3-17-2006

Adjusting for Covariate Effects on Classification Accuracy Using the Covariate-Adjusted ROC Curve

Holly Janes
Johns Hopkins University, hjanes@jhsph.edu

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

1 Introduction

Research into new markers for disease diagnosis, screening, and prognosis has exploded in recent years. In each of these settings, the primary question is of classification accuracy: How well does the marker distinguish between the two groups of individuals, the "cases" and the "controls"?

The ROC curve plays a central role in evaluating classification accuracy (Baker, 2003; Pepe et al., 2001). It displays the tradeoff between false-positive and false negative error rates associated with classification rules based on the marker, Y. Let D denote the binary group variable, "disease status", and Y_D and $Y_{\bar{D}}$ case and control observations with survivor functions $S_D(y) = P[Y_D > y]$ and $S_{\bar{D}} = P[Y_{\bar{D}} > y]$. The ROC curve is a plot of the true-positive fraction (TPF) (sensitivity) versus the false-positive fraction (FPF) (1 - specificity) for the rules which classify an individual as "test-positive" if $Y > c$, where the threshold c varies over all possible values. Equivalently, at a FPF $= t$, $\text{ROC}(t) = P[Y_D > S_{\bar{D}}^{-1}(t)] = S_D(S_{\bar{D}}^{-1}(t))$ (Pepe, 2003).

There are commonly factors which affect test accuracy. Understanding these effects helps to determine how the test should be used in practice. It may be that the definition of testing positive on the basis of the marker should depend on covariates, or it may be that the accuracy of the test is less than optimal in certain settings (Pepe, 2003 [p. 48-49]). Patient characteristics, such as age, gender, or race, often impact marker measurements. For example, younger women have more dense breasts, which leads to more false positive errors when using a mammogram. Factors which affect the test itself, such as the expertise of the test operator, or variations in how the test is performed, may also affect test accuracy. The manner in which a biological

specimen is collected, processed, or stored can greatly affect the assay. Storage time may be an important factor. Characteristics of disease also commonly affect accuracy. More advanced disease is often easier to detect in cases, and controls may have related conditions that increase the likelihood of false positive errors.

While the concept of covariate adjustment has been well studied in epidemiological and clinical research, as well as in statistics more broadly, it has not been developed in the classification context. In this paper, we propose a covariate-adjusted summary measure of classification accuracy. We begin by motivating covariate adjustment in the classification setting. In Section 3, we define and give several interpretations for the covariate-adjusted ROC curve. Section 4 proposes and provides distribution theory for two novel estimators. Their small-sample performance is evaluated in Section 5. In section 6, we illustrate these methods using data from the Physicians' Health Study.

2 Background and Motivation

2.1 Covariate Effects on Classification Accuracy

There are two dimensions to the ROC curve, and hence two ways in which a covariate can impact test accuracy (Pepe, 2003 [pp. 131 – 132]). Consider the traditional "pooled" ROC curve, which ignores covariates by combining all case observations together, regardless of covariate value, and all control observations together, regardless of covariate value. The pooled ROC describes the accuracy of rules that use a common threshold to define test-positive. That is, the same threshold is used for all

marker observations, regardless of their covariate values. If a covariate is associated with marker observations among controls, then the use of a common threshold will yield varying FPF's across covariate groups. Hence, varying the covariate value has the effect of moving horizontally along the ROC curve. This is illustrated in Figure 1, adapted from Janes and Pepe (unpublished manuscript), which shows data for a hypothetical marker, Y, and binary covariate, Z. The two points on the common covariate-specific ROC curve are the operating characteristics of the positivity criterion '$Y > 2.5$' in the $Z = 0$ and $Z = 1$ populations. But the covariate may also affect the inherent discriminatory accuracy of the marker, i.e., the separation between the Y_D and $Y_{\bar{D}}$ distributions (the ROC curve) may vary with covariate value. This is analogous to effect modification in the association setting. We initially focus on covariates with only the first type of covariate effect. That is, we assume that the separation between the Y_D and $Y_{\bar{D}}$ distributions is the same in different covariate populations, as in Figure 1(a). For example, in a multi-center study, variations in equipment or testing procedures may affect marker levels equally in cases and controls so that marker performance is similar across the study sites. More generally, any covariate that causes a monotone transformation of Y that is independent of disease status will not affect ROC performance.

2.2 What is Covariate Adjustment for ROC Curves?

In therapeutic research, the covariate-adjusted treatment effect is the effect of treatment within a population with fixed covariate value. Similarly, in classic etiologic epidemiology, the covariate-adjusted odds ratio is the odds associated with an expo-

sure (or risk factor) among subjects with the same covariate values. In the absence of effect modification by covariates, the covariate-adjusted effect of treatment or exposure is defined to be the effect that is common across covariate strata. Conceptually, we stratify. In practice, covariate adjustment may be achieved by stratification, when covariates are discrete, or using regression methods.

We define covariate adjustment for ROC curves using an analogous approach. The covariate-adjusted ROC curve for Y is the covariate-stratified ROC curve. In other words, it is the ROC curve which characterizes the separation between Y_D and $Y_{\bar{D}}$ distributions in a population with fixed covariate value. In Figure 1, this is the common covariate-specific ROC curve (solid line).

We emphasize that covariate adjustment is different from other roles for covariates in marker evaluation, such as: 1) the performance of the covariate-adjusted risk score for Y; 2) the incremental value of Y over Z; 3) the performance of Y in a study where controls are matched to cases with respect to Z; and 4) ROC regression which allows the performance of Y to vary with Z.

Consider first the covariate-adjusted risk score for Y, $P[D = 1|Y, Z]$. For example, a logistic regression model for Y with adjustment for Z yields a linear predictor $\beta_1 Y + \beta_2 Z$. The ROC curve for the linear predictor is not the ROC curve for Y adjusted for Z, but rather it captures the ability of the combination of marker and covariates to discriminate between cases and controls. Observe that this combination may perform well even if Y is a poor classifier if Z discriminates well. Figure 2 shows two examples where (Y, Z) is bivariate normal with mean $(0, 0)$ and variance-covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ in controls, and mean (μ_Y, μ_Z) and the same variance-

covariance in cases. Under this model, the risk score, the optimal combination of Y and Z for discrimination (McIntosh and Pepe, 2002), is (a monotone function of) a linear combination of Y and Z. In Figure 2(a), Z is a good classifier ($\mu_Z = 1.5$) but Y is not ($\mu_Y = 0.5$), and the two are relatively uncorrelated ($\rho = 0.1$). The linear predictor performs well, but the covariate-adjusted ROC curve for Y, i.e. the ROC curve for Y stratified by Z, is low because it relates to the discriminatory accuracy of Y. In Figure 2(b), both Y and Z are good classifiers ($\mu_Y = \mu_Z = 1.5$), but are highly correlated ($\rho = 0.9$). The linear predictor performs well, as expected since it should be at least as good as either marker on its own. However, after adjustment for Z the ROC curve for Y is low because within a population where Z is fixed, Y is not a good discriminator. Most of its marginal discrimination is explained by Z, with which it is highly correlated.

Consider the incremental value of the marker over the covariates. This is quantified by comparing the ROC curve for the (Y, Z) combination to the ROC curve for Z alone. This answers yet another question: How much does discriminatory accuracy improve with the addition of Y to Z? It is easy to find examples in which the incremental value of Y is low, but the covariate-adjusted performance of Y is good, and where the incremental value is large, but the covariate-adjusted performance of Y is poor (Janes and Pepe, unpublished manuscript).

Matching of controls to cases is a design strategy commonly used to account for covariate effects on classification accuracy. But the performance of Y in a study matched on Z does not reflect its covariate-adjusted performance either. It is widely appreciated in epidemiologic research that the analysis in a matched study must adjust for the matching covariates in order to appropriately estimate exposure or risk

factor effects. Analyses that do not adjust for matching covariates produce biased estimates. A similar result was recently shown to hold for evaluation of classification accuracy from matched studies (Janes and Pepe, unpublished manuscript): the unadjusted ROC curve is biased downwards. Matching does not in and of itself adjust for covariates. Rather, the analysis must also make these adjustments.

Finally, we note that ROC regression (Tosteson and Begg, 1988; Toledano and Gatsonis, 1995; Pepe, 1998; Faraggi, 2003; Schisterman et al., 2004; Le, 1997; Pepe, 2000; Alonzo and Pepe, 2002; Cai and Pepe, 2002) is a methodology that investigates if and how the discriminatory accuracy of the marker (the ROC curve) depends on covariates. This is analogous to effect modification in epidemiologic research and is not the same as covariate adjustment. Figure 1 demonstrates that covariate adjustment may be necessary even when the ROC curve does not vary with covariates.

2.3 Why Adjust for Covariates?

The pooled or unadjusted ROC curve has a number of drawbacks when there are covariate effects on test accuracy. Observe that the pooled ROC describes the performance of the marker, including the portion of performance that is due to the covariates. This is illustrated in Figure 1, scenario 1, wherein the pooled ROC lies above the common covariate-specific ROC curve. For a real data example, consider prostate-specific antigen (PSA), prostate cancer screening biomarker. The pooled ROC curve for PSA is overly optimistic because it includes the portion of discriminatory accuracy due to age: cases tend to be older than controls, and older subjects tend to have higher PSA levels (Oesterling et al. 1993; Baillargeon et al., 2005). The

performance of PSA conditional on age is of much more interest.

The use of a common threshold to define test-positive is another undesirable attribute of the pooled ROC. For example, in the PSA setting, the use of a common threshold will yield much higher FPF's in older populations than in younger populations. This suggests that age-specific thresholds should be used to control the FPF across age groups, as has been suggested in the literature (Oesterling et al. 1993).

