Biostatistics (2004), **5**, 1, *pp*. 45–60
Printed in Great Britain

Semi-parametric ROC regression analysis with placement values

TIANXI CAI
Department of Biostatistics, Harvard University, Boston, MA 02115, USA
tcai@hsph.harvard.edu

SUMMARY

Advances in technology provide new diagnostic tests for early detection of disease. Frequently, these tests have continuous outcomes. One popular method to summarize the accuracy of such a test is the Receiver Operating Characteristic (ROC) curve. Methods for estimating ROC curves have long been available. To examine covariate effects, Pepe (1997, 2000) and Alonzo and Pepe (2002) proposed distribution-free approaches based on a parametric regression model for the ROC curve. Cai and Pepe (2002) extended the parametric ROC regression model by allowing an arbitrary non-parametric baseline function. In this paper, while we follow the same semi-parametric setting as in that paper, we highlight a new estimator that offers several improvements over the earlier work: superior efficiency, the ability to estimate the covariate effects without estimating the non-parametric baseline function and easy implementation with standard software. The methodology is applied to a case control dataset where we evaluate the accuracy of the prostate-specific antigen as a biomarker for early detection of prostate cancer. Simulation studies suggest that the new estimator under the semi-parametric model, while always being more robust, has efficiency that is comparable to or better than the Alonzo and Pepe (2002) estimator from the parametric model.

Keywords: Diagnostic accuracy; Estimating equation; Semi-parametric transformation model; U-process.

1. INTRODUCTION

Considerable research has focused on the development of new technologies to facilitate non-invasive diagnosis of serious diseases recently. Advances in genetics and immunology have yielded a multitude of biomarkers such as gene expression micro-arrays and protein mass spectrometry. Promising tumor biomarkers for cancer screening, such as prostate-specific antigen (PSA) for prostate cancer and CA-125 for ovarian cancer have become available. While diagnostic methods continue to improve, we note that most of the non-invasive tests are imperfect. We can characterize the imperfection of these tests by noting that two types of errors can occur. A diseased subject may produce a normal biomarker reading or a disease-free subject may test abnormal. Statistical methodologies to evaluate the accuracy of biomarkers or diagnostic tests are important in the medical field, and will become more so.

A well accepted measure of accuracy for diagnostic tests is the Receiver Operating Characteristic (ROC) curve (Swets and Pickett, 1982; Hanley, 1989; Begg, 1991). The ROC curve is a plot of the true positive rates (TPR) versus the false positive rates (FPR) at various threshold values for defining a positive test result. Specifically, let Y denote the test result with the convention that large values of Y are more indicative of disease. We use Y_{D} and $Y_{\bar{\mathrm{D}}}$ to denote the test result for diseased and non-diseased subjects, respectively. For any threshold value c, we classify subjects as diseased if $Y \geqslant c$ and disease free if $Y < c$. Then the true and false positive rates associated with this decision criterion are

$$\mathrm{TPR}(c) = 1 - F_{\mathrm{D}}(c) = P(Y_{\mathrm{D}} \geqslant c), \qquad \text{and} \qquad \mathrm{FPR}(c) = 1 - F_{\bar{\mathrm{D}}}(c) = P(Y_{\bar{\mathrm{D}}} \geqslant c),$$

respectively. At any attainable false positive rate u, the corresponding true positive rate is

$$\mathrm{ROC}(u) = 1 - F_{\mathrm{D}}\left\{F_{\bar{\mathrm{D}}}^{-1}(1-u)\right\} = P\{Y_{\mathrm{D}} \geqslant F_{\bar{\mathrm{D}}}^{-1}(1-u)\}, \qquad u \in (0,1). \tag{1.1}$$

The ROC curve displays the range of possible trade-offs between the true and false positive rates. This allows us to choose the threshold value in any particular setting depending on the costs of false positive decisions and false negative decisions.

ROC analysis has become a popular technique for summarizing the accuracy of diagnostic tests in biomedical applications, especially in radiology (Swets and Pickett, 1982; Metz, 1986; Gatsonis *et al.*, 1995). Methods for estimating individual ROC curves and for comparing ROC curves have long been available. To evaluate possible covariate effects on the accuracy of a diagnostic test, roughly two different approaches to ROC regression have emerged from the literature: (1) modeling the distribution of Y_{D} and $Y_{\bar{\mathrm{D}}}$ (Tosteson and Begg, 1988; Toledano and Gatsonis, 1995) to induce regression models for the ROC curves and (2) direct modeling of the ROC curves (Pepe, 1997, 2000; Alonzo and Pepe, 2002; Cai and Pepe, 2002). In the second approach, the regression parameters correspond to covariate effects on ROC curves. In general, they do not in the first approach.

In this paper, we focus on the direct modeling approach. Pepe (2000) has previously proposed parametric ROC-GLM models:

$$\mathrm{ROC}_{\mathbf{x}}(u) = P\left\{Y_{\mathrm{D}} \geqslant F_{\bar{\mathrm{D}},\mathbf{x}}^{-1}(1-u) \mid \mathbf{x}\right\} = g\left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + h(u)\right\}, \qquad 0 < u < 1, \tag{1.2}$$

where $\mathbf{x}$ denotes the covariate vector, $F_{\bar{\mathrm{D}},\mathbf{x}}(c) = P(Y_{\bar{\mathrm{D}}} < c \mid \mathbf{x})$, g^{-1} is a link function, and the baseline function h is specified parametrically. Cai and Pepe (2002) extended the parametric ROC-GLM model by allowing arbitrary non-parametric baseline functions. They estimated the baseline function $h(\cdot)$ and the covariate effect $\boldsymbol{\beta}$ simultaneously.

