_{ij}(\beta)/\partial\beta \mid Y_i^D\} \overrightarrow{p} E\{\partial\mathbf{S}_{ij}(\beta)/\partial\beta\}.$$

Therefore, for $\epsilon > 0$, $\Pr[|\frac{1}{n_D}\sum_i E\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)|Y_i^D\} - E\{\frac{\partial}{\partial\beta}\mathbf{S}_{ij}(\beta)\}| > \epsilon/2] \to 0$ as $N \to \infty$. Hence the two terms in (A.2) $\overrightarrow{p}$ 0, and the result follows.

A.3 Proof of Theorem 3

We show that the least squares estimators from the proposed, derived-variables, and jackknife-AUC methods are the same. To simplify, we assume that $n_D = n_{\bar D} = n$, although the result holds for any n_D and $n_{\bar D}$, as long as they do not vary with k. Recall that test results of nondiseased and diseased subjects are paired within a given covariate level.

For the proposed model, $E(U_{ijk}) = \beta_0^P + \beta_1^P X_k$, let $U_{ijk} \equiv I(Y_{ik}^D > Y_{jk}^{\bar D})$, $\overline{U} \equiv \frac{1}{Kn^2}\sum_{ijk} U_{ijk}$, $\overline{X} \equiv \frac{1}{2nK}\sum_{ijk} X_{ijk} = \frac{1}{K}\sum_k X_k$, $S(U, X) \equiv \sum_{ijk} U_{ijk}X_k - Kn^2\overline{U}\overline{X}$, and $S(X, X)^P \equiv \sum_{ijk} X_{ijk}^2 - Kn^2\overline{X} = n^2\{\sum_k X_k^2 - K\overline{X}\}$. The least squares estimators are given by

$$\hat\beta_0^P = \overline{U} - \hat\beta_1^P\overline{X} \qquad \text{and} \qquad \hat\beta_1^P = \frac{S(U, X)}{S(X, X)^P}.$$

First, we show the estimators from the derived-variables method, with model $E(\widehat{AUC}_k) = \beta_0^d + \beta_1 X_k^d$, and are the same. Observe that $\overline{\widehat{AUC}_k} \equiv \frac{1}{Kn^2}\sum_{ijk} U_{ijk} = \overline{U}$ and $S(X, X)^d \equiv \sum_k X_k^2 - K\overline{X}^2 = \frac{1}{n^2} S(X, X)^P$. A little algebra shows that $S(\widehat{AUC}, X) \equiv \sum_k(\widehat{AUC}_k X_k) - K\overline{\widehat{AUC}}\overline{X} = \frac{1}{n^2}S(U, X)$. It follows that $\hat\beta_1^d = S(\widehat{AUC}, X)/ S(X, X)^d = \hat\beta_1^P$ and $\hat\beta_0^d = \hat\beta_0^P$.

The jackknifed-AUC model is $E(A_{lk}) = \beta_0^J + \beta_1^J X_k$, where A_{lk} denotes the jackknifed-AUC value at the kth covariate level for $l = 1, \ldots, 2n$. We use A_{lk} to denote a jackknifed AUC value from the combined vector, $(A_{1k}^{\bar D}, \ldots, A_{n_{\bar D}k}^{\bar D}, A_{(n_{\bar D}+1)k}^D, \ldots, A_{(n_{\bar D}+n_D)k}^D)$. We also denote the vector as $\{A_{ik}^{\bar D} : i = 1, \ldots, n_D, A_{jk}^D : j = 1, \ldots, n_{\bar D}\}$. Note that the superscript in $A_{ik}^{\bar D}$ indicates that this term is averaged across all nondiseased observations and random with respect to a given observation from the diseased observations.