In certain settings, failing to adjust for covariates will attenuate the ROC curve. In particular, if the covariate affects the marker in the same way in cases and controls and is independent of disease status, the pooled ROC curve will lie below the common, covariate-specific ROC curve (Pepe, 2003 [p. 133–134]). This is illustrated in Figure 1, scenario 2. In the radiology literature, attenuation of the ROC curve associated with pooling data from multiple readers who use the rating scales differently is well known (Swets and Pickett, 1982 [p. 65]; Hanley, 1989; Rutter and Gatsonis, 2001). Recently this phenomenon has been highlighted as a general issue in matched case-control studies (Janes and Pepe, unpublished manuscript).

3 The $\mathcal{A}$ROC

Consider a continuous marker, Y, and continuous covariate, Z. Let Z_D and $Z_{\bar{D}}$ denote case and control covariate observations with cumulative distribution functions (CDF) P_{Z_D} and $P_{Z_{\bar{D}}}$. Denote by $S_{DZ}(y) = P[Y_D > y|Z]$ and $S_{\bar{D}Z} = P[Y_{\bar{D}} > y|Z]$ the continuous survivor functions for Y conditional on Z, f_{DZ} and $f_{\bar{D}Z}$ the corresponding densities, and $\mathrm{ROC}(t) = S_{DZ}(S^{-1}_{\bar{D}Z}(t))$ the common covariate-specific ROC curve. Our methods generalize naturally to a discrete covariate or multiple covariates.

3.1 Definition and Interpretations

The covariate-adjusted ROC curve is defined as the common covariate-specific ROC curve for Y, and denoted by $\mathcal{A}\text{ROC} = S_{DZ}(S^{-1}_{\bar{D}Z}(t))$ to emphasize its adjusted or stratified nature. Mathematically,

$$\mathcal{A}\text{ROC}(t) = P[Y_D > S^{-1}_{\bar{D}Z_D}(t)], \tag{1}$$

where the covariate-specific thresholds, $S^{-1}_{\bar{D}Z}(t)$, yield FPF $= t$ among controls with covariate value Z. In other words, the $\mathcal{A}$ROC is a plot of the TPF versus the FPF for the set of rules that classify a subject with covariate value Z as positive if $Y > c_Z$, where $c_Z = S^{-1}_{\bar{D}Z}(t)$ is the Z-specific threshold associated with a FPF of t. Using these rules, the marginal FPF is also equal to t.

Several other interpretations can be provided for the covariate-adjusted ROC curve. We write the $\mathcal{A}$ROC as

$$\mathcal{A}\text{ROC}(t) = P[S_{\bar{D}Z_D}(Y_D) \leq t]. \tag{2}$$

This reveals that the $\mathcal{A}$ROC is the CDF of $S_{\bar{D}Z_D}(Y_D)$, the placement of a case observation relative to a reference distribution of controls with the same covariate value as the case. Contrast this with the unadjusted or pooled ROC curve, $\text{ROC}(t) = P[S_{\bar{D}}(Y_D) \leq t]$, which is the CDF of a case observation standardized relative to the general control distribution (Pepe and Cai, 2002).

Another interpretation for the $\mathcal{A}$ROC follows from marker standardization. Let $Y^* = 1 - S_{\bar{D}Z}(Y)$ be the percentile for Y in the control population with the appropriate covariate value. Such standardization is used, for example, to standardize children's weights relative to height and gender (Hammill et al., 1977). The $\mathcal{A}$ROC is simply the pooled ROC curve for Y^* (this follows because $1 - S_{\bar{D}Z_{\bar{D}}}(Y_{\bar{D}}) \sim \text{Uniform}[0, 1]$).

The $\mathcal{A}$ROC has some attractive mathematical properties. It is invariant with respect to monotone increasing transformations of Y and/or Z. It is also unaffected by control covariate-dependent sampling (e.g., matching). This follows because such a design samples controls randomly conditional on Z, and cases are a simple random sample from the case population.

Exploring the ordering of the pooled and adjusted ROC curves is useful for identifying scenarios in which failing to adjust for covariates leads to bias, and for determining the direction and magnitude of the bias. The mathematical relationship between the two ROC curves is complex. In one trivial case, they are the same: if the distribution of $Y_{\bar{D}}$ is independent of Z, the Z-specific thresholds associated with a fixed FPF do not vary. If, on the other hand, Z is independent of D and does not affect the discriminatory capacity of Y, the pooled ROC curve will lie below the $\mathcal{A}$ROC (Pepe, 2003 [p. 135]). More generally, the ordering of the two ROC curves depends on the distributions of Y and Z and the associations between them and of each with disease status. In a classical distributional case (the binormal model), the ordering is somewhat intuitive (Janes and Pepe, 2006). The pooled ROC will lie above the $\mathcal{A}$ROC if the association between Z and D is stronger than the association between Y and D, since it includes the portion of discriminatory accuracy due to Z. However, the $\mathcal{A}$ROC rule will yield gains in accuracy if the association between Y and D is stronger than the association between Z and D, and if in addition the correlation between $Y_{\bar{D}}$ and $Z_{\bar{D}}$ is large. Larger correlation between Y_D and Z_D will further improve the $\mathcal{A}$ROC relative to the pooled ROC.

3.2 When Covariates Affect Discrimination

When Z affects discrimination, covariate-specific ROC curves, $\text{ROC}^Z(t) = S_{DZ}(S^{-1}_{\bar{D}Z}(t))$, are of interest. A wide variety of methods for estimating covariate-specific ROC curves have been proposed (see, e.g., Tosteson and Begg, 1988; Toledano and Gatsonis, 1995; Pepe, 1998; Faraggi, 2003; Schisterman et al., 2004; Le, 1997; Pepe, 2000; Alonzo and Pepe, 2002; Cai and Pepe, 2002). These methods allow for covariate effects on both the FPF's (or thresholds) and on the ROC curve itself.

Interestingly, the $\mathcal{A}$ROC is a simple summary of covariate-specific ROC curves:

$$\begin{aligned}\mathcal{A}\text{ROC}(t) &= \int P[Y_D > S^{-1}_{\bar{D}Z_D}(t) \mid Z_D = Z]\, dP_{Z_D}(Z) \\ &= \int \text{ROC}^Z(t)\, dP_{Z_D}(Z). \qquad (3)\end{aligned}$$

Equivalently, $\mathcal{A}\text{ROC}(t) = E_{Z_D}[\text{ROC}^{Z_D}(t)]$. The $\mathcal{A}$ROC reports a weighted average of covariate-specific TPF's, holding the covariate-specific FPF's constant. This is a useful summary of covariate-adjusted accuracy, particularly for small studies where covariate-specific ROC curves cannot be estimated with precision. It also provides a single summary of covariate-adjusted accuracy with which to compare markers.

4 Estimation of the $\mathcal{A}$ROC

4.1 Estimators

We propose two estimators for $\mathcal{A}\text{ROC}(t) = P[Y_D > S^{-1}_{\bar{D}Z_D}(t)]$ using n_D and $n_{\bar{D}}$ case and control observations, where n_{DZ} and $n_{\bar{D}Z}$ are the numbers of each with covariate value Z. In both instances, we estimate the outside probability empirically. The remaining task is estimation of the control quantiles, $S^{-1}_{\bar{D}Z}(t)$. With the non-parametric

estimator, valid for a discrete covariate ($Z = 1, \ldots, K$), we use empirical quantiles in each stratum. With the semi-parametric estimator, the quantiles are estimated based on a model for the distribution of $Y_{\bar{D}}$ as a function of $Z_{\bar{D}}$. Here we lay out the general framework for the $\mathcal{A}$ROC estimator, of which the non-parametric estimator is a special case.

Suppose we assume the quantile model, $Y_{\bar{D}} = f(Z_{\bar{D}}, \epsilon; \theta)$, where ϵ is random error and θ are parameters. With the semi-parametric $\mathcal{A}$ROC estimator, this model may be parametric, such as a normal linear model, or semi-parametric (see, e.g., Heagerty and Pepe, 1999). The model induces a form for the control quantiles. Let $q_Z(t;\theta) = S^{-1}_{\bar{D}Z}(t;\theta)$ be the function which extracts the $1-t$ quantile from the set of control quantiles with covariate value Z, where $q_Z(t;\hat{\theta}) = S^{-1}_{\bar{D}Z}(t;\hat{\theta})$ is the estimated quantile. We write

$$\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t) = \frac{1}{n_D}\sum_{i=1}^{n_D} \mathrm{I}\left[Y_{Di} > q_{Z_{D_i}}(t;\hat{\theta})\right].$$

With the non-parametric estimator, $\theta = \left(S^{-1}_{\bar{D}Z=1}(t), \ldots, S^{-1}_{\bar{D}Z=K}(t)\right)^T$ are the quantiles themselves, $\hat{S}_{\bar{D}Z}(y) = n^{-1}_{\bar{D}Z}\sum_{i=1}^{n_{\bar{D}Z}} \mathrm{I}\left[Y_{\bar{D}Z_i} > y\right]$, and $q_Z(t;\hat{\theta}) = \hat{S}^{-1}_{\bar{D}Z}(t) = \inf_{s\in[0,1]}\left\{\hat{S}_{\bar{D}Z}(s) \geq t\right\}$. This estimator depends only on the ranks of the data, and thus is invariant with respect to monotone transformations.

4.2 Asymptotic Distribution Theory

We assume the following conditions in establishing asymptotic distribution theory. Recall that the distribution of Y_D is not a function of θ.