Here, while we consider the same semi-parametric setting as in Cai and Pepe (2002), we propose a new approach to the estimation of the parameters. The key observation is that the semi-parametric ROC-GLM model can be treated as a transformation model for the placement values, where placement values are defined as a particular standardization of the raw measurement relative to that of the non-diseased population (Pepe and Cai, 2002). This leads to the idea of using pairwise comparisons of the placement values to estimate the regression parameters. The main advantages of this new approach are that (1) we do not need to estimate the infinite-dimensional parameter h in order to estimate $\boldsymbol{\beta}$, (2) it yields greatly improved efficiency in estimating $\boldsymbol{\beta}$ and (3) estimation is easy to implement in practice.

In the next section, we describe the regression modeling framework. Procedures for making inference about $\boldsymbol{\beta}$ are given in Section 3. The estimation procedures for the ROC curve of a continuous test with a specific set of covariate values are outlined in Section 4. For illustration, we applied our method to a data set from a case control study that evaluates the accuracy of PSA as a biomarker for prostate cancer in men. Simulation studies were performed to compare the new procedure to the existing methods. The results of the example and the simulation studies are summarized in Section 5. In Section 6, we outline an alternative approach to estimation for the special case $g(x) = 1 - \exp\{-\exp(x)\}$. Some closing remarks are made in Section 7.

2. Semi-parametric ROC regression model

Suppose that the data for analysis are organized as N_{D} data records for n_{D} subjects with disease

$$\mathfrak{R}_{\text{D}} = \{(Y_{\text{D}ik}, \mathbf{X}_{ik} = (\mathbf{Z}_{ik}^{\mathsf{T}}, \mathbf{Z}_{\text{D}ik}^{\mathsf{T}})^{\mathsf{T}}),\ k = 1, \dots, K_i,\ i = 1, \dots, n_{\text{D}}\},$$

and $N_{\bar{\text{D}}}$ data records for $n_{\bar{\text{D}}}$ subjects without disease

$$\mathfrak{R}_{\bar{\text{D}}} = \{(Y_{\bar{\text{D}}jl}, \mathbf{Z}_{jl}),\ l = 1, \dots, K_j,\ j = n_{\text{D}} + 1, \dots, n_{\text{D}} + n_{\bar{\text{D}}}\},$$

where $N_{\text{D}} = \sum_{i=1}^{n_{\text{D}}} K_i$ and $N_{\bar{\text{D}}} = \sum_{j=n_{\text{D}}+1}^{n_{\text{D}}+n_{\bar{\text{D}}}} K_j$. $\mathbf{Z}$ denotes covariates that are relevant to all subjects, while $\mathbf{Z}_{\text{D}}$ denotes covariates that are specific to subjects with disease, but not relevant to non-diseased subjects. See Pepe (2000) or Alonzo and Pepe (2002) for discussion of difference covariate types in these models. As indicated by our notation, the data records may be clustered. For example, if multiple biomarkers are measured on a subject or if the same biomarker is measured at different times, then each subject may have several data records in the analysis.

Pepe and Cai (2002) define the placement value of $Y_{\text{D}ik}$ with covariate $\mathbf{X}_{ik}$ as

$$U_{\text{D}ik} = 1 - F_{\bar{\text{D}},\mathbf{Z}_{ik}}(Y_{\text{D}ik}).$$

In essence the non-diseased population is considered a reference population with respect to which Y is standardized by the placement value transformation. The placement value $U_{\text{D}ik}$ is the proportion of the non-diseased population with values exceeding $Y_{\text{D}ik}$. The cumulative distribution of $U_{\text{D}ik}$ was shown to be identical to the ROC curve:

$$P(U_{\text{D}ik} \leqslant u \mid \mathbf{X}_{ik}) = P\left\{Y_{\text{D}ik} \geqslant F_{\bar{\text{D}},\mathbf{Z}_{ik}}^{-1}(1-u)\right\} \equiv \text{ROC}_{\mathbf{X}_{ik}}(u).$$

Therefore the placement values directly relate to the discrimination capacity of Y. Moreover, regression models for ROC curves can be treated as regression models for the placement values. For example, we can interpret the parametric ROC-GLM as a parametric transformation model for the placement values:

$$h(U_{\text{D}ik}; \boldsymbol{\alpha}) = -\beta_0^{\mathsf{T}}\mathbf{Z}_{ik} + \epsilon_{ik}$$

where ϵ_{ik} has a specified distribution function g and $h(\cdot; \boldsymbol{\alpha})$ is specified up to $\boldsymbol{\alpha}$. Pepe and Cai (2002) proposed a pseudo-likelihood approach to the estimation of the regression parameters in this model.