The least squares estimators from the jackknifed-AUC model depend on the random variables $\{A_{lk} : l = 1, \ldots, 2n\}$. Observe that $\overline{X}^J \equiv \frac{1}{K2n}\sum_{k=1}^K\sum_{l=1}^{2n} X_{lk} = \overline{X}$ and $S(X, X)^J \equiv \sum_{k=1}^K\sum_{l=1}^{2n} X_{lk}^2 - 2nK\overline{X} = 2n(\sum_k X_k^2 - K\overline{X}) = \frac{2}{n}S(X, X)^P$. Next, we show that $\overline{A} \equiv \frac{1}{K2n}\sum_{k=1}^K\sum_{l=1}^{2n} A_{lk}$ equals $\overline{U}$. The mean of the jackknifed AUC at covariate level k can be written as $\overline{A_k} = \frac{1}{2n}\sum_l A_{lk} = \frac{1}{2n}\sum_{i=1}^n A_{ik}^{\bar D} + \frac{1}{2n}\sum_{j=1}^n A_{jk}^D$. Define $\widehat{F}_k^{\bar D}(Y_{ik}^D) = \frac{1}{n}\sum_j I(Y_{ik}^D > Y_{jk}^{\bar D})$ and $\widehat{F}_k^D(Y_{jk}^{\bar D}) = \frac{1}{n}\sum_i I(Y_{ik}^D > Y_{jk}^{\bar D})$, where $\widehat{F}$ is the empirical cdf. Note that these are the empirical placement value estimators. To illustrate the relationship between the U_{ijk} terms and A_{lk}, we use a result from Hanley and Haijan-Tilaki (1997),

$$A_{ik}^{\bar D} = \frac{2n-1}{n-1}\widehat{F}_k^{\bar D}(Y_{ik}^D) - \frac{n}{n-1}\widehat{AUC}_k \qquad \text{and}$$

$$A_{jk}^D = \frac{2n-1}{n-1}\widehat{F}_k^D(Y_{jk}^{\bar D}) - \frac{n}{n-1}\widehat{AUC}_k.$$

The mean of the $A_{ik}^{\bar D}$'s is given by

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n A_{ik}^{\bar D} &= \frac{1}{n}\sum_{i=1}^n\left\{\frac{2n-1}{n-1}\widehat{F}_k^{\bar D}(Y_{ik}^D) - \frac{n}{n-1}\widehat{AUC}_k\right\}\\
&= \frac{2n-1}{n(n-1)}\sum_{i=1}^n \widehat{F}_k^{\bar D}(Y_{ik}^D) - \frac{n}{n-1}\widehat{AUC}_k\\
&= \frac{2n-1}{n-1}\widehat{AUC}_k - \frac{n}{n-1}\widehat{AUC}_k = \widehat{AUC}_k.
\end{aligned}$$

Using a similar argument, the mean of the A_{jk}^D's can be shown to equal $\widehat{AUC}_k$. Hence $\overline{A_k} = \widehat{AUC}_k$ and $\overline{A} = \frac{1}{k}\sum_k \widehat{AUC}_k = \overline{U}$. Now consider the term $S(A, X)$,

$$\begin{aligned} S(A,X) &\equiv \sum_{k=1}^{K}\sum_{l=1}^{2n} A_{lk}X_k - 2nK\overline{UX} \\ &= \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n} A_{ik}^{\overline{D}}X_k}_{\text{(c)}} + \underbrace{\sum_{k=1}^{K}\sum_{j=1}^{n} A_{jk}^{D}X_k}_{\text{(d)}} - 2Kn\overline{A}\,\overline{X}. \qquad \text{(A.4)} \end{aligned}$$

The term (c) in the expression in (A.4) is equal to

$$\begin{aligned} \sum_{k=1}^{K}\sum_{i=1}^{n} A_{ik}^{\overline{D}}X_k &= \sum_{k=1}^{K}\sum_{i=1}^{n}\left(\frac{2n-1}{n-1}\widehat{F}_k^{D}(Y_{ik}^{\overline{D}}) - \frac{n}{n-1}\widehat{AUC}_k\right)X_k \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{2n-1}{n-1}\left(\frac{1}{n}\sum_{j=1}^{n} U_{ijk}X_k\right) \\ &\quad - \frac{n^2}{n-1}\sum_{k}\widehat{AUC}_k X_k \\ &= \frac{2n-1}{n(n-1)}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ijk}X_k \\ &\quad - \frac{n^2}{n-1}\sum_{k}\widehat{AUC}_k X_k. \qquad \text{(A.5)} \end{aligned}$$