C(1) Random sampling conditional on D, $n_D + n_{\bar{D}} \to \infty$, and $\frac{n_D}{n_{\bar{D}}} \to \lambda \in (0,1)$.

C(2) $\sqrt{n_{\bar{D}}}\,(\hat{\theta} - \theta) \xrightarrow{d} N(\theta, \Sigma_\theta)$ as $n_{\bar{D}} \to \infty$.

C(3) $\mathcal{A}\text{ROC}_\theta(t)$ is differentiable, and hence continuous, in θ.

C(4) $\lim_{n_{\bar{D}}\to\infty} P[\mathcal{A}\text{ROC}_{\hat{\theta}}(t) \notin \{0,1\}] = 1$, where $\mathcal{A}\text{ROC}_{\hat{\theta}}(t) = P[Y_D > q_{Z_D}(t;\hat{\theta}) \mid \hat{\theta}]$ is the $\mathcal{A}$ROC based on estimated quantiles.

C(5) $t \notin \{0,1\}$.

We note in relation to **C(1)** that covariate-dependent sampling can also be accommodated (see Section 4.4). A wide variety of quantile models satisfy **C(2)**, including parametric (Cole, 1990; Cole and Green, 1992; Pepe, 2003 [p. 140]), semi-parametric (Heagerty and Pepe 1999; Zheng 2002), empirical (proven in appendix A.2), and any $\hat{\theta}$ based on unbiased estimating equations satisfying standard regularity conditions. **C(3)** is also valid for a diversity of quantile and ROC models, such as the location-scale quantile model (Heagerty and Pepe, 1999) with bounded $\frac{\partial}{\partial t}\text{ROC}^Z(t)$ and $E(Z_D) < \infty$ (Janes and Pepe, 2006). **C(4)** is violated if the support of the case distribution is entirely above or below the estimated quantile of interest. This will not occur as long as the support of the Y_D distribution includes the support of the $Y_{\bar{D}}$ distribution, or if the support of the Y_D distribution is unbounded (e.g.,the normal distribution). We also require that $t \notin \{0,1\}$, but by definition $\mathcal{A}\text{ROC}(0) = 0$ and $\mathcal{A}\text{ROC}(1) = 1$. Finally, imposing continuity of $S_{DZ}(y)$ and $S_{\bar{D}Z}(y)$ implies that $\text{ROC}^Z(t) = S_{DZ}(S^{-1}_{\bar{D}Z}(t))$ and $\mathcal{A}\text{ROC}(t) = E_{Z_D}[\text{ROC}^{Z_D}(t)]$ are continuous in t.

Theorem 1 Under **C(1)**-**C(5)**, $\sqrt{n_D}\,(\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t) - \mathcal{A}\text{ROC}_\theta(t)) \xrightarrow{d} N(0, V(t))$ as $n_D, n_{\bar{D}} \to \infty$, where

$$V(t) = \mathcal{A}\text{ROC}_\theta(t)\,(1 - \mathcal{A}\text{ROC}_\theta(t)) + \lambda \cdot \frac{\partial}{\partial\theta}\mathcal{A}\text{ROC}_\theta(t)\ \Sigma_\theta\ \frac{\partial}{\partial\theta}\mathcal{A}\text{ROC}_\theta(t)^T \qquad (4)$$

(proven in appendix A.1).

The form of $V(t)$ is quite intuitive. The second component comes from estimating the Z-specific quantiles, while the first is a binomial variance associated with estimating the TPF, given the quantiles.

For the non-parametric $\mathcal{A}$ROC estimator, **C(2)** and **C(3)** are satisfied when

C(6) $f_{\bar{D}Z}(y)$ is continuous and positive in a neighborhood of $S^{-1}_{\bar{D}Z}(t)\ \forall Z$,

and $V(t)$ reduces to

$$V(t) = \mathcal{A}\text{ROC}_\theta(t)\left(1 - \mathcal{A}\text{ROC}_\theta(t)\right) + \lambda \cdot \sum_{Z=1}^{K} \frac{p^2_{Z_D}(Z)}{p_{Z_{\bar{D}}}(Z)} \cdot \frac{f_{DZ}(S^{-1}_{\bar{D}Z}(t))^2}{f_{\bar{D}Z}(S^{-1}_{\bar{D}Z}(t))^2} \cdot t(1-t), \quad (5)$$

where $p_{Z_D}(Z)$ and $p_{Z_{\bar{D}}}(Z)$ are the probability mass functions for Z_D and $Z_{\bar{D}}$ (proven in appendix A.2).

4.3 Consistent Variance Estimation

We propose two variance estimators. The first can be used to estimate the variance of the semi-parametric $\mathcal{A}$ROC estimator, (4). The semi-parametric estimator is consistent by Theorem 1. We assume that a consistent estimator of Σ_θ exists (e.g., if $\hat{\theta}$ is based on a set of unbiased estimating equations, a sandwich-type variance estimator can be used). The j^{th} component of $\frac{\partial}{\partial\theta}\mathcal{A}\text{ROC}_\theta(t)$ is estimated by $\dfrac{\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}+h^j(n)}(t) - \widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}-h^j(n)}(t)}{2h(n)}$, where $h(n)$ is $o(n_D^{-1/3})$, and $\hat{\theta}+h^j(n)$ $(\hat{\theta}-h^j(n))$ denotes the vector $\hat{\theta}$ with $h(n)$ added to (subtracted from) the j^{th} component only. The composite variance estimator is shown to be consistent in appendix A.3, under **C(1)**-**C(5)**, and

C(7) $\lim_{n_{\bar{D}}\to\infty} P[\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}+h^j(n)}(t) \notin \{0,1\}] = \lim_{n_{\bar{D}}\to\infty} P[\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}-h^j(n)}(t) \notin \{0,1\}] = 1,\ \forall j.$

With small sample sizes, the $\frac{\partial}{\partial\theta_j}\mathcal{A}\mathrm{ROC}_\theta(t)$ estimate may be sensitive to the choice of bandwidth, $h(n)$. We have used $h(n) = 0.04$ in applications and simulations; in one example this ensured $\dfrac{\mathcal{A}\mathrm{ROC}_{\theta+h^j(n)}(t) - \mathcal{A}\mathrm{ROC}_{\theta-h^j(n)}(t)}{2h(n)} \approx \dfrac{\partial}{\partial\theta_j}\mathcal{A}\mathrm{ROC}_\theta(t)$. This value has worked well. We leave exploration of the optimal choice of $h(n)$ for future research.

Our second variance estimator can be used to estimate the variance of the non-parametric $\mathcal{A}$ROC estimator, (5). Here, $\mathcal{A}\mathrm{ROC}_\theta(t)$ is estimated using the non-parametric estimator, $p_{Z_D}(Z)$ and $p_{Z_{\bar{D}}}(Z)$ using binomial proportions, $S^{-1}_{\bar{D}Z}(t)$ empirically, and $f_{DZ}(y)$ and $f_{\bar{D}Z}(y)$ with uniformly consistent kernel density estimates (Silverman 1986 [Section 3.7]). In appendix A.4, we prove that the composite function is consistent under **C(1)**-**C(6)**, and

C(8) $f_{DZ}(y)$ and $f_{\bar{D}Z}(y)$ are continuous density functions $\forall Z$.

Bootstrap variance estimation is a simple alternative which accommodates clustered sampling and performs well in practice and in small sample simulations (see Section 5).

4.4 Sampling Based on Covariates

In many situations, sampling may depend on both D and Z. Two simple examples are matching, in which controls are sampled to have the same Z distribution as the cases, and sampling subjects in a specified Z range, say conditional on $Z > z_0$. With such designs, our results continue to hold, but all population distributions should be replaced with sampling distributions in the asymptotic distributions of the estimators. For example, if sampling is conditional on D and $Z > z_0$, $S_D(y)$ and $S_{\bar{D}}(y)$ should be

replaced with $P[Y_D > y \mid Z > z_0]$ and $P[Y_{\bar{D}} > y \mid Z > z_0]$, respectively.

4.5 Estimation of the $\mathcal{A}$ROC using ROC-GLM

The $\mathcal{A}$ROC can also be estimated using ROC-GLM, a method originally proposed for estimating covariate-specific ROC curves (Pepe, 2000; Alonzo and Pepe, 2002). ROC-GLM requires estimating the covariate-specific control quantiles using any of the existing approaches, and specifying and fitting a model for the ROC curve, typically as a function of covariates:

$$g(\mathrm{ROC}^Z(t)) = g\left(P[S_{\bar{D}Z_D}(Y_D) \leq t \mid Z_D = Z]\right) = h_0(t) + \beta Z,$$

where g and h_0 are monotone functions on (0,1). A model for the $\mathcal{A}$ROC is obtained by including Z in the quantile calculations, while omitting Z from the ROC model, $g(\mathcal{A}\mathrm{ROC}(t)) = h_0(t)$. An example is the binormal model, $\mathcal{A}\mathrm{ROC}(t) = \Phi(\alpha + \beta\Phi^{-1}(t))$, where Φ is the standard normal CDF. This approach assumes a parametric form for the $\mathcal{A}$ROC, but the marker distributions remain unspecified. A smooth estimate of the $\mathcal{A}$ROC results. The version of ROC-GLM in which $h_0(t)$ is estimated empirically (Cai and Pepe, 2002) reduces to our semi-parametric estimator of the $\mathcal{A}$ROC.