Here, we consider a semi-parametric transformation model for the placement values:

$$h_0(U_{\text{D}ik}) = -\beta_0^{\mathsf{T}}\mathbf{X}_{ik} + \epsilon_{ik} \tag{2.1}$$

where $h_0(\cdot)$ is a completely unspecified increasing function. This model is essentially equivalent to the semi-parametric ROC model proposed by Cai and Pepe (2002):

$$\text{ROC}_{\mathbf{X}_{ik}}(u) = g\left\{\beta_0^{\mathsf{T}}\mathbf{X}_{ik} + h_0(u)\right\}, \qquad 0 < u < 1. \tag{2.2}$$

If the response variable $U_{\text{D}ik}$ is observable and $K_i = 1$, methods for making statistical inference based on the semi-parametric transformation models are already available (Dabrowska and Doksum, 1988; Cheng *et al.*, 1995, 1997; Scharfstein *et al.*, 1998). However, in our setting, since the distribution of $Y_{\bar{\text{D}}}$ is usually unknown, the placement values $U_{\text{D}ik}$ can only be estimated, rather than observed. In addition, we allow for clustered data, that is, each subject may have multiple data records in the analysis. In the next section, we present estimation procedures for the regression parameters in the presence of such uncertainty for clustered placement values.

3. Inference procedure for regression parameters

It is important to note that the transformation h in (2.1) is an increasing function and therefore preserves the order of $U_{\mathrm{D}ik}$. Analogous to the estimation procedure proposed by Cheng *et al.* (1995) for analyzing univariate survival data using semi-parametric transformation models, we base our inference for $\boldsymbol{\beta}$ on the pairwise comparisons of the placement values between diseased subjects. Under model (2.1), we have

$$P(U_{\mathrm{D}ik} \leqslant U_{\mathrm{D}\iota\kappa} \mid \mathbf{X}_{ik}, \mathbf{X}_{\iota\kappa}) = P\{\epsilon_{ik} - \epsilon_{\iota\kappa} \leqslant \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}} \mid \mathbf{X}_{ik}, \mathbf{X}_{\iota\kappa}\} = \xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}}), \quad \text{for } i \neq \iota,$$

where $\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}} = \mathbf{X}_{ik} - \mathbf{X}_{\iota\kappa}$, and

$$\xi(x) = P(\epsilon_{ik} - \epsilon_{\iota\kappa} \leqslant x) = \int_{-\infty}^{\infty} g(x+y)\,\mathrm{d}g(x).$$

If $F_{\bar{\mathrm{D}},\mathbf{Z}}$ is known, mimicking the score equation for binary regression, one can solve

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n_{\mathrm{D}}}\sum_{k=1}^{K_i}\sum_{\iota<i}\sum_{\kappa=1}^{K_\iota} w_\xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}}\left\{I(U_{\mathrm{D}ik} \leqslant U_{\mathrm{D}\iota\kappa}) - \xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\right\} = 0 \tag{3.1}$$

to obtain a consistent estimate of $\boldsymbol{\beta}$, where for any function f, $w_f(x) = \dot{f}(x)/[f(x)\{1-f(x)\}]$, and $\dot{f}(x) = \mathrm{d}f(x)/\mathrm{d}x$.

However, in practice, $F_{\bar{\mathrm{D}},\mathbf{Z}}$ is usually unknown and therefore the placement values $U_{\mathrm{D}ik}$ are not available. We use data from the non-diseased population, $\mathfrak{R}_{\bar{\mathrm{D}}}$, to estimate $F_{\bar{\mathrm{D}},\mathbf{Z}}$. We suggest to use non-parametric empirical estimates when applicable, for instance, when $\mathbf{Z}$ is discrete. When continuous covariates are included, we recommend marginal semi-parametric regression models for $F_{\bar{\mathrm{D}},\mathbf{Z}}(y)$. For example, one could assume a flexible semi-parametric location scale model (Pepe, 1998; Heagerty and Pepe, 1999). Other types of semi-parametric models such as linear transformation models (Han, 1997; Cai *et al.*, 2000) could also be considered. Here, we do not assume any specific form for the modeling of $F_{\bar{\mathrm{D}},\mathbf{Z}}(y)$, but require that the resulting estimate of $F_{\bar{\mathrm{D}},\mathbf{Z}}(y)$ is consistent and has a $n_{\bar{\mathrm{D}}}^{\frac{1}{2}}$ convergence rate. We note that one does not need to estimate $F_{\bar{\mathrm{D}},\mathbf{Z}}$ if it is independent of covariates, that is $F_{\bar{\mathrm{D}},\mathbf{Z}}(y) = F_0(y)$ for some distribution function F_0. $\boldsymbol{\beta}$ can be estimated by solving (3.1) directly since

$$I(U_{\mathrm{D}ik} \leqslant U_{\mathrm{D}\iota\kappa}) = I\{1 - F_0(Y_{\mathrm{D}ik}) \leqslant 1 - F_0(Y_{\mathrm{D}\iota\kappa})\} = I(Y_{\mathrm{D}ik} \geqslant Y_{\mathrm{D}\iota\kappa})$$

which does not involve observations from the non-diseased population.

