1-27-2003

Semi-parametric Regression for the Area Under the Receiver Operating Characteristic Curve

Lori E. Dodd
National Institute of Health, doddl@mail.nih.gov

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

1 INTRODUCTION

The performance of a binary classifier with continuous output is often evaluated with Receiver Operating Characteristic (ROC) Curve analysis (Zhu et al., 2002; Brusic et al. 2002; Pepe, 2000). For two states D and $\bar{D}$, which are typically diseased and non-diseased states in medicine, and classifier output Y, let $Y > c$ indicate classification into state D. The ROC curve plots $(P(Y > c \mid \bar{D}), P(Y > c \mid D))$ 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 towards $(0, 1)$ indicates better performance. When under development, a classifier's optimal threshold is not known. Since 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 as $P(Y^D > Y^{\bar{D}})$, 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, since $AUC = \int_0^1 ROC(t)dt$. A perfect classifier has $AUC = 1$, while 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 paper, 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 et al., 1996). The strength of the DPOAE output, measured by DPOAE amplitude, indicates auditory function. Since 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 too mentally disabled for the behavioral gold standard test.

One goal of the study was to determine if 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 better diagnoses the most severely impaired ears than those with mild impairment. Exploration of the relationship between severity of impairment and diagnostic accuracy yields information about the types of cases 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 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 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 modelling of ROC curves have been proposed (see Pepe, 1998 for a review), and we will contrast them briefly with AUC regression in Section 7. First, we develop our approach.

2 AUC BINARY REGRESSION

2.1 The Model

Although D and $\bar{D}$ may be any two states, we use terminology from diagnostic testing for them, so D is referred to as "disease" and $\bar{D}$ is referred to as "non-disease." We use X to denote covariates and Y to denote classifier output. Let (Y_i^D, X_i^D) and $(Y_j^{\bar{D}}, X_j^{\bar{D}})$ denote observations from D and $\bar{D}$, with $(i = 1, ..., n_D)$ and $(j = 1, ..., n_{\bar{D}})$, respectively. The result of Bamber (1975) suggests that we can write the covariate-specific AUC as $P(Y_i^D > Y_j^{\bar{D}} |\ X_i^D, X_j^{\bar{D}}) \equiv \theta_{ij}$. The parameter θ_{ij} compares the results from diseased population with covariates X_i^D to those from non-diseased with covariates $X_j^{\bar{D}}$. To simplify notation, let X_{ij} denote $(X_i^D, X_j^{\bar{D}})$, or a specified function of them. For a vector of parameters β and a monotone increasing link function g, we propose the following AUC regression model:

$$g(\theta_{ij}) = X_{ij}^T \beta. \tag{1}$$

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 $AUC/(1 - AUC) = P(Y^D > Y^{\bar{D}})/P(Y^D < Y^{\bar{D}})$. Since larger AUCs are associated with increasing accuracy, AUC odds greater than one 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 non-

diseased 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 $\beta_1 > 0$, the test is better at distinguishing between diseased and non-diseased women than between diseased and non-diseased men.

When covariates are specific to the diseased group (e.g., stage of disease), the AUC is modelled as a function of the covariate X_i^D. That is, the covariate-specific AUC is defined as $P(Y_i^D > Y_j^{\bar{D}} | X_i^D) \equiv \theta_i$. 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 non-diseased groups changes. Consider, for example, the covariate *age*. Computation of an AUC for diseased subjects of age 80 and non-diseased subjects of age 50 is not scientifically relevant, while an AUC for diseased and non-diseased subjects both of age 80 (*or* of age 50) is of interest. The goal is to understand how the AUC, *for diseased and non-diseased subjects of the same age*, changes as age varies. The parameter β_1 in the model $logit(\theta_{ij}) = \beta_0 + \beta_1 X_i^D + \beta_2 (X_i^D - X_j^{\bar{D}})$ describes this relationship. If the covariate is age in years, $exp(\beta_1)$ is the ratio of AUC odds associated with a one-year increase in age for diseased and non-diseased subjects of the same age. If this value is greater than one, 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 $U_{ij} = I(Y_i^D > Y_j^{\bar{D}})$, and let $N = n_D + n_{\bar{D}}$. Note that $E(U_{ij} | X_{ij}) = P\left(Y_i^D > Y_j^{\bar{D}} \mid X_{ij}\right) = \theta_{ij}$. This suggests that our model, $g(\theta_{ij}) = X_{ij}^T \beta$, is a generalized linear regression model for the binary variables U_{ij}. The

following estimating function:

$$S_N(\beta) = \sum_{i}^{n_D} \sum_{j}^{n_{\bar{D}}} \frac{\partial \theta_{ij}}{\partial \beta} \nu(\theta_{ij})^{-1} (U_{ij} - \theta_{ij}) \equiv \sum_{i}^{n_D} \sum_{j}^{n_{\bar{D}}} S_{ij}(\beta) \tag{2}$$

is the classic estimating function for binary regression, except the U_{ij}'s are not independent. The term $\partial\theta_{ij}/\partial\beta$ is a $(p \times 1)$ vector of the partial derivatives of θ_{ij} with respect to the model parameters β. The term $\nu(\theta_{ij})$ is the variance function, while 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 i^{th} 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 observations are independent, and, to borrow language from Generalized Estimating Equations (GEE), 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 non-diagonal WCM often fails in diagnostic testing applications with repeated measures and would result in inconsistent estimates (Diggle et al., 2002; pg.255). Furthermore, in applications in which the above condition is met, the dimensionality of the non-diagonal WCM may be prohibitively large. For example, in the application here the matrix would be of dimension 72708×72708.

2.3 Implementation

Data are observed as follows: $\{(Y_1^D, X_1^D), ..., (Y_{n_D}^D, X_{n_D}^D), (Y_1^{\bar{D}}, X_1^{\bar{D}}), ..., (Y_{n_{\bar{D}}}^{\bar{D}}, X_{n_{\bar{D}}}^{\bar{D}})\}$. Section 2.2 suggests that all pairs are included in (2), but one only needs to include (and model) subsets of pairs. First, note that if covariates are categorical and there are sufficient observations at each covariate level, 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, X_i^D) and $(Y_j^{\bar{D}}, X_j^{\bar{D}})$ must be created for subjects with different covariate values. It may not be appropriate to pair (Y_i^D, X_i^D) and $(Y_j^{\bar{D}}, X_j^{\bar{D}})$ for all (i, j), as it 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, e.g., create a pair if $|X_i^D - X_j^{\bar{D}}| \leq \zeta$. If covariates for the $(i, j)^{th}$ pair are farther than ζ apart, 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 $|X_i^D - X_j^{\bar{D}}| \leq \zeta$. The number of pairs depends on ζ and the distribution of covariates. For a given i the number of observations from non-diseased subjects paired with Y_i^D is denoted $n_{\bar{D}}(\zeta, i)$. Here, the estimating function is the sum $\sum_i^{n_D} \sum_j^{n_{\bar{D}}(\zeta,i)} S_{ij}(\beta)$. 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 non-diseased results. There is a trade-off between bias and efficiency as ζ varies. For a small ζ, much of the data is 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 enough 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 zero. If the link function is chosen to be the identity, closed-form expressions for $\hat{\beta}$ 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 $\zeta = \infty$ 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 j^{th} result of non-diseased, and similaly for $n_{\bar{D}}(\zeta, i)$. Then, as long as $n_D(\zeta, j) = O(N)$ and $n_{\bar{D}}(\zeta, j) = O(N)$ the theory applies. If ζ is fixed and does not get smaller as N increases, these conditions should be satisfied. In other words, as long as each diseased is paired with a proportion of the non-diseased subjects the theory outlined below applies.

To derive theory, we assume the following conditions (C1) $\{(Y_i^D, X_i^D) : i = 1, ..., n_D\}$ are i.i.d., $\{(Y_j^{\bar{D}}, X_j^{\bar{D}}) : j = 1, ..., n_{\bar{D}}\}$ are i.i.d., and both vectors are mutually independent; (C2) $\lim_{N\to\infty} n_D/N \to \lambda$, where $0 < \lambda < 1$ and $N = n_D + n_{\bar{D}}$; (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 $\beta \in N_\delta(\beta_0) \equiv \{\beta : \| \beta - \beta_0 \| < \delta\}$; (C5) the covariate space is bounded; (C6) the matrix $E(\partial S_{ij}(\beta_0)/\partial\beta)$ is negative definite.