5 Small Sample Performance of Proposed Estimators

In this section, we evaluate the finite sample properties of the $\mathcal{A}$ROC estimators using simulations. We first evaluate the non-parametric estimator and its variance, which can be used for discrete Z. We assume Y is normally distributed conditional on a

binary covariate, $Z = 0, 1$, where $Y_{\bar{D}} \mid Z = 0 \sim N(0, 1)$, $Y_{\bar{D}} \mid Z = 1 \sim N(\mu_{\bar{D}_1}, 1)$, $Y_D \mid Z = 0 \sim N(\mu_{D_0}, 1)$, and $Y_D \mid Z = 1 \sim N(\mu_{D_1}, 1)$. The induced $\mathcal{A}$ROC is

$$\mathcal{A}\mathrm{ROC}_\theta(t) = P[Z_D = 0]\, \Phi(\mu_{D_0} + \Phi^{-1}(t)) + P[Z_D = 1]\, \Phi(\mu_{D_1} - \mu_{\bar{D}_1} + \Phi^{-1}(t)), \quad (6)$$

and the asymptotic variance of the non-parametric estimator is

$$\frac{V(t)}{n_D} = \frac{\mathcal{A}\mathrm{ROC}_\theta(t)\,(1 - \mathcal{A}\mathrm{ROC}_\theta(t))}{n_D} + \frac{P[Z_D = 0]^2}{P[Z_{\bar{D}} = 0]} \left(\frac{\phi(\mu_{D_0} + \Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} \right)^2 \cdot \frac{t(1-t)}{n_{\bar{D}}} + \frac{P[Z_D = 1]^2}{P[Z_{\bar{D}} = 1]} \left(\frac{\phi(\mu_{D_1} - \mu_{\bar{D}_1} + \Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} \right)^2 \cdot \frac{t(1-t)}{n_{\bar{D}}}, \quad (7)$$

where ϕ is the standard normal density function. Due to the invariance of the $\mathcal{A}$ROC with respect to monotone transformations, this model simply assumes that there exists a monotone increasing transformation which makes Y normal in cases and controls, conditional on Z. All of the assumptions laid out in Section 4 are satisfied under this model. We set $\mu_{\bar{D}_1} = 0.2$, $\mu_{D_0} = 0.9$, $\mu_{D_1} = 0.9$, $P[Z_{\bar{D}} = 1] = 0.7$, and $P[Z_D = 1] = 0.3$ and consider estimation at $t = 0.05, 0.10, 0.20, 0.50$. The $\mathcal{A}$ROC values are $0.21, 0.33, 0.50, 0.80$ and the two components of asymptotic variance are $n_D^{-1}(0.33, 0.44, 0.50, 0.32)$ and $n_{\bar{D}}^{-1}(1.41, 1.40, 1.14, 0.40)$.

We simulated 5,000 datasets, where $n_D = n_{\bar{D}}$ varies between 100 and 1,000 (see Table 1). Note that, with $n_D = n_{\bar{D}} = 100$ and $P[Z_D = 1] = 0.3$, there are approximately 30 cases with $Z = 1$. In terms of percent bias, defined as $\dfrac{avg(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t)) - \mathcal{A}\mathrm{ROC}_\theta(t)}{\mathcal{A}\mathrm{ROC}_\theta(t)}$, where $avg(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t))$ is the average $\mathcal{A}$ROC estimate, the $\mathcal{A}$ROC estimator performs very well, except for some modest bias when both t and $n_D = n_{\bar{D}}$ are small. The percent bias in the non-parametric variance estimate (using rectangular kernel density estimates) is defined as $\dfrac{median\,(\hat{V(t)}) - \widehat{Var}(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t))}{\widehat{Var}(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t))}$, where the median variance estimate is calculated because of the skewed distribution of the variance estimates, and $\widehat{Var}(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t))$ is the sample variance in the $\mathcal{A}$ROC estimates. The

variance estimator tends to underestimate the true variance, and most of this bias comes from estimating the second component of variance. There is substantial bias when t is small, but this disappears for larger t. The percent difference between the asymptotic and sample variances of the $\mathcal{A}$ROC estimates, $\frac{V(t) - \widehat{Var}(\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t))}{\widehat{Var}(\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t))}$, shows that they tend to be close, with differences only when both t and $n_D = n_{\bar{D}}$ are small. Coverage probabilities based on non-parametric variance estimates are provided. Coverage based on *logit* transformations, which have been shown to improve coverage for the pooled ROC when t is close to 0 or 1 (Pepe, 2003 [p. 102]), are also shown. Only *logit*-based coverage is shown when both t and $n_D = n_{\bar{D}}$ are small, since $\mathcal{A}$ROC estimates are frequently close to zero. We find that coverage can be low with small t, but is very good for moderate t.

We also evaluate the performance of bootstrap variance estimates. Data is resampled 100 times conditional on D, and the sample variance of the $\mathcal{A}$ROC estimates is calculated. The percent bias in the bootstrap variance estimate, defined as with the non-parametric variance estimates, shows substantially less bias. Bootstrap coverage also tends to be better; coverage is good except when both t and $n_D = n_{\bar{D}}$ are small.

We compare non-parametric $\mathcal{A}$ROC estimates with semi-parametric estimates, based on a normal linear quantile model. Table 2 displays the percent difference in the estimates, defined as $avg\left(\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta};semi}(t) - \widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t)\right)/\mathcal{A}\text{ROC}_{\theta}(t)$, where $\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t)$ is the non-parametric estimate, $\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta};semi}(t)$ is the semi-parametric estimate, and the average is taken over 5,000 simulations. The estimates agree quite well. The estimated relative efficiency of the two estimators, $\frac{\widehat{Var}(\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta};semi}(t))}{\widehat{Var}(\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t))}$, where the variance is estimated over the 5,000 simulations, is also shown. The semi-parametric estimator yields substantial gains in efficiency, with larger gains for smaller

t and larger $n_D = n_{\bar{D}}$.

We explore the performance of the semi-parametric $\mathcal{A}$ROC estimator and its variance under the double binormal model (Lin and Jeon, 2003),

$$\begin{pmatrix} Y_{\bar{D}} \\ Z_{\bar{D}} \end{pmatrix} \sim BVN\left(\begin{pmatrix} \mu_{Y_{\bar{D}}} \\ \mu_{Z_{\bar{D}}} \end{pmatrix}, \begin{pmatrix} \sigma^2_{Y_{\bar{D}}} & \sigma_{Y_{\bar{D}}}\sigma_{Z_{\bar{D}}}\rho_{\bar{D}} \\ \sigma_{Y_{\bar{D}}}\sigma_{Z_{\bar{D}}}\rho_{\bar{D}} & \sigma^2_{Z_{\bar{D}}} \end{pmatrix}\right)$$
$$\begin{pmatrix} Y_D \\ Z_D \end{pmatrix} \sim BVN\left(\begin{pmatrix} \mu_{Y_D} \\ \mu_{Z_D} \end{pmatrix}, \begin{pmatrix} \sigma^2_{Y_D} & \sigma_{Y_D}\sigma_{Z_D}\rho_D \\ \sigma_{Y_D}\sigma_{Z_D}\rho_D & \sigma^2_{Z_D} \end{pmatrix}\right). \quad (8)$$

This is an extension of the classic binormal model for the pooled ROC curve (Swets, 1986; Hanley, 1988, 1996). The induced $\mathcal{A}$ROC is a binormal ROC curve with intercept and slope parameters $\frac{\mu_{Y_D}-\mu_{Y_{\bar{D}}}}{s\cdot\sigma_{Z_D}} - \frac{\rho_{\bar{D}}\sigma_{Y_{\bar{D}}}(\mu_{Z_D}-\mu_{Z_{\bar{D}}})}{s\cdot\sigma_{Z_{\bar{D}}}\sigma_{Z_D}}$ and $\frac{\frac{\sigma_{Y_{\bar{D}}}}{\sigma_{Z_D}}\sqrt{1-\rho^2_{\bar{D}}}}{s}$, where $s = \sqrt{\frac{\sigma^2_{Y_D}}{\sigma^2_{Z_D}}(1-\rho^2_D) + \left(\rho_{\bar{D}}\frac{\sigma_{Y_{\bar{D}}}}{\sigma_{Z_{\bar{D}}}} - \rho_D\frac{\sigma_{Y_D}}{\sigma_{Z_D}}\right)^2}$ (Janes and Pepe, 2006). Again, this model is more general than it first appears; it stipulates that there exists a monotone, increasing function which transforms (Y, Z) to bivariate normality in cases and controls (Janes and Pepe, 2006). All of the assumptions laid out in Section 4 are satisfied under this model. We apply the semi-parametric $\mathcal{A}$ROC estimator using a normal linear quantile model; this is the true model for $Y_{\bar{D}}$ given $Z_{\bar{D}}$. We set $\mu_{Y_{\bar{D}}} = \mu_{Z_{\bar{D}}} = 0$, $\sigma_{Y_{\bar{D}}} = \sigma_{Y_D} = 1$, $\sigma_{Z_{\bar{D}}} = \sigma_{Z_D} = 1.5$, $\rho_D = 0.6$, $\rho_{\bar{D}} = 0.2$, $\mu_{Y_D} = 0.7$, and $\mu_{Z_D} = 0.5$. The $\mathcal{A}$ROC values at $t = 0.05, 0.10, 0.20, 0.50$ are $0.16, 0.25, 0.39, 0.67$. The two components of asymptotic variance are $n_D^{-1}(0.24, 0.37, 0.49, 0.36)$ and $n_{\bar{D}}^{-1}(0.35, 0.53, 0.56, 0.23)$.