To estimate $\boldsymbol{\beta}$ in general, we use the estimates of the placement values $\widehat{U}_{\mathrm{D}ik} = 1 - \widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}(Y_{\mathrm{D}ik})$ and substitute these estimates into the estimating equation (3.1) to form our final set of estimating equations:

$$\widehat{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n_{\mathrm{D}}}\sum_{k=1}^{K_i}\sum_{\iota<i}\sum_{\kappa=1}^{K_\iota} w_\xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}}\left\{I(\widehat{U}_{\mathrm{D}ik} \leqslant \widehat{U}_{\mathrm{D}\iota\kappa}) - \xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\right\} = 0\,, \tag{3.2}$$

Let $\widehat{\boldsymbol{\beta}}$ denote the solution to (3.2). In Appendix A, we show that $\widehat{\boldsymbol{\beta}}$ is unique for large n and is consistent. We show in Appendix B that the distribution of $\widehat{\boldsymbol{\beta}}$ can be approximated by a normal distribution with mean $\boldsymbol{\beta}_0$ and covariance matrix $n_{\mathrm{D}}^{-1}\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} = 4n_{\mathrm{D}}^{-3}\sum_{i=1}^{n_{\mathrm{D}}}\sum_{\iota<i}\sum_{\substack{\iota'<i\\ \iota'\neq\iota}}\mathcal{U}_{\mathrm{D}i\iota}\mathcal{U}_{i\iota'} + n_{\bar{\mathrm{D}}}^{-1}\sum_{j=n_{\mathrm{D}}+1}^{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}}\mathcal{U}_{\bar{\mathrm{D}}j}^2, \tag{3.3}$$

$$\mathcal{U}_{\mathrm{D}i\iota} = \mathbb{A}^{-1}\sum_{k=1}^{K_i}\sum_{\kappa=1}^{K_\iota} w_\xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}}\left\{I(U_{\mathrm{D}ik} \leqslant U_{\mathrm{D}\iota\kappa}) - \xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\right\}, \tag{3.4}$$

$\mathcal{U}_{\bar{\mathrm{D}}j} = \int \mathbf{W}_{\bar{\mathrm{D}}j}(u)\,\mathrm{d}u$, $\mathbb{A}$ and $\mathbf{W}_{\bar{\mathrm{D}}j}(u)$ are defined in appendix B. An estimate of $\boldsymbol{\Sigma}$ can be obtained by replacing all the theoretical quantities in $\boldsymbol{\Sigma}$ with their empirical counterparts.

4. Estimating the ROC curve

To estimate the ROC curve for a test in a setting specified by particular values of the covariates, it remains to estimate $h_0(u)$. At any given u, we consider the score equation for binary regression based on $I(U_{\mathrm{D}ik} \leqslant u)$ under the independence working assumption. Specifically, for any $u \in [a, b]$, we solve the following estimating equation for $h(u)$:

$$\sum_{i=1}^{n_{\mathrm{D}}}\sum_{k=1}^{K_i} w_g\left\{h(u) + \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_{ik}\right\}\left[I\left(\widehat{U}_{\mathrm{D}ik} \leqslant u\right) - g\left\{h(u) + \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_{ik}\right\}\right] = 0, \qquad (4.1)$$

where $a, b \in (0, 1)$ are constants such that $P(U_{\mathrm{D}11} > a)$ and $P(U_{\mathrm{D}11} < b)$ are positive. Let $\widehat{h}(\cdot)$ be the solution to (4.1). It follows from the uniform consistency of $\widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}}(\cdot)$ and similar arguments as in Appendix A that $\widehat{h}(u)$ is consistent for $u \in [a, b]$. Based on $\{\widehat{\boldsymbol{\beta}}, \widehat{h}(\cdot)\}$, a consistent estimate of the ROC curve for a test with covariates $\mathbf{x}$ is $\widehat{\mathrm{ROC}_{\mathbf{x}}}(u) = g\left\{\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{x} + \widehat{h}(u)\right\}$.

To obtain the distribution of $\widehat{\mathrm{ROC}_{\mathbf{x}}}(u)$, in Appendix C, we show that $R(u; \mathbf{x}) = n_{\mathrm{D}}^{\frac{1}{2}}\left[g^{-1}\left\{\widehat{\mathrm{ROC}_{\mathbf{x}}}(u)\right\} - g^{-1}\left\{\mathrm{ROC}_{\mathbf{x}}(u)\right\}\right]$ and $\widetilde{R}(u; \mathbf{x})$ converge weakly to the same Gaussian process, where

$$\widetilde{R}(u; \mathbf{x}) = n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\frac{e_i(u)}{\boldsymbol{a}_0(u)} + n_{\mathrm{D}}^{-\frac{3}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\sum_{\iota<i}\left\{\mathbf{x} - \frac{\boldsymbol{a}_1(u)}{\boldsymbol{a}_0(u)}\right\}^{\mathsf{T}}\mathcal{U}_{\mathrm{D}i\iota} + n_{\bar{\mathrm{D}}}^{-\frac{1}{2}}\sum_{j=n_{\mathrm{D}}+1}^{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}}\left[\frac{\mathcal{P}_j(u)}{\boldsymbol{a}_0(u)} + \left\{\mathbf{x} - \frac{\boldsymbol{a}_1(u)}{\boldsymbol{a}_0(u)}\right\}^{\mathsf{T}}\mathcal{U}_{\bar{\mathrm{D}}j}\right], \qquad (4.2)$$