In a similar manner, we can show that (d) in (A.4) equals the expression shown in (A.5), and that expression (A.4) equals

$$\begin{aligned} &2\left\{\frac{2n-1}{n(n-1)}\sum_{ijk} U_{ijk}X_k - \frac{n^2}{n-1}\sum_{k}\widehat{AUC}_k X_k\right\} - 2Kn\overline{UX} \\ &\quad = 2\left\{\frac{2n-1}{n(n-1)}\sum_{ijk} U_{ijk}X_k - \frac{1}{n-1}\sum_{ijk} U_{ijk}X_k\right\} - 2Kn\overline{UX} \\ &\quad = \frac{2}{n}\sum_{ijk} U_{ijk}X_k - Kn\overline{UX} \\ &\quad = \frac{2}{n}S(U,X)^P. \end{aligned}$$

The least squares estimators for the jackknife AUC method are $\hat\beta_1^J = S(A,X)/S(X,X)^J = \hat\beta_1^P$ and $\hat\beta_0^J = \overline{A} - \hat\beta_1^J\overline{X}^J = \hat\beta_0^P$.

[Received April 2002. Revised November 2002.]

REFERENCES

Bamber, D. (1975), “The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph,” *Journal of Mathematical Psychology*, 12, 387–415.

Brusic, V., Rudy, G., Honeyman, M., Hammer, J., and Harrison, L. (1998), “Prediction of MHC Class II–Binding Peptides Using an Evolutionary Algorithm and Artificial Neural Network,” *Bioinformatics*, 14, 121–130.

DeLong, E. R., DeLong, D., and Clarke-Pearson, D. (1988), “Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach,” *Biometrics*, 44, 837–845.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002), *Analysis of Longitudinal Data*, Oxford, U.K.: Oxford University Press.

Dodd, L. (2001), “Regression Methods for Areas and Partial Areas Under the Receiver-Operating Characteristic Curve,” unpublished doctoral thesis, University of Washington.

Dorfman, D., Berbaum, K., and Metz, C. (1992), “Receiver Operating Characteristic Analysis: Generalization to the Population of Readers and Patients With the Jackknife Method,” *Investigative Radiology*, 27, 723–731.

Fine, J., and Bosch, R. (2000), “Risk Assessment via a Robust Probit Model With Application to Toxicology,” *Journal of the American Statistical Association*, 95, 375–382.

Foulkes, A., and De Gruttola, V. (2002), “Characterizing the Relationship Between HIV-1 Genotype and Phenotype: Prediction-Based Classification,” *Biometrics*, 58, 145–156.

Foutz, R. (1977), “On the Unique Solution to the Likelihood Equations,” *Journal of the American Statistical Association*, 72, 147–148.

Hanley, J., and Hajian-Tilaki, K. (1997), “Sampling Variability of Nonparametric Estimates of the Areas Under Receiver Operating Characteristic Curves: An Update,” *Academic Radiology*, 17, 49–58.

McCullagh, P., and Nelder, J. (1997), *Generalized Linear Models* (2nd ed.), New York: Chapman & Hall.

Pepe, M. (1997), “A Regression Modelling Framework for Receiver Operating Characteristic Curves in Medical Diagnostic Testing,” *Biometrika*, 84, 595–608.

——— (1998), “Three Approaches to Regression Analysis of Receiver Operating Characteristic Curves for Continuous Test Results,” *Biometrics*, 54, 124–135.

——— (2000), “Receiver Operating Characteristic Methodology,” *Journal of the American Statistical Association*, 95, 307–311.

Shapiro, D. (1999), “The Interpretation of Diagnostic Tests,” *Statistical Methods in Medical Research*, 8, 113–134.

Stover, L., Gorga, M., and Neely, S. (1996), “Toward Optimizing the Clinical Utility of Distortion Product Otoacoustic Emission Measurements,” *Journal of the Acoustical Society of America*, 100, 956–967.

Thompson, M. L., and Zucchini, W. (1989), “On the Statistical Analysis of ROC Curves,” *Statistics in Medicine*, 8, 1277–1290.

Tosteson, A. N., and Begg, C. B. (1988), “A General Regression Methodology for ROC Curve Estimation,” *Medical Decision Making*, 8, 204–215.

Zhu, H., Beling, P., and Overstreet, G. (2002), “A Bayesian Framework for the Combination of Classifier Outputs,” *Journal of the Operational Research Society*, 53, 719–727.