We simulated 5,000 datasets, where $n_D = n_{\bar{D}}$ varies between 100 and 1,000. The $\mathcal{A}$ROC estimator performs very well, except for some modest small sample bias for very small $n_D = n_{\bar{D}}$ and t (see Table 3). The semi-parametric variance estimator exhibits moderate small sample bias for the smallest sample sizes; the variance is consistently overestimated. This is primarily due to bias in the second component of

variance, which involves $\frac{\partial}{\partial\theta}\mathcal{A}\text{ROC}_\theta(t)$. Yet, coverage is reasonable. The asymptotic and sample variances agree quite well, except for some minor differences with the smallest sample sizes. Bootstrap variance estimates are good alternatives: they tend to exibit less bias, and have excellent coverage.

In summary, using quite general simulation models, we have found that the $\mathcal{A}$ROC estimators perform reasonably well in small samples. Varying parameter choices have produced similar or improved performance.

6 Illustration

We illustrate our methods using data from the Physicians' Health Study (PHS) (Gann et al., 2002). The PHS was a randomized, placebo-controlled study of aspirin and β-carotene among 22,071 US male physicians ages 40 to 84 years in 1982. A blood sample taken at enrollment was stored. For 429 men diagnosed with prostate cancer up to 12 years after enrollment (most before PSA was widely used for screening), and for 1,287 controls not diagnosed with prostate cancer during 12 years of follow-up, the serum was assayed for PSA. Controls were matched to cases with respect to age; for each case, three controls were selected who were within one year of age (Gann et al., 2002; Etzioni et al., 2004).

The goal of this sub-study is to determine how well PSA discriminates between men who did and did not go on to develop prostate cancer. The pooled ROC curve in the matched data is not of practical interest (Janes and Pepe, unpublished manuscript). It describes the ability of PSA to distinguish between cases and age-matched controls, an artificially constructed control group. More importantly, this

ROC curve is attenuated by matching on age in the design. We use the $\mathcal{A}$ROC to summarize the age-adjusted discriminatory accuracy of PSA.

Age-specific ROC curves for PSA, estimated using a binormal ROC-GLM model (Alonzo and Pepe, 2002), with quantiles based on a linear location-scale model (Heagerty and Pepe, 1999), are shown in Figure 3(a). Observe that there is very little variation in discrimination due to age. Hence, the $\mathcal{A}$ROC represents the common, age-specific ROC curve for PSA, and is a good summary of PSA performance.

The $\mathcal{A}$ROC for PSA is shown in Figure 3(b), estimated both using the semi-parametric estimator and using a binormal ROC-GLM model, where the control quantiles are estimated using a linear location-scale model (Heagerty and Pepe, 1999) for both methods. Bootstrapping is used for inference, and *logit*-based confidence intervals are overlaid at $t = 0.025$ and $t = 0.05$. The $\mathcal{A}$ROC describes the ability of PSA to discriminate between cases and controls of the same age. Using ROC-GLM, we estimate that 18% of cases can be detected (95% CI: 14% to 23%) when the age-specific FPF is held at 0.025, and 27% cases can be detected (95% CI: 23% to 32%) when the common FPF is increased to 0.05.

7 Discussion

We have proposed the $\mathcal{A}$ROC as a measure of covariate-adjusted discriminatory accuracy. This is the common covariate-specific ROC curve when the covariate does not affect discrimination, and a weighted average of covariate-specific ROC curves when the covariate does affect discrimination. Asymptotic distribution theory was developed for our non-parametric and semi-parametric $\mathcal{A}$ROC estimators, which perform

reasonably well in small samples. The consistent variance estimators also have good small sample performance, but bootstrap variance estimation is easier to implement and provides improved coverage. We have used the asymptotic variance expressions to investigate efficient study design (Janes and Pepe, unpublished manuscript). An intriguing result is that matching of controls to cases is optimal when covariates affect the marker but not discrimination. The optimal case-control ratio also follows from the variance expressions.

Covariate adjustment is important for covariates which affect marker observations but not discrimination. Their effects must be adjusted for to avoid bias in ROC estimation. However, covariates which are markers in their own right might be better combined with the marker in the risk score in order to examine the value of the combination or the incremental value of the marker over the covariates. Covariates which affect discrimination should be used to estimate covariate-specific ROC curves. The $\mathcal{A}$ROC can be used in such situations to summarize the covariate-specific ROC curves. This may be particularly useful for comparing the performance of markers. Methods to compare covariate-adjusted ROC curves are under development.

The $\mathcal{A}$ROC is a simple vertical average of covariate-specific ROC curves. This is just one of many possible ways of summarizing covariate-specific ROC curves. Our approach is appealing because it results in a true ROC curve, i.e., a plot of the TPF versus FPF for a set of classification rules. Many potential summary measures are not true ROC curves. The $\mathcal{A}$ROC also makes sense in applications where controlling the FPF across covariate groups is desirable (e.g., cancer screening). There are applications, however, where controlling the covariate-specific false negative fractions is more appropriate; this suggests averaging horizontally. The vertical and horizontal

ROC averages are exactly the same when the covariates do not affect discrimination, but differ more generally. The horizontal version, a simple extension of our methods, describes the accuracy of rules which classify using covariate-specific thresholds that control the covariate-specific false negative fractions.

The area under the $\mathcal{A}$ROC, the $\mathcal{A}$-AUC, can be interpreted as the probability of correctly ordering a randomly chosen case and control observation with the same covariate value, $\mathcal{A}\text{-AUC} = P[Y_D > Y_{\bar{D}Z_D}]$. This statistical summary deserves further development and might serve as the basis of tests to compare covariate-adjusted ROC curves for different markers.

References

1. T. A. Alonzo and M. S. Pepe. Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, 3:421–32, 2002.

2. T. Cai and M. S. Pepe. Semi-parametric ROC analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association*, 97:1099–107, 2002.

3. T. J. Cole. The LMS method for constructing normalized growth standards. *European Journal of Clinical Nutrition*, 44:45–60, 1990.

4. T. J. Cole and P. J. Green. Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, 11:1305–19, 1992.

5. R. Etzioni, Falcon S., P. H. Gann, C. L. Kooperberg, D. F. Penson, and M. J. Stampfer. Prostate-specific antigen and free prostate-specific antigen in the early detection of prostate cancer: Do combination tests improve detection? *Cancer Epidemiology, Biomarkers and Prevention*, 13:1640–5, 2004.

6. D. Faraggi. Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician*, 52(Part 2):179–92, 2003.

7. T. S. Ferguson. *A Course in Large Sample Theory*. Chapman and Hall, 1996.

8. P. H. Gann, C. H. Hennekens, and M. J. Stampfer. A prospective evaluation of plasma prostate-specific antigen for detection of prostatic cancer. *Journal of the American Medical Association*, 273:289–94, 1995.

9. P. V. Hammill, T. A. Drizd, C. L. Johnson, R. B. Reed, and A. F. Roche. NCHS growth curves for children birth - 18 years. Technical Report 11, United States, Vital Health Statistics, 1977. pp. 1-74.

10. J. A. Hanley. The robustness of the 'binormal' assumptions used in fitting ROC curves. *Medical Decision Making*, 8:197–203, 1988.

11. J. A. Hanley. The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine*, 15:1575–85, 1996.

12. P. Heagerty and M. S. Pepe. Semi-parametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics*, 48:533–51, 1999.

13. H. Janes and M. S. Pepe. Matching in studies of classification accuracy: Implications for bias, efficiency, and assessment of incremental value. *Biometrics*, (revision requested).

14. H Janes and MS Pepe. Adjusting for covariate effects on classification accuracy using the covariate-adjusted ROC curve. Technical Report 283, University of Washington, Department of Biostatistics Working Paper Series, 2006.

15. C. T. Le. Evaluation of confounding effects in ROC studies. *Biometrics*, 53:998–1007, 1997.

16. Y. Lin and Y. Jeon. Discriminant analysis through a semi-parametric model. *Biometrika*, 90:379–92, 2003.

17. M. S. Pepe. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, 54:124–35, 1998.

18. M. S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

19. M. S. Pepe, R. Etzioni, Z. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M . Winget, and Y. Yatsui. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, 93(14):1054–61, 2001.

20. M.S. Pepe. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56:352–9, 2000.

21. E. F. Schisterman, D. Faraggi, and B. Reiser. Adjusting the generalized ROC curve for covariates. *Statistics in Medicine*, 23:3319–31, 2004.

22. B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

23. J. A. Swets. Indices of discrimination or diagnostic accuracy: Their ROCs and implied methods. *Psychological Bulletin*, 99:100–17, 1986.

24. A. Toledano and C. A. Gatsonis. Regression analysis of correlated receiver operating characteristic data. *Academic Radiology*, 2(Supplement 1):S30–6, 1995.

25. A. A. N. Tosteson and C. B. Begg. A general regression methodology for ROC curve estimation. *Medical Decision Making*, 8:204–15, 1988.

26. Y. Zheng. *Semi-Parametric Methods for Longitudinal Diagnostic Accuracy*. PhD thesis, University of Washington, 2002.