$e_i(u) = \sum_{k=1}^{K_i} w_{g_{ik}}(u)\left[I(U_{\mathrm{D}ik} \leqslant u) - g\left\{h_0(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\right\}\right]$, $w_{g_{ik}}(u) = w_g\{h_0(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\}$, $\boldsymbol{a}_k(u)$ and $\mathcal{P}_j(u)$ are defined in Appendix C. The equivalence between $R(u; \mathbf{x})$ and $\widetilde{R}(u; \mathbf{x})$ allows us to approximate the distribution of the process $R(u; \mathbf{x})$ using re-sampling techniques (Parzen *et al.*, 1994) in practice. In essence, one can simulate random samples $\mathcal{L} = \{L_i,\ i = 1, \ldots, n_{\mathrm{D}} + n_{\bar{\mathrm{D}}}\}$ from the standard normal distribution, and for each set of $\mathcal{L}$, obtain

$$\widehat{R}_{\mathcal{L}}(u; \mathbf{x}) = n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\frac{\widehat{e}_i(u)}{\widehat{\boldsymbol{a}}_0(u)}L_i + n_{\mathrm{D}}^{-\frac{3}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\sum_{\iota<i}\left\{\mathbf{x} - \frac{\widehat{\boldsymbol{a}}_1(u)}{\widehat{\boldsymbol{a}}_0(u)}\right\}^{\mathsf{T}}\widehat{\mathcal{U}}_{\mathrm{D}i\iota}(L_i + L_\iota) + n_{\bar{\mathrm{D}}}^{-\frac{1}{2}}\sum_{j=n_{\mathrm{D}}+1}^{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}}\left[\frac{\widehat{\mathcal{P}}_j(u)}{\widehat{\boldsymbol{a}}_0(u)} + \left\{\mathbf{x} - \frac{\widehat{\boldsymbol{a}}_1(u)}{\widehat{\boldsymbol{a}}_0(u)}\right\}^{\mathsf{T}}\widehat{\mathcal{U}}_{\bar{\mathrm{D}}j}\right]L_j, \qquad (4.3)$$

where $\widehat{e}_i$, $\widehat{\boldsymbol{a}}_k$, $\widehat{\mathcal{U}}_{\mathrm{D}i\iota}$, $\widehat{\mathcal{P}}_j$ and $\widehat{\mathcal{U}}_{\bar{\mathrm{D}}j}$ are obtained by replacing all the theoretical quantities in e_i, $\boldsymbol{a}_k$, $\mathcal{U}_{\mathrm{D}i\iota}$, $\mathcal{P}_j$ and $\mathcal{U}_{\bar{\mathrm{D}}j}$ by their empirical counterparts, respectively. Confidence bands for the ROC curve can subsequently be obtained based on the generated samples of $\{\widehat{R}_{\mathcal{L}}(u; \mathbf{x})\}$.

Table 1. *PSA data: estimated β_t and β_z and their standard errors*

Method	$\widehat{\beta_t}$ $(\widehat{\sigma}_t)$	$\widehat{\beta_z}$ $(\widehat{\sigma}_z)$
New Approach	−0.111(0.025)	−0.0001(0.008)
Cai and Pepe	−0.120(0.041)	−0.014 (0.020)
Parametric	−0.119(0.041)	−0.020 (0.019)

5. NUMERICAL STUDIES

5.1 *Example: evaluating PSA as a biomarker for prostate cancer*

PSA levels in serum have been widely used to screen men for prostate cancer. A retrospective longitudinal case control study was conducted in an effort to quantify the accuracy of PSA in distinguishing men with cancer from those without prior to onset of clinical symptoms. This study was nested in the Beta-Carotene and Retinol efficacy Trial (CARET) (Thornquist *et al.*, 1993). The trial enrolled a total of 12 025 men at high risk of lung cancer, and evaluated the efficacy of beta-carotene and retinol in preventing lung cancer. As part of the protocol, subjects had serum samples drawn at baseline and at two-year intervals thereafter. By the end of the trial, 354 subjects developed prostate cancer. All 88 of those who had at least three blood samples taken were included in the PSA case control study. Serum samples stored prior to their diagnosis with cancer were assayed for PSA, together with samples from 88 age-matched controls. A full description of the study design is available in Etzioni *et al.* (1999).

Age, denoted by z, is a factor that can potentially affect the discriminatory ability of PSA. Another factor we included is T, years between the onset of symptoms and when the serum sample was drawn. The following model was fit to the data:

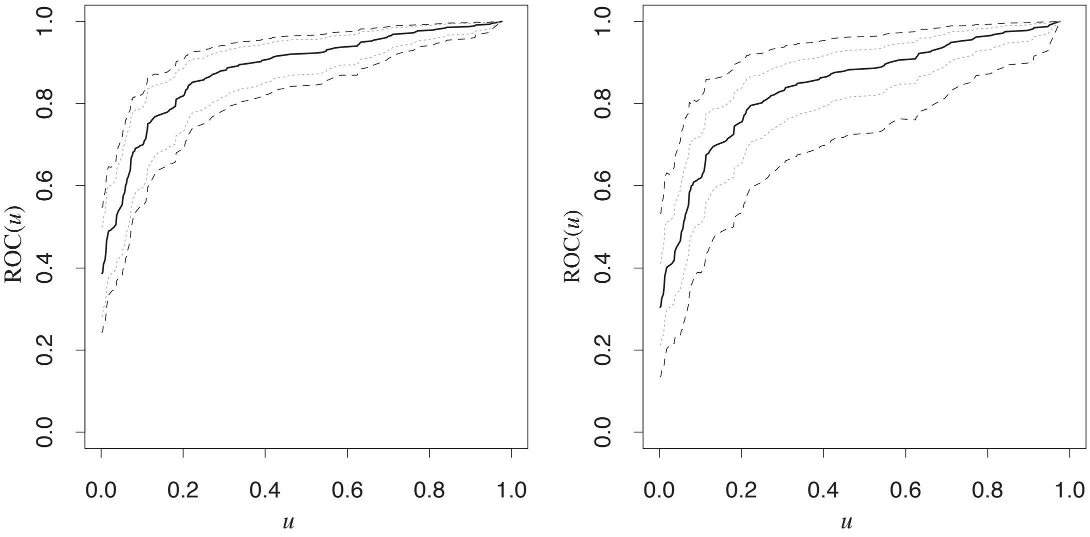
$$\text{ROC}_{T,z}(u) = \Phi\left\{\beta_t T + \beta_z z + h(u)\right\}.$$

With the proposed semi-parametric approach, the estimated β_t is -0.111 with standard error 0.025, and the estimated age effect β_z is -0.0001 with standard error 0.008. The negative coefficient for T indicates that the discrimination improves as the PSA measurement time approaches the time of diagnosis. The discriminatory capacity of PSA does not seem to vary with the subject's age. For comparison, shown in Table 1 are the estimates from the Cai and Pepe (2002) semi-parametric method and the Alonzo and Pepe (2002) parametric distribution-free approach with baseline ROC function taking the form $h(u) = \alpha_0 + \alpha_1\Phi^{-1}(u)$. The estimates of β_t are fairly close to each other, but the estimated age effect from this approach is even closer to 0. The new approach gives much smaller estimated standard errors. This is consistent with simulation studies reported below.

In Figures 1(a) and (b), we show the estimated ROC curves and their 95% point-wise confidence intervals and simultaneous confidence bands at $T = 2$ and 4 years prior to diagnosis for patients who are 60 years old when PSA is measured. These plots also indicate that PSA is more accurate at distinguishing diseased subjects from non-diseased when PSA is measured closer to diagnosis. For example, at a false positive rate 20%, the estimated true positive rate is 82% (SE $=$ 3.9%) if PSA is measured 2 years prior to diagnosis and 76% (SE $=$ 4.7%) if measured 4 years prior to diagnosis.

5.2 *Simulation studies*

We conducted simulation studies to evaluate the performance of the new procedure. First, we mimicked the PSA example by including both a covariate z common to diseased and non-diseased subjects and a



(a) $T = 2$ years

(b) $T = 4$ years

Fig. 1. Estimated ROC curves (solid curves) and their 95% point-wise (dotted curves) and simultaneous (dashed lines) confidence intervals for patients at age 60 whose PSA are measured at 2 and 4 years prior to diagnosis, based on the new approach.

diseased subject specific covariate T. In particular, data for the diseased subjects were generated from the linear model

$$Y_{\mathrm{D}ik} = 12 + \beta_t T_{ik} + (2 + \beta_z) z_i + \epsilon_{\mathrm{D}ik}\,, \qquad k = 1, \ldots, K,\ i = 1, \ldots, n_{\mathrm{D}}, \tag{5.1}$$

and data for non-diseased subjects followed the model

$$Y_{\bar{\mathrm{D}}jl} = 10 + 2z_j + \sigma_{\bar{\mathrm{D}}}\epsilon_{\bar{\mathrm{D}}jl}\,, \qquad l = 1, \ldots, K,\ j = n_{\mathrm{D}} + 1, \ldots, n_{\mathrm{D}} + n_{\bar{\mathrm{D}}}, \tag{5.2}$$

where $(\epsilon_{\mathrm{D}i1}, \ldots, \epsilon_{\mathrm{D}iK})$ and $(\epsilon_{\bar{\mathrm{D}}j1}, \ldots, \epsilon_{\bar{\mathrm{D}}jK})$ follow multivariate normal distribution with zero mean, variance 1 and covariance 0.3. Here, T and z are random variables with probability distributions that are exponential (rate = 1) and Bernoulli (probability = 0.5), respectively. This configuration induces the ROC curve

$$\mathrm{ROC}(u; T, z) = \Phi\left\{h(u) + \beta_t T + \beta_z z\right\},$$

where $h(u) = 2 + \sigma_{\bar{\mathrm{D}}}\Phi^{-1}(u)$.

For each simulated data set, we obtained point estimates of β_t and β_z with three approaches: (1) the new semi-parametric approach; (2) the Alonzo and Pepe (2002) parametric distribution-free (PDF) approach; and (3) the Cai and Pepe (2002) semi-parametric approach. For the semi-parametric approaches, we do not specify $h(\cdot)$. In contrast, for the parametric approach, we assume that $h(u) = \alpha_0 + \alpha_1\Phi^{-1}(u)$ with (α_0, α_1) unspecified. In Table 2, we present the bias and mean square error of the estimators from these methods based on 500 simulations.