It follows from (C3)-(C6) that $\frac{1}{n_D n_{\bar{D}}} \frac{\partial}{\partial\beta} S_N(\beta)$, $\frac{1}{n_D n_{\bar{D}}} \frac{\partial^2}{\partial\beta\partial\beta^T} S_N(\beta)$, and $\frac{\partial}{\partial\beta} E\left(\frac{1}{n_D n_{\bar{D}}} \frac{\partial}{\partial\beta} S_N(\beta)\right)$ are bounded uniformly for $\beta \in N_\delta(\beta_0)$. To see this one must show that each of the elements in $\frac{1}{n_D n_{\bar{D}}} \frac{\partial}{\partial\beta} S_N(\beta)$ and $\frac{1}{n_D n_{\bar{D}}} \frac{\partial^2}{\partial\beta\partial\beta^T} S_N(\beta)$ has a bound independent of β. The boundedness condition of $\frac{\partial}{\partial\beta} E\left(\frac{1}{n_D n_{\bar{D}}} \frac{\partial}{\partial\beta} S_N(\beta)\right)$ is slightly more involved, and requires demonstrating that its limit is equal to that of $E\left(\frac{1}{n_D n_{\bar{D}}} \frac{\partial^2}{\partial\beta\partial\beta^T} S_N(\beta)\right)$, whose bound does not depend on β. We refer to this as property (B). Proofs of lemmas are found in the appendix.

3.1 Consistency

Theorem 1. Under $(C1)-(C6)$, as $N \to \infty$, solutions to $S_N(\beta) = 0$ are unique with probability converging to 1 and $\hat{\beta} \xrightarrow{p} \beta_0$.

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) $\partial S_N(\beta)/\partial\beta$ exists and is continuous for $\beta \in N_\delta(\beta_0)$, (F2) $(n_D n_{\bar{D}})^{-1}\, \partial S_N(\beta)/\partial\beta \xrightarrow{p} E\left(\partial S_{ij}(\beta)/\partial\beta\right)$ uniformly for $\beta \in N_\delta(\beta_0)$ as $N \to \infty$, (F3) $(n_D n_{\bar{D}})^{-1}\partial S_N(\beta_0)/\partial\beta$ is negative definite with probability converging to one as $N \to \infty$, and (F4) $ES_N(\beta_0) = 0$. The assumptions listed above and the following two lemmas are sufficient for establishing the Foutz conditions.

Lemma 1. Under property (B), and if, for each fixed $\beta \in N_\delta(\beta_0)$, $\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial}{\partial\beta}S_N(\beta)\right)$ converges to $E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)\right)$ in probability as $N \to \infty$, then convergence of $\frac{1}{n_D n_{\bar{D}}}\frac{\partial}{\partial\beta}S_N(\beta)$ to $E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)\right)$ is uniform for $\beta \in N_\delta(\beta_0)$.

Lemma 2. $\frac{1}{n_D n_{\bar{D}}}\frac{\partial}{\partial\beta}S_N(\beta) \xrightarrow{p} E\frac{\partial S_{ij}(\beta)}{\partial\beta}$ as $N \to \infty$.

Condition (F1) follows trivially from the assumptions above by the existence of third derivatives of the elements of $S_N(\beta)$. 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' condition (F2) is satisfied. Condition (F3) follows since $(n_D n_{\bar{D}})^{-1}\partial S_N(\beta_0)/\partial\beta \xrightarrow{p} E(\partial S_{ij}(\beta_0)/\partial\beta)$ by Lemma 2, which by assumption is a negative definite matrix. Finally, since 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 $S_N(\beta)$ 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:

$$\begin{aligned} E\left(U_{ij}|Y_i^D = y_i^D, X_i^D, X_j^{\bar{D}}\right) &= E\left(I(y_i^D > Y_j^{\bar{D}})|X_i^D, X_j^{\bar{D}}\right) \\ &= P_{X_j^{\bar{D}}}(y_i^D > Y^{\bar{D}}) \equiv F^{\bar{D}}_{X_j^{\bar{D}}}(y_i^D). \end{aligned}$$

This notation denotes the probability of observing a value of y_i^D or lower in the distribution of test results of non-diseased that have covariate pattern $X_j^{\bar{D}}$. We refer to $1 - F^{\bar{D}}_{X_j^{\bar{D}}}(y_i^D)$ as placement values. They indicate the "place" the diseased observation has in the distribution of non-diseased test results with covariate pattern $X_j^{\bar{D}}$. For a given y_i^D, a value of $F^{\bar{D}}_{X_j^{\bar{D}}}(y_i^D)$ closer to 1 indicates that most of the non-diseased test results fall below it. Note that $E(F^{\bar{D}}_{X_j^{\bar{D}}}(Y_i^D)|X_i^D) = P(Y_i^D > Y_j^{\bar{D}}|X_i^D, X_j^{\bar{D}}) = \theta_{ij}$. If the Y^D's, on average, fall in the upper tail of the distribution of $Y^{\bar{D}}$, then the AUC will be larger.