Appendix

A.1 Proof of Theorem 1 We write

$$\begin{aligned}
\sqrt{n_D}\,(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat\theta}(t) - \mathcal{A}\mathrm{ROC}_\theta(t)) &= \sqrt{n_D}\,\left(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat\theta}(t) - \mathcal{A}\mathrm{ROC}_{\hat\theta}(t)\right) \\
&\quad + \sqrt{n_D}\,(\mathcal{A}\mathrm{ROC}_{\hat\theta}(t) - \mathcal{A}\mathrm{ROC}_\theta(t)) \\
&= \frac{1}{\sqrt{n_D}} \sum_{i=1}^{n_D} \mathrm{I}\left[Y_{D_i} > q_{Z_{D_i}}(t;\hat\theta)\right] - P[Y_D > q_{Z_D}(t;\hat\theta) \mid \hat\theta] \\
&\quad + \sqrt{n_D}\,\left(P[Y_D > q_{Z_D}(t;\hat\theta) \mid \hat\theta] - P[Y_D > q_{Z_D}(t;\theta)]\right) \\
&\equiv A_n + B_n.
\end{aligned}$$

Note that $B_n = \sqrt{n_D}\,(g(\hat\theta) - g(\theta))$, and by **C(1)-C(3)**, the delta method (Ferguson 1996 [p. 45]) and Slutsky's Theorem, $B_n \xrightarrow{d} N(0, \sigma_b^2)$ as $n_D, n_{\bar D} \to \infty$, where $\sigma_b^2 = \lambda\, \frac{\partial}{\partial\theta}\mathcal{A}\mathrm{ROC}_\theta(t)\, \Sigma_\theta\, \frac{\partial}{\partial\theta}\mathcal{A}\mathrm{ROC}_\theta(t)^T$. Now, we write $A_n = \frac{1}{\sqrt{n_D}} \sum_{i=1}^{n_D} A_{n_i}$ and find its asymptotic distribution conditional on $\hat\theta$, using the Lindeberg-Feller Central Limit Theorem (LFCLT). First, note that $E[A_{n_i}|\,\hat\theta\,] = 0$ and $Var[A_{n_i}|\,\hat\theta\,] = \mathcal{A}\mathrm{ROC}_{\hat\theta}(t)\ (1-$

$\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)$). Convergence under the LFCLT requires that

$$\frac{1}{n_D\,\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)\,(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t))}\sum_{i=1}^{n_D} E[A_{n_i}^2\mathrm{I}\left[|A_{n_i}|\ \geq \epsilon\, n_D \mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)\,(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t))\right]] \tag{1 a}$$

converges to zero as $n_D \to \infty$ for all $\epsilon > 0$. But $A_{n_i}^2\mathrm{I}[|\ A_{n_i}\ | \geq \epsilon \cdot n_D\ \mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)$ $(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t))]$ takes the value $(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t))^2 \cdot \mathrm{I}\left[\frac{1}{\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)} \geq \epsilon n_D\right]$ with probability $\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)$, and $\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)^2 \cdot \mathrm{I}\left[\frac{1}{1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)} \geq \epsilon n_D\right]$, with probability 1 $-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)$. Hence,(1a) becomes $(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)) \cdot \mathrm{I}\left[\frac{1}{\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)} \geq \epsilon n_D\right] + \mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t) \cdot$ $\mathrm{I}\left[\frac{1}{1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)} \geq \epsilon n_D\right]$. **C(4)** and **C(5)** ensure that this converges to zero. Thus, conditional on $\hat{\theta}$, $\frac{A_n}{\sqrt{\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)\,(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t))}} \xrightarrow{d} N(0,1)$ as $n_D \to \infty$. Finally, the asymptotic distribution of $\frac{A_n}{\sqrt{\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)\,(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t))}}$ conditional on $\hat{\theta}$ is the same as that of $\frac{A_n}{\sqrt{\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)\,(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t))}}$ conditional on B_n, since it is functionally independent of $\hat{\theta}$ and $B_n = \sqrt{n_D}\ (g(\hat{\theta}) - g(\theta))$. By **C(2)** and **C(3)**, $\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)\,(1-\mathcal{A}\mathrm{ROC}_{\hat{\theta}}(t)) \xrightarrow{P} \mathcal{A}\mathrm{ROC}_{\theta}(t)\,(1-\mathcal{A}\mathrm{ROC}_{\theta}(t))$ as $n_{\bar{D}} \to \infty$. By Slutsky's Theorem, $\begin{pmatrix} \frac{A_n}{\sqrt{\mathcal{A}\mathrm{ROC}_{\theta}(t)\,(1-\mathcal{A}\mathrm{ROC}_{\theta}(t))}} \\ B_n \end{pmatrix} \xrightarrow{d} BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}\right)$ as $n_D, n_{\bar{D}} \to \infty$. The continuous mapping theorem then yields the desired result.

A.2 Non-Parametric Estimation of the $\mathcal{A}$ROC We prove that **C(2)** and **C(3)** are satisfied with empirical quantile estimates. Under **C(6)**, by standard empirical process theory,(Ferguson 1996 [p. 91]), for a fixed stratum Z and conditional on $n_{\bar{D}Z}$, $\sqrt{n_{\bar{D}Z}}\left(\hat{S}_{\bar{D}Z}^{-1}(t) - S_{\bar{D}Z}^{-1}(t)\right) \xrightarrow{d} N(0,\sigma_Z^2)$ as $n_{\bar{D}Z} \to \infty$, where $\sigma_Z^2 = \frac{t(1-t)}{f_{\bar{D}Z}^2(S_{\bar{D}Z}^{-1}(t))}$. By

C(1), $\frac{n_{\bar{D}Z}}{n_{\bar{D}}} \xrightarrow{P} p_{Z_{\bar{D}}}(Z)$ as $n_{\bar{D}} \to \infty$. Hence, for all ϵ, there exists N such that

$$\begin{aligned}
P[\sqrt{n_{\bar{D}Z}}\ \left(\hat{S}^{-1}_{\bar{D}Z}(t) - S^{-1}_{\bar{D}Z}(t)\right) \le y] &= E[P[\sqrt{n_{\bar{D}Z}}\ \left(\hat{S}^{-1}_{\bar{D}Z}(t) - S^{-1}_{\bar{D}Z}(t)\right) \le y \mid n_{\bar{D}Z}]] \\
&= E[P[\sqrt{n_{\bar{D}Z}}\ \left(\hat{S}^{-1}_{\bar{D}Z}(t) - S^{-1}_{\bar{D}Z}(t)\right) \le y \mid n_{\bar{D}Z}]\, \mathrm{I}\,[n_{\bar{D}Z} > N]] \\
&\quad + E[P[\sqrt{n_{\bar{D}Z}}\ \left(\hat{S}^{-1}_{\bar{D}Z}(t) - S^{-1}_{\bar{D}Z}(t)\right) \le y \mid n_{\bar{D}Z}]\, \mathrm{I}\,[n_{\bar{D}Z} \le N]] \\
&< (\Phi(y/\sigma_Z) + \epsilon) \cdot P[n_{\bar{D}Z} > N] \\
&\quad + E[P[\sqrt{n_{\bar{D}Z}}\ \left(\hat{S}^{-1}_{\bar{D}Z}(t) - S^{-1}_{\bar{D}Z}(t)\right) \le y \mid n_{\bar{D}Z}]\, \mathrm{I}\,[n_{\bar{D}Z} \le N]].
\end{aligned}$$

Since $n_{\bar{D}Z} \to \infty$, the second term can be made arbitrarily small, and $P[n_{\bar{D}Z} > N]$ arbitrarily close to 1, by choosing N large enough. Thus, $\sqrt{n_{\bar{D}Z}}(\hat{S}^{-1}_{\bar{D}Z}(t) - S^{-1}_{\bar{D}Z}(t)) \xrightarrow{d} N(0, \sigma^2_Z)$, and by Slutsky's Theorem, $\sqrt{n_{\bar{D}}}(\hat{S}^{-1}_{\bar{D}Z}(t) - S^{-1}_{\bar{D}Z}(t)) \xrightarrow{d} N(0, \frac{\sigma^2_Z}{p_{Z_{\bar{D}}(Z)}})$ as $n_{\bar{D}} \to \infty$. Because observations in different strata are independent, marginal convergence implies joint asymptotic normality, with variance-covariance matrix $\Sigma_\theta = diag\left(\frac{\sigma^2_Z}{p_{Z_{\bar{D}}(Z)}}\right)$. We also calculate the form of $\frac{\partial}{\partial\theta}\mathcal{A}\mathrm{ROC}_\theta(t)$. We have

$$\begin{aligned}
\frac{\partial}{\partial\theta_Z}\mathcal{A}\mathrm{ROC}_\theta(t) &= \frac{\partial}{\partial S^{-1}_{\bar{D}Z}(t)} E_{Z_D}[P[Y_D > S^{-1}_{\bar{D}Z_D}(t) \mid Z_D]] \\
&= -E_{Z_D}[f_{DZ_D}(S^{-1}_{\bar{D}Z_D}(t)) \cdot \frac{\partial}{\partial S^{-1}_{\bar{D}Z}(t)} S^{-1}_{\bar{D}Z_D}(t)] \\
&= -f_{DZ}\left(S^{-1}_{\bar{D}Z}(t)\right) \cdot p_{Z_D}(Z),
\end{aligned}$$

and $V(t)$ reduces to (5).