The results show that the new estimator outperforms both the Cai and Pepe (2002) estimator and the parametric estimator that assume a parametric ROC-GLM model. At a sample size of 200, the empirical efficiencies of the parametric estimator and the Cai and Pepe (2002) estimator relative to the new estimator are (74%, 74%) for β_t and (50%, 49%) for β_z when there is a single measurement for each subject

Table 2. *Bias and Mean Square Error (MSE) of the estimated regression parameters based on accordingly specified parametric model (with the PDF approach) and semi-parametric models (with the new approach and the Cai and Pepe approach) when* $\sigma_{\bar{\mathrm{D}}} = 1$

	$n_{\mathrm{D}} = n_{\bar{\mathrm{D}}}$		$\beta_t = -1$			β_z=2		
			New	PDF	Cai and Pepe	New	PDF	Cai and Pepe
$K = 1$	100	Bias	0.025	−0.034	−0.034	0.052	0.118	−0.072
		MSE	0.019	0.027	0.032	0.104	0.190	0.193
	200	Bias	−0.009	0.022	−0.024	0.016	0.060	0.052
		MSE	0.011	0.016	0.016	0.053	0.109	0.112
$K = 2$	100	Bias	−0.009	−0.012	−0.014	0.012	0.038	0.030
		MSE	0.011	0.015	0.014	0.064	0.094	0.083
	200	Bias	0.007	0.001	0.013	0.012	0.036	0.026
		MSE	0.006	0.007	0.007	0.034	0.050	0.051

($K = 1$). The relative efficiencies change to (82%, 84%) for β_t and (69%, 67%) for β_z when each subject has two measurements in that setting.

To compare the ROC estimation procedures, we first take the estimated ROC curves from both the new approach and the parametric approach, then find the sample averages and their empirical 95% confidence intervals. The results presented in Figure 2 showed that both yield estimators with essentially no bias. The new semi-parametric method has comparable efficiency with the parametric method in estimating the ROC curve as shown in Figure 2(c).

The efficiency gain in estimating the regression parameter can in part be attributed to the elimination of the need to estimate h. We expect that the amount of efficiency gain varies from setting to setting. To investigate more on this, we used models (5.1) and (5.2) to simulate data, but included only covariate z by setting $\beta_t = 0$. As shown in Table 3, at sample size 200, the empirical efficiency of the parametric estimator for β_z relative to the semi-parametric estimator is 99% when $\sigma_{\bar{\mathrm{D}}} = 2$ and 58% when $\sigma_{\bar{\mathrm{D}}} = 1$.

6. Special case: $g(x) = 1 - \exp\{-\exp(x)\}$

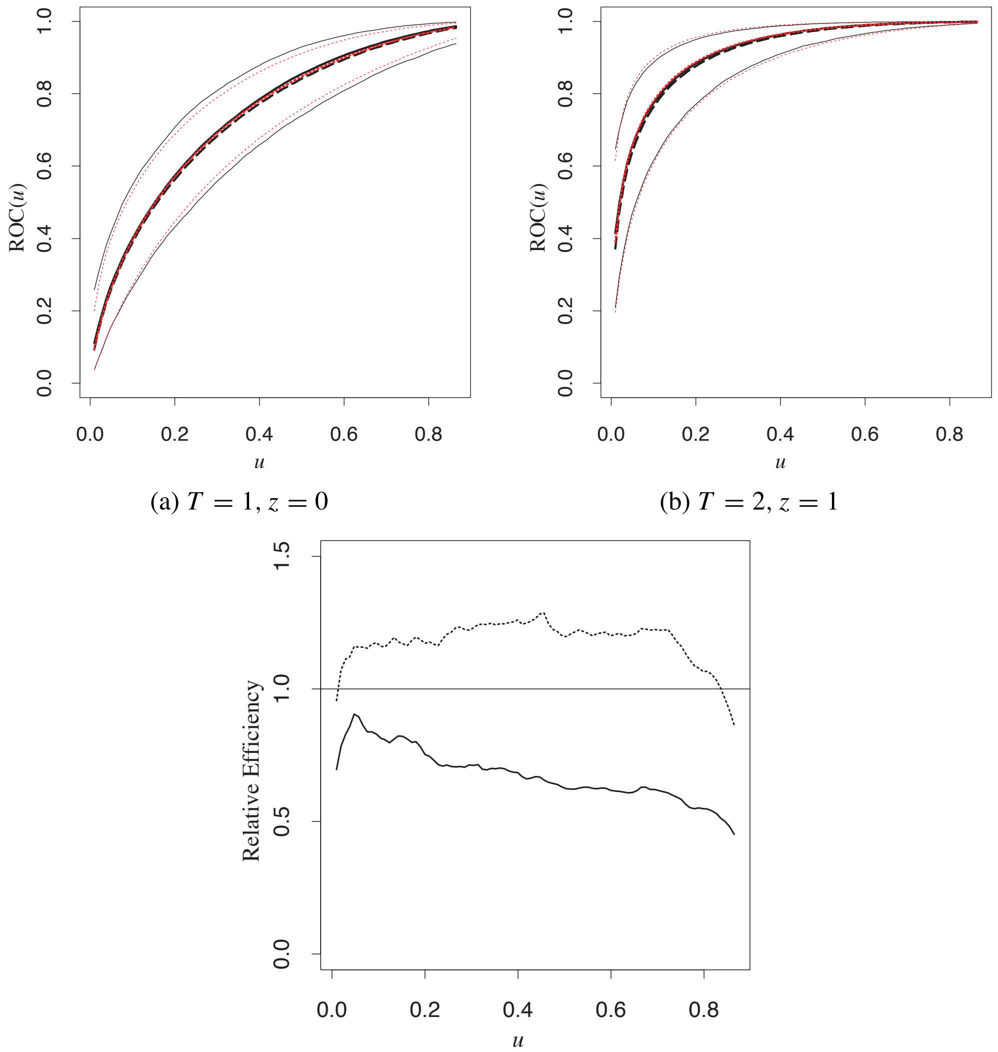
When the link function $g(x) = 1 - \exp\{-\exp(x)\}$, model (2.1) corresponds to the Cox proportional hazards model (Cox, 1972) with baseline function $h = \log \Lambda_0$, where Λ_0 is the baseline cumulative hazard function. For the Cox model, simple yet efficient estimating procedures have long been available based on partial likelihood. Inference procedures have also been generalized to accommodate clustered data (Lee *et al*., 1992). Here, we take a further step to allow for estimated response variables. We consider the marginal log partial likelihood:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n_{\mathrm{D}}} \sum_{k=1}^{K_i} \left\{ \boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}_{ik} - \sum_{\iota=1}^{n_{\mathrm{D}}} \sum_{\kappa=1}^{K_\iota} I(U_{\mathrm{D}\iota\kappa} \geqslant U_{\mathrm{D}ik}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}_{ik}) \right\}$$