An analogous entity is defined by conditioning on a non-diseased observation as follows: $E\big(U_{ij}|Y_j^{\bar{D}} = y_j^{\bar{D}}, X_i^D, X_j^{\bar{D}}\big) = (1 - F^D_{X_i^D}(y_j^{\bar{D}})) \equiv \bar{F}^D_{X_i^D}(y_j^{\bar{D}})$. The interpretation is similar to the placement value concept for y_i^D. We define the following sum:

$$S_{N,P}(\beta) = \sum_i^{n_D} \sum_j^{n_{\bar{D}}} \omega_{ij} \left\{ \left(F_{X_j^{\bar{D}}}(Y_i^D) - \theta_{ij} \right) + \left(\bar{F}_{X_i^D}(Y_j^{\bar{D}}) - \theta_{ij} \right) \right\}, \tag{3}$$

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

Theorem 2. Under (C1)-(C6), $\sqrt{\frac{n_D n_{\bar{D}}}{N}}(\hat{\beta} - \beta_0) \xrightarrow{d} Z \sim N(0, I(\beta_0)^{-1} \not{\Sigma} I(\beta_0)^{-1})$ as $N \to \infty$, where $I(\beta_0) \equiv -E\left(\frac{\partial}{\partial\beta} S_{ij}(\beta_0)\right)$ and

$$\begin{aligned}
\not{\Sigma} &= \lim_{N\to\infty} \left\{ \frac{n_D}{N} \left[\frac{1}{n_{\bar{D}}} \sum_j^{n_{\bar{D}}} \frac{1}{n_D^2} \sum_i^{n_D} \sum_k^{n_D} \omega_{ij}\omega_{kj}^T cov\left(\bar{F}^D_{X_i^D}(Y_j^{\bar{D}}), \bar{F}^D_{X_k^D}(Y_j^{\bar{D}}) \right) \right] \right\} \\
&\quad + \lim_{N\to\infty} \left\{ \frac{n_{\bar{D}}}{N} \left[\frac{1}{n_D} \sum_i^{n_D} \frac{1}{n_{\bar{D}}^2} \sum_j^{n_{\bar{D}}} \sum_l^{n_{\bar{D}}} \omega_{ij}\omega_{il}^T cov\left(F^{\bar{D}}_{X_j^{\bar{D}}}(Y_i^D), F^{\bar{D}}_{X_l^{\bar{D}}}(Y_i^D) \right) \right] \right\} \\
&\equiv \lambda \not{\Sigma}_{\bar{D}} + (1-\lambda) \not{\Sigma}_D. \qquad (4)
\end{aligned}$$

Observe that the asymptotic variance is comprised of one component that depends on variability in Y^D and another that depends on $Y^{\bar{D}}$, with each weighted by its relative contribution to the overall sample size. To obtain variance estimates, models for $F^D_{X^D}$ and $F^{\bar{D}}_{X^{\bar{D}}}$ 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 $F^D_{X^D}$ and $F^{\bar{D}}_{X^{\bar{D}}}$ this formula could be applied. The theory is extended to repeated measures data when the number of diseased and non-diseased subjects gets large. To show this, we identify all 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 the bootstrap to obtain appropriate standard errors.

4 RELATIONSHIPS WITH EXISTING METHODS

4.1 Comparing two AUCs

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 $U_{ijk} = I(Y_{ik}^D > Y_{jk}^{\bar{D}})$ and let $\nu(\theta_{ij}) = 1$. The estimating function is simply:

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

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

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

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

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

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

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

4.2 Comparison with existing AUC regression methods

4.2.1 Derived Variables Approach

Thompson and Zucchini (1989) propose 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 k^{th} covariate level as $\hat{\theta}_k$. The derived variables AUC regression model is given by:

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

Since 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 modelled. Further, since different numbers of subjects often contribute to AUC estimates across covariate levels, the regression assumption of equal variances will frequently may fail.

4.2.2 Jackknifed AUC Approach

Dorfman, Berbaum, and Metz (1992) propose 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 $\hat{\theta}_k$ and N_k denote, respectively, the AUC estimate and the total number of observations at the k^{th} covariate level. Jackknifed AUC values for the i^{th} 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 i^{th} 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 i^{th} subject to the AUC estimate at covariate level k. The regression model is given by $E(\theta_{ik}^*) = \beta_0^J + \beta_1^J X_k$. Since $E(\theta_{ik}^*) \in (0,1)$, we again consider using non-linear regression methods to estimate models of the form:

$$g\left(E(\theta_{ik}^*)\right) = \beta_0^J + \beta_1^J X_k,$$

where the function g is defined as before. Like the derived variables AUC regression method, a major limitation of this approach also is that continuous covariates are not allowed.

4.2.3 Analytical Comparisons

Theorem 3. When $n_{Dk} = n_D$ and $n_{\bar{D}k} = n_{\bar{D}}$ for all k, $\widehat{\theta}_k = \frac{1}{n_D n_{\bar{D}}} \sum_{ij} U_{ijk}$, 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 non-identity 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 §4.2. Next, we evaluate the small-sample performance of the proposed method under a model for continuous covariates. We generate data such that $Y_i^D \sim N(\mu_{D,X}, \sigma_D^2)$ and $Y_j^{\bar{D}} \sim N(\mu_{\bar{D},X}, \sigma_{\bar{D}}^2)$, 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:

$$\theta_X = \Phi\left(\frac{\mu_{D,X} - \mu_{\bar{D},X}}{\sqrt{\sigma_{\bar{D}}^2 + \sigma_D^2}}\right) = \Phi\left(\frac{\gamma_2 X}{\sqrt{\sigma_{\bar{D}}^2 + \sigma_D^2}}\right) = \Phi\left(\beta X\right), \tag{6}$$

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

5.1 Comparison with Existing AUC 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 $\mu_{D,X} = 0.5X, \sigma_D =$

$1.2, \mu_{\bar{D},X} = 0$, and $\sigma_{\bar{D}} = 1$ so that the model is $\Phi^{-1}(\theta_k) = 0.32X_k$. 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 are found in (Dodd, 2001). Our method produce 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 1000 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 $X \sim Uniform(0, 10)$. Parameter estimates are obtained from generating U_{ij}'s for all pairs of disease and non-disease test results. 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 30-200 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 anti-conservative for $n = 30$.

6 ASSESSMENT OF DEVICE FOR DIAGNOSING HEARING LOSS

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 continuous. 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, since 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 42% for every 10 dB increase in stimulus intensity (AUC odds = 0.58, 95% CI = 0.43,0.79) 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 better discriminates severely impaired ears from normal ears than mildly impaired ears from normal ears. Lastly, increasing the frequency

setting appears to increase the AUC odds 7% for every 100 Hz increase(AUC odds = 1.07, 95% CI = 0.99,1.16), but this result is not statistically significant.

Graphical methods, such as plots of fitted versus empirical AUCs, were used to evaluate model fit (Figure 1). Severity was categorized into four categories. Note the cloud of points in the upper right quadrant (Figure 1a). Plots of frequency for fixed severity and intensity suggested a lack of fit(not shown). Hence, the model was re-fit with frequency as dummy variables and the fit is somewhat better (Figure 1b). 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 is likely not linear, but more data are necessary for its characterization. Finally, note that although 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 has recently been proposed with different nomenclature by Fine and Bosch (2000) for use in toxicology and by Foulkes and De Gruttola (2002) for predicting HIV 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 the regression methods we have proposed here 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 or 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, as 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 modelling 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), where a regression model for the ROC curve is stipulated, and that stemming from work by Tosteson and Begg (1987) that models the probability distributions for the test results Y^D and $Y^{\bar{D}}$. The latter approach, modelling probability distributions, requires the strongest assumptions, while Pepe's approach, that models the *relationship* between those distributions as characterized by the ROC curve, requires fewer. Our approach requires fewer assumptions still because covariate effects on a summary index need only be specified. We will investigate if this leads to robustness for our approach over others in future work. We refer to 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. Attractions 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 method lead 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.

8 APPENDIX: PROOFS

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

8.1 Proof: Lemma 1

We show that under (C1)-(C6), if the sum $(n_D n_{\bar{D}})^{-1}\partial S_N(\beta)/\partial\beta \xrightarrow{p} ES_{ij}(\beta)$ as $N \to \infty$, then convergence of $(n_D n_{\bar{D}})^{-1}\partial S_N(\beta)/\partial\beta$ to its expectation is uniform for $\beta \in N_\delta(\beta_0)$.

We find a finite union of intervals with a 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:

$$
\begin{aligned}
&\sup_{\beta\in N_\delta(\beta_0)}\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta)}{\partial\beta}-E\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta)}{\partial\beta}\right)\right| \\
&= \max_k \sup_{\beta\in C_k}\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta)}{\partial\beta}-\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}+E\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}\right)\right. \\
&\quad \left.-E\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta)}{\partial\beta}\right)+\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}-E\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}\right)\right| \\
&\leq \max_k \sup_{\beta\in C_k}\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta)}{\partial\beta}-\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}\right| + \\
&\quad \max_k \sup_{\beta\in C_k}\left|E\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}\right)-E\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta)}{\partial\beta}\right)\right| \\
&\quad +\max_k \sup_{\beta\in C_k}\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}-E\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}\right)\right| \\
&= A_{1,N}+A_{2,N}+A_{3,N} \qquad (7)
\end{aligned}
$$

The Mean Value Theorem gives the following result for the first term in (7).

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

since 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 $\left(\frac{\partial}{\partial\beta}E\left(\partial S_N(\beta^\star)/\partial\beta\right)\right)$ similarly imply $A_{2,N} < \psi M_2$ where $M_2 < \infty$. Finally, since $(n_D n_{\bar{D}})^{-1}\partial S_N(\beta_k)/\partial\beta$ converges in probability to its expectation, for a given k, we can find an N_ϵ such that when $N > N_\epsilon$ then

$$
P\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}-E\left(\frac{1}{n_D n_{\bar{D}}}\frac{\partial S_N(\beta_k)}{\partial\beta}\right)>\epsilon/2\right)<\gamma/K.
$$

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

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

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

8.2 Proof: Lemma 2

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

$$
\begin{aligned}
P\left\{\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial}{\partial\beta}S_N(\beta) - E\frac{\partial S_{ij}(\beta)}{\partial\beta}\right| > \epsilon\right\} &= P\left\{\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial}{\partial\beta}S_N(\beta) - \frac{1}{n_D}\sum_i E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)|Y_i^D\right)\right.\right. \\
&\qquad \left.\left. + \frac{1}{n_D}\sum_i E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)|Y_i^D\right) - E\frac{\partial S_{ij}(\beta)}{\partial\beta}\right| > \epsilon\right\} \\
&\le P\left\{\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial}{\partial\beta}S_N(\beta) - \frac{1}{n_D}\sum_i E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)|Y_i^D\right)\right| > \epsilon/2\right\} \\
&\quad + P\left\{\left|\frac{1}{n_D}\sum_i E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)|Y_i^D\right) - E\frac{\partial S_{ij}(\beta)}{\partial\beta}\right| > \epsilon/2\right\} \qquad (8)
\end{aligned}
$$

Consider the first term on the right-hand side (RHS) of the inequality in (8):

$$
\begin{aligned}
E\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial}{\partial\beta}S_N(\beta) - \frac{1}{n_D}\sum_i E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)|Y_i^D\right)\right| &= E\left|\frac{1}{n_D}\sum_i\left(\frac{1}{n_{\bar{D}}}\sum_j \frac{\partial}{\partial\beta}S_{ij} - E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)|Y_i^D\right)\right)\right| \\
&\le \frac{1}{n_D}\sum_i E\left|\left(\frac{1}{n_{\bar{D}}}\sum_j \frac{\partial}{\partial\beta}S_{ij}(\beta) - E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)|Y_i^D\right)\right)\right| \qquad (9)
\end{aligned}
$$

The terms inside the expectation in (9) are i.i.d. across j for fixed i. Hence, by the weak law of large numbers (WLLN), (9) $\to_p 0$. Since convergence in mean implies convergence in probability, it

follows that for all $\epsilon > 0$, $P\left(\left|\frac{1}{n_D n_{\bar{D}}}\frac{\partial}{\partial\beta}S_N(\beta) - \frac{1}{n_D}\sum_i E\left(\frac{\partial}{\partial\beta}S_{ij}|Y_i^D\right)\right| > \epsilon/2\right) \to 0$ as $N \to \infty$.

Now consider the second term on the RHS in (8). The terms $E(\partial S_{ij}(\beta)/\partial\beta|Y_i^D)$ are independent and have finite expectation. From the WLLN $1/n_D \sum_i E(\partial S_{ij}(\beta)/\partial\beta|Y_i^D) \xrightarrow{p} E\left(\partial S_{ij}(\beta)/\partial\beta\right)$.

Therefore, for $\epsilon > 0$, $P\left(\left|\frac{1}{n_D}\sum_i E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)|Y_i^D\right) - E\left(\frac{\partial}{\partial\beta}S_{ij}(\beta)\right)\right| > \epsilon/2\right) \to 0$ as $N \to \infty$. Hence the two terms in (8) $\xrightarrow{p} 0$, and the result follows.

8.3 Proof: Theorem 3

We show that the least-squares estimates 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 non-diseased 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 $U_{ijk} \equiv I(Y_{ik}^D > Y_{jk}^{\bar{D}}), \bar{U} \equiv \frac{1}{Kn^2}\sum_{ijk} U_{ijk}, \bar{X} \equiv \frac{1}{2nK}\sum_{ijk} X_{ijk} = \frac{1}{K}\sum_k X_k, S(U,X) \equiv \sum_{ijk} U_{ijk}X_k - Kn^2\bar{U}\bar{X}$, and $S(X,X)^P \equiv \sum_{ijk} X_{ijk}^2 - Kn^2\bar{X} = n^2\left\{\sum_k X_k^2 - K\bar{X}\right\}$ The least-squares estimators are given by:

$$\hat{\beta}_0^P = \bar{U} - \hat{\beta}_1^P\bar{X} \text{ and } \hat{\beta}_1^P = \frac{S(U,X)}{S(X,X)^P}.$$

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 $\overline{\widehat{AUC}_k} \equiv \frac{1}{Kn^2}\sum_{ijk} U_{ijk} = \bar{U}$ and $S(X,X)^d \equiv \sum_k X_k^2 - K\bar{X}^2 = \frac{1}{n^2}S(X,X)^P$. A little algebra shows that $S(\widehat{AUC},X) \equiv \sum_k(\widehat{AUC}_k X_k) - K\overline{\widehat{AUC}}\bar{X} = \frac{1}{n^2}S(U,X)$. 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 (JA) at the k^{th} covariate level for $l = 1, ..., 2n$. We use A_{lk} to denote an JA from the combined vector, $(A_{1k}^{\bar{D}}, ..., A_{n_{\bar{D}}k}^{\bar{D}}, A_{(n_{\bar{D}}+1)k}^{D}, ..., A_{(n_{\bar{D}}+n_D)k}^{D})$. We also denote the vector as $\{A_{ik}^{\bar{D}} : i = 1, ..., n_D, A_{jk}^{D} : j = 1, ..., n_{\bar{D}}\}$. Note that the superscript in $A_{ik}^{\bar{D}}$ indicates that this term is averaged

across all non-diseased observations and random with respect a given observation from diseased.

The least-squares estimators from the jackknife-AUC model depend on the random variables $\{A_{lk} : l = 1, ..., 2n\}$. Observe that $\bar{X}^J \equiv \frac{1}{K2n}\sum_{k=1}^{K}\sum_{l=1}^{2n} X_{lk} = \bar{X}$ and $S(X,X)^J \equiv \sum_{k=1}^{K}\sum_{l=1}^{2n} X_{lk}^2 - 2nK\bar{X} = 2n\left\{\sum_k X_k^2 - K\bar{X}\right\} = \frac{2}{n}S(X,X)^P$. Next, we show that $\bar{A} \equiv \frac{1}{K2n}\sum_{k=1}^{K}\sum_{l=1}^{2n} A_{lk}$ equals $\bar{U}$. The mean of the jackknifed AUC at covariate level k can be written as $\bar{A}_k = \frac{1}{2n}\sum_l A_{lk} = \frac{1}{2n}\sum_{i=1}^{n} A_{ik}^{\bar{D}} + \frac{1}{2n}\sum_{j=1}^{n} A_{jk}^{D}$. Define $\hat{F}_k^{\bar{D}}(Y_{ik}^D) = \frac{1}{n}\sum_j I(Y_{ik}^D > Y_{jk}^{\bar{D}})$ and $\hat{F}_k^{D}(Y_{jk}^{\bar{D}}) = \frac{1}{n}\sum_i I(Y_{ik}^D > Y_{jk}^{\bar{D}})$, where $\hat{F}$ 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):

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

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

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

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

$$S(A,X) = \equiv \sum_{k=1}^{K}\sum_{l=1}^{2n} A_{lk}X_k - 2nK\bar{U}\bar{X}\sum_{k=1}^{K} = \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n} A_{ik}^{\bar{D}}X_k}_{(c)} + \underbrace{\sum_{k=1}^{K}\sum_{j=1}^{n} A_{jk}^{D}X_k}_{(d)} - 2Kn\bar{A}\bar{X} \qquad (10)$$

The term (c) in the expression in (10) is equal to:

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{n} A_{ik}^{\bar{D}} X_k &= \sum_{k=1}^{K}\sum_{i=1}^{n}\left(\frac{2n-1}{n-1}\hat{F}_k^{\bar{D}}(Y_{ik}^{D}) - \frac{n}{n-1}\widehat{AUC}_k\right) X_k \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{2n-1}{n-1}\left(\frac{1}{n}\sum_{j=1}^{n} U_{ijk}X_k\right) - \frac{n^2}{n-1}\sum_{k}\widehat{AUC}_k X_k \\ &= \frac{2n-1}{n(n-1)}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ijk}X_k - \frac{n^2}{n-1}\sum_{k}\widehat{AUC}_k X_k \qquad (11)\end{aligned}$$

In a similar manner, one can show that (d) in equation (10) equals the expression shown in (11), and expression (10) equals:

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

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 $\hat{\beta}_0^J = \bar{A} - \hat{\beta}_1^J \bar{X}^J = \hat{\beta}_0^P$.

(a) (b)

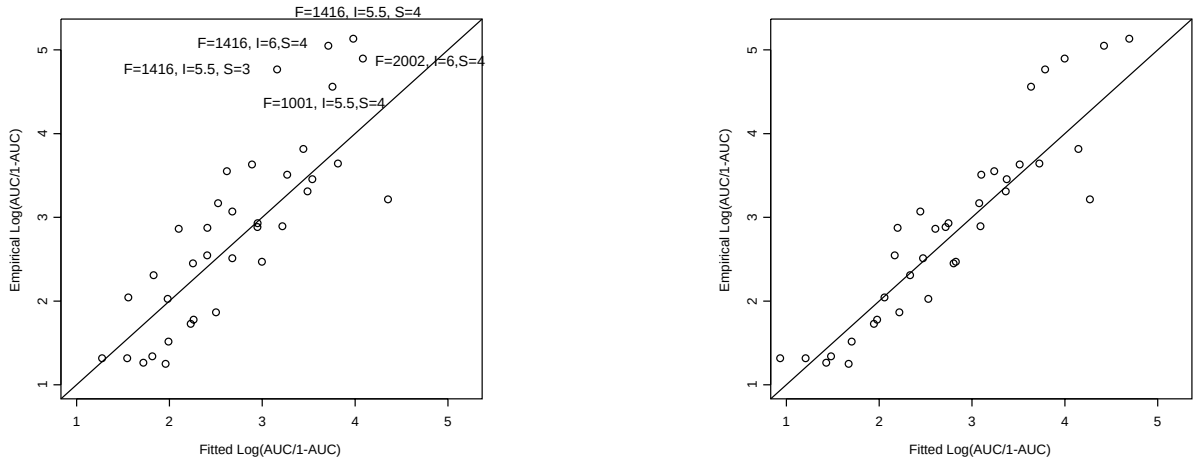

Figure 1: (a) empirical versus fitted AUCs on the log-odds scale with frequency as continuous. F=frequency, I=intensity, S=severity category. (b) empirical versus fitted log AUC-odds with frequency as dummy variables.

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
$\hat{\beta}_1$	0.326	0.338	0.341	0.329	0.332	0.360
% Bias	2.0	5.5	6.6	2.7	3.7	12.6
Relative Efficiency	1	0.88	0.43	1	0.76	0.14

NOTES: True β_1= 0.320, The balanced design sampled equal numbers at each covariate level. The unbalanced design sampled 50%, 10%,10%, 10%, 20% within covariate levels $X = 1, 2, 3, 4, 5$, respectively. Results represent 1000 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	0.442	6.3%	0.166	0.180	-8.0%	0.930
50	0.433	4.1%	0.140	0.133	5.2%	0.950
100	0.427	2.5%	0.090	0.086	5.1%	0.955
200	0.417	0.2%	0.062	0.060	3.0%	0.953

NOTES: Confidence intervals computed assuming normality with bootstrapped standard error estimates. SE = Standard Error. CI = Confidence Interval. Results represent 1000 realizations of the model and 200 bootstrap samples each.

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(2):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 University Press.

Dodd, L. (2001). *Regression Methods for Areas and Partial Areas Under the Receiver-Operating Characteristic Curve*. PhD thesis, University of Washington, Seattle, WA 98195.

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*, 72:147–48.

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*. Chapman and Hall, second edition.

Pepe, M. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84:595–608.

Pepe, M. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, 54:124–135.

Pepe, M. (2000). Receiver operating characteristic methodology. *Journal 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 emision measurements. *J. Acoust. Soc. Am.*, 100:956–967.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine*, 8:1277–1290.

Zhu, H., Beling, P., and Overstreet, G. (2002). A Bayesian framework for the combination of classifier outputs. *Journal of the Operational Research Society*.