A.3 We prove consistency of the estimated asymptotic variance of the semi-parametric $\mathcal{A}$ROC estimator. We write the estimate of $\frac{\partial}{\partial\theta_j}\mathcal{A}\mathrm{ROC}_\theta(t)$ as

$$\begin{aligned}
&\frac{\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}+h^j(n)}(t) - \mathcal{A}\mathrm{ROC}_{\theta+h^j(n)}(t)}{2h(n)} + \frac{\mathcal{A}\mathrm{ROC}_{\theta+h^j(n)}(t) - \mathcal{A}\mathrm{ROC}_{\theta-h^j(n)}(t)}{2h(n)} \\
&\quad - \frac{\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}-h^j(n)}(t) - \mathcal{A}\mathrm{ROC}_{\theta-h^j(n)}(t)}{2h(n)}. \qquad (2\text{ a})
\end{aligned}$$

Consider the first component. We claim that $\sqrt{n_D}(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}+h^j(n)}(t) - \mathcal{A}\mathrm{ROC}_{\theta+h^j(n)}(t)) \xrightarrow{d} N(0, V(t))$. The proof of this fact is very similar to the proof that $\sqrt{n_D}(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t) - \mathcal{A}\mathrm{ROC}_{\theta}(t))$ is asymptotically normal, and hence $O_p(1)$, proven in appendix A.1. Hence, $\dfrac{\sqrt{n_D}(\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}+h^j(n)}(t) - \mathcal{A}\mathrm{ROC}_{\theta+h^j(n)}(t))}{2\sqrt{n_D}h(n)} \xrightarrow{P} 0$, since the denominator converges to ∞. A similar argument can be used to prove that the third term in (2a) converges to 0. Finally, $\dfrac{\mathcal{A}\mathrm{ROC}_{\theta+h^j(n)}(t) - \mathcal{A}\mathrm{ROC}_{\theta-h^j(n)}(t)}{2h(n)} \xrightarrow{P} \dfrac{\partial}{\partial\theta}\mathcal{A}\mathrm{ROC}_{\theta}(t)$, by continuity of $\mathcal{A}\mathrm{ROC}_{\theta}(t)$ in θ (assumption **C(3)**). Hence, our estimate of $\frac{\partial}{\partial\theta_j}\mathcal{A}\mathrm{ROC}_{\theta}(t)$ is consistent. Now, with $\hat{\Sigma}_{\theta} \xrightarrow{P} \Sigma_{\theta}$, $\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t) \xrightarrow{P} \mathcal{A}\mathrm{ROC}_{\theta}(t)$, and consistency of the derivative estimator, we have consistency of the composite variance estimator.

A.4 We prove consistency of the estimated asymptotic variance of the non-parametric $\mathcal{A}$ROC estimator. We have $\hat{p}_{Z_D}(Z) \xrightarrow{P} p_{Z_D}(Z)$ and $\hat{p}_{Z_{\bar{D}}}(Z) \xrightarrow{P} p_{Z_{\bar{D}}}(Z)$ for all $Z = 1, \ldots, K$, and by standard empirical process theory, (Ferguson 1996 [p. 91]) under **C(6)**, $\hat{S}^{-1}_{\bar{D}Z}(t) \xrightarrow{P} S^{-1}_{\bar{D}Z}(t)$ as $n_{\bar{D}} \to \infty$. We write

$$
\begin{aligned}
&\mid \hat{f}_{DZ}(\hat{S}^{-1}_{\bar{D}Z}(t)) - f_{DZ}(S^{-1}_{\bar{D}Z}(t)) \mid = \\
&\mid \hat{f}_{DZ}(\hat{S}^{-1}_{\bar{D}Z}(t)) - f_{DZ}(\hat{S}^{-1}_{\bar{D}Z}(t)) + f_{DZ}(\hat{S}^{-1}_{\bar{D}Z}(t)) - f_{DZ}(S^{-1}_{\bar{D}Z}(t)) \mid \\
&\leq \mid \hat{f}_{DZ}(\hat{S}^{-1}_{\bar{D}Z}(t)) - f_{DZ}(\hat{S}^{-1}_{\bar{D}Z}(t)) \mid + \mid f_{DZ}(\hat{S}^{-1}_{\bar{D}Z}(t)) - f_{DZ}(S^{-1}_{\bar{D}Z}(t)) \mid .
\end{aligned}
$$

The first term converges in probability to zero by the uniform consistency of $\hat{f}_{DZ}$, while the second term converges in probability to zero by the consistency of $\hat{S}^{-1}_{\bar{D}Z}(t)$, **C(8)**, and the continuous mapping theorem. Hence, $\hat{f}_{DZ}(\hat{S}^{-1}_{\bar{D}Z}(t)) \xrightarrow{P} f_{DZ}(S^{-1}_{\bar{D}Z}(t))$ as $n_D, n_{\bar{D}} \to \infty$. A similar argument shows $\hat{f}_{\bar{D}Z}(\hat{S}^{-1}_{\bar{D}Z}(t))$ is also consistent. The variance estimator is a continuous function of these components, and under **C(6)** is consistent.

Table 1: Small sample performance of the non-parametric estimator of the $\mathcal{A}$ROC, under model (6), based on 5,000 simulations. The sample size, $n_D = n_{\bar{D}}$, varies between 100 and 1,000. The non-parametric variance estimator uses rectangular kernel density estimates. Percent bias in $\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t)$, percent difference between asymptotic and sample variances, percent bias in the non-parametric variance estimator, and coverage are shown at four FPF's of interest. Only coverage based on *logit*-transformations is shown for small t and $n_D = n_{\bar{D}}$, since $\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t)$ is frequently close to zero. Bootstrap variance estimates (based on 100 bootstrap samples) and associated coverage are also shown.

			$t = 0.05$					
	% Bias in $\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t)$	% Diff. between Asym. and Sample Var.	% Bias in Non-Par. Var. Est.	95% CI Coverage	*logit* Coverage	% Bias in Boot. Var. Est.	95% CI Coverage	*logit* Coverage
$n_D = 100$	6.15	21.34	-30.65	–	89.36	-25.25	–	87.32
$n_D = 200$	10.05	7.83	-29.67	–	88.14	-6.30	–	91.68
$n_D = 500$	2.40	6.88	-20.25	–	91.58	4.36	–	94.86
$n_D = 1000$	1.71	2.28	-16.24	91.90	92.14	2.92	94.40	94.18
			$t = 0.10$					
	% Bias in $\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t)$	% Diff. between Asym. and Sample Var.	% Bias in Non-Par. Var. Est.	95% CI Coverage	*logit* Coverage	% Bias in Boot. Var. Est.	95% CI Coverage	*logit* Coverage
$n_D = 100$	7.83	15.48	-19.75	–	91.22	-7.32	–	93.38
$n_D = 200$	4.37	7.06	-17.53	–	91.88	0.79	–	94.82
$n_D = 500$	0.75	3.88	-12.94	–	93.22	2.27	–	94.60
$n_D = 1000$	0.94	1.39	-10.53	92.86	92.18	1.68	94.10	94.40
			$t = 0.20$					
	% Bias in $\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t)$	% Diff. between Asym. and Sample Var.	% Bias in Non-Par. Var. Est.	95% CI Coverage	*logit* Coverage	% Bias in Boot. Var. Est.	95% CI Coverage	*logit* Coverage
$n_D = 100$	2.21	12.70	-4.42	91.92	95.36	1.04	93.88	97.08
$n_D = 200$	1.45	6.64	-6.00	92.76	94.38	3.08	94.26	95.68
$n_D = 500$	0.24	3.51	-3.96	93.74	94.42	2.96	94.44	95.08
$n_D = 1000$	0.33	4.26	-1.12	94.14	94.42	5.04	94.16	94.56
			$t = 0.50$					
	% Bias in $\widehat{\mathcal{A}\text{ROC}}_{\hat{\theta}}(t)$	% Diff. between Asym. and Sample Var.	% Bias in Non-Par. Var. Est.	95% CI Coverage	*logit* Coverage	% Bias in Boot. Var. Est.	95% CI Coverage	*logit* Coverage
$n_D = 100$	0.02	2.05	-3.67	92.16	96.98	8.41	95.00	97.26
$n_D = 200$	-0.11	-0.10	-0.47	93.42	95.86	4.38	94.18	95.98
$n_D = 500$	-0.05	-1.23	1.78	94.78	95.98	0.71	94.64	95.42
$n_D = 1000$	-0.01	2.48	6.37	95.36	95.88	3.76	94.78	95.20

Table 2: Comparison of the non-parametric and semi-parametric estimates of the $\mathcal{A}$ROC, under model (6), based on 5,000 simulations. The sample size, $n_D = n_{\bar{D}}$, varies between 100 and 1,000. The semi-parametric estimator uses a normal linear quantile model. The percent difference in the $\mathcal{A}$ROC estimates and relative efficiency are shown at four FPF's of interest.

	$t = 0.05$		$t = 0.10$		$t = 0.20$		$t = 0.50$	
	% Diff. in Estimates	Relative Efficiency	% Diff. in Estimates	Relative Efficiency	% Diff. in Estimates	Relative Efficiency	% Diff. in Estimates	Relative Efficiency
$n_D = 100$	-0.20	0.61	-5.33	0.68	-2.24	0.76	-0.75	0.82
$n_D = 200$	-6.87	0.54	-2.98	0.64	-1.25	0.72	-0.33	0.81
$n_D = 500$	-1.90	0.51	-0.88	0.61	-0.47	0.71	-0.17	0.79
$n_D = 1000$	-0.92	0.50	-0.52	0.61	-0.12	0.71	-0.04	0.82

Table 3: Small sample performance of the semi-parametric estimator of the $\mathcal{A}$ROC under model (8), based on 5,000 simulations. The sample size, $n_D = n_{\bar{D}}$, varies between 100 and 1,000. The quantiles are estimated using a normal linear model, and the semi-parametric variance estimate uses a bandwidth of $h = 0.04$. Percent bias in $\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t)$, percent difference between asymptotic and sample variances, percent bias in the semi-parametric variance estimator, and coverage are shown at four FPF's of interest. Only coverage based on *logit*-transformations is shown for small t and $n_D = n_{\bar{D}}$, since $\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t)$ is frequently close to zero. Bootstrap variance estimates (based on 100 bootstrap samples) and associated coverage are also shown.