as in Lee *et al*. (1992) and substitute the estimated 'event times' $\widehat{U}_{\mathrm{D}ik}$ into $\ell(\boldsymbol{\beta})$ to form a *pseudo* log partial likelihood, $\widehat{\ell}(\boldsymbol{\beta})$. Let $\widetilde{\boldsymbol{\beta}} = \operatorname{argmax}_{\beta} \widehat{\ell}(\boldsymbol{\beta})$. Given the uniform convergence of $\widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}}(\cdot)$, one can show that

$$\sup_{\beta \in D_\beta \text{ with } \beta} \left| \widehat{\ell}(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}) \right| \to 0$$

almost surely. It then follows from arguments similar to those in Appendix A that $\widetilde{\boldsymbol{\beta}}$ is a consistent estimate of $\boldsymbol{\beta}_0$.



(a) $T = 1, z = 0$

(b) $T = 2, z = 1$

(c) Relative efficiency

Fig. 2. The true (dashed lines) and sampling average of the estimated ROC curves (thicker lines) and their empirical 95% confidence intervals (thinner lines) from the semi-parametric model (solid lines) and parametric model (dotted lines). Shown also are the empirical efficiencies of the semi-parametric method relative to the parametric method for estimating the ROC curves (solid curve: $T = 1,\ z = 0$; dotted curve: $T = 2,\ z = 1$). Results are based on $K = 1$ and $n_{\mathrm{D}} = n_{\bar{\mathrm{D}}} = 200$.

The baseline function h can be estimated based on Breslow's estimator (Andersen *et al.*, 1995) for Λ_0:

$$\widetilde{h}(u) = \log \int_0^u \frac{\mathrm{d}\left\{\sum_{i=1}^{n_{\mathrm{D}}} \sum_{k=1}^{K_i} I(\widehat{U}_{\mathrm{D}ik} \leqslant v)\right\}}{\sum_{i=1}^{n_{\mathrm{D}}} \sum_{k=1}^{K_i} I(\widehat{U}_{\mathrm{D}ik} \geqslant u) \exp(\widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{Z}_{ik})}.$$

Table 3. *Bias and Mean Square Error (MSE) of the estimated β_z based on a semi-parametric model with the new approach (New) and on a parametric model with the PDF approach at sample size $n_D = n_{\bar{D}} = 200$ with $K = 1$. Here, the true value of β_z is 1*

	$\sigma_{\bar{D}} = 1$		$\sigma_{\bar{D}} = 2$	
	New	PDF	New	PDF
Bias	0.011	0.034	0.002	0.005
MSE	0.045	0.078	0.100	0.101

Table 4. *Bias and Mean Square Error (MSE) of the estimated regression parameters based on the generalized estimating equations (GEE) given in Section 3 and based on the maximum marginal partial likelihood (MMPLE)*

$n_D = n_{\bar{D}}$		$\beta_t = -1$		$\beta_z = 3$		$\beta_t = -1$		$\beta_z = 1$	
		GEE	MMPLE	GEE	MMPLE	GEE	MMPLE	GEE	MMPLE
100	Bias	−0.040	−0.021	0.092	0.058	−0.044	−0.029	0.022	0.020
	MSE	0.045	0.028	0.250	0.165	0.050	0.032	0.116	0.092
200	Bias	−0.007	−0.001	0.035	0.025	−0.012	−0.005	0.008	0.006
	MSE	0.018	0.012	0.110	0.081	0.021	0.012	0.059	0.049

The consistency of $\widetilde{h}$ follows from the consistency of $\widetilde{\boldsymbol{\beta}}$ and $\