			$t = 0.05$					
	% Bias in $\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t)$	% Diff. between Asym. and Sample Var.	% Bias in Semi-Par. Var. Est.	95% CI Coverage	*logit* Coverage	% Bias in Boot. Var. Est.	95% CI Coverage	*logit* Coverage
$n_D = 100$	14.35	-7.64	1.50	–	91.52	6.25	–	93.20
$n_D = 200$	6.47	0.28	5.17	–	93.88	6.67	–	94.64
$n_D = 500$	2.97	0.21	3.70	–	94.60	1.97	–	95.40
$n_D = 1000$	0.85	1.64	1.95	94.32	95.08	2.20	94.42	95.30
			$t = 0.10$					
	% Bias in $\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t)$	% Diff. between Asym. and Sample Var.	% Bias in Semi-Par. Var. Est.	95% CI Coverage	*logit* Coverage	% Bias in Boot. Var. Est.	95% CI Coverage	*logit* Coverage
$n_D = 100$	7.08	2.27	8.25	–	93.30	2.47	–	94.96
$n_D = 200$	3.23	3.26	4.87	–	94.28	3.79	–	95.48
$n_D = 500$	1.34	-2.22	-2.72	–	94.22	-3.09	–	94.82
$n_D = 1000$	0.34	2.47	1.48	94.66	95.10	2.48	95.00	95.44
			$t = 0.20$					
	% Bias in $\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t)$	% Diff. between Asym. and Sample Var.	% Bias in Semi-Par. Var. Est.	95% CI Coverage	*logit* Coverage	% Bias in Boot. Var. Est.	95% CI Coverage	*logit* Coverage
$n_D = 100$	2.46	5.14	9.02	92.54	94.02	-0.12	93.50	95.86
$n_D = 200$	1.26	2.31	4.54	93.80	94.70	0.25	94.38	95.54
$n_D = 500$	0.37	-0.55	-0.26	94.10	95.42	-1.71	94.30	94.72
$n_D = 1000$	0.03	-0.33	0.79	94.48	94.72	-0.17	94.42	94.66
			$t = 0.50$					
	% Bias in $\widehat{\mathcal{A}\mathrm{ROC}}_{\hat{\theta}}(t)$	% Diff. between Asym. and Sample Var.	% Bias in Semi-Par. Var. Est.	95% CI Coverage	*logit* Coverage	% Bias in Boot. Var. Est.	95% CI Coverage	*logit* Coverage
$n_D = 100$	-0.62	3.66	13.89	93.52	95.98	9.06	94.70	96.76
$n_D = 200$	-0.44	1.08	9.75	94.58	95.44	4.63	94.92	95.68
$n_D = 500$	-0.21	1.38	5.50	95.02	95.26	4.35	94.82	95.18
$n_D = 1000$	-0.14	0.45	3.13	95.00	95.24	3.24	94.68	94.88

Figure 1: Fictitious data for a marker Y and binary covariate $Z = 0, 1$. Under scenario 1, $P[Z = 1|D = 0] = 0.10$ and $P[Z = 1|D = 1] = 0.50$. Under scenario 2, $P[Z = 1|D = 0] = P[Z = 1|D = 1] = 0.50$. (a) The densities of Y conditional on $Z = 0$, conditional on $Z = 1$, marginally under scenario 1, and marginally under scenario 2. The solid line represents the case density, and the dashed line the control density. A common threshold of 2.5 is indicated. (b) The common covariate-specific ROC curve (solid line), the pooled ROC curve under scenario 1 (dotted line) and the pooled ROC curve under scenario 2 (dashed line) The performances of the common threshold are indicated.

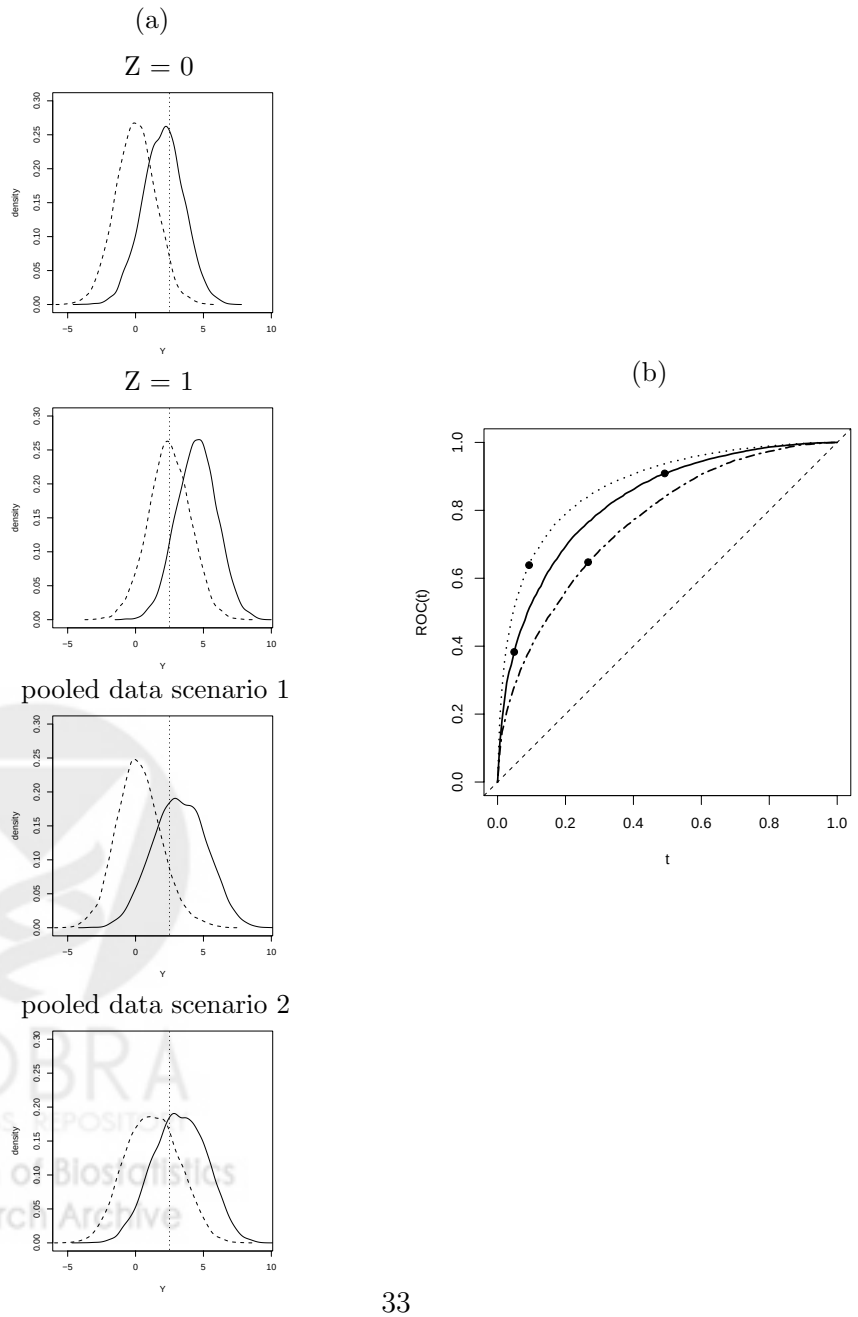

Figure 2: Two examples to illustrate that the ROC curve for the risk score, $R = P[D = 1|Y, Z]$ is different from the common covariate-specific ROC curve. In both examples, (Y, Z) is bivariate normal. The ROC curve for R (dotted line) and the common covariate-specific ROC curve (solid line) are shown. (a) Z is a good classifier but Y is not, and the two are relatively uncorrelated. (b) Both Y and Z are good classifiers, and are highly correlated.

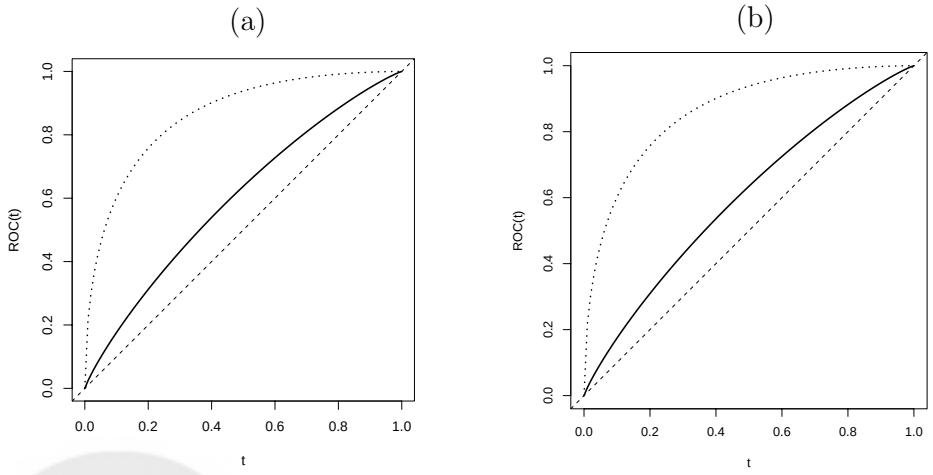

Figure 3: ROC curves for PSA in the PHS data. (a) Age-specific ROC curves, estimated using ROC-GLM. (b) The age-adjusted ROC curve, estimated using the semi-parametric estimator (solid line) and ROC-GLM (dashed line). 95% confidence intervals, based on bootstrapped variance estimates, are overlaid at $t = 0.025$ and $t = 0.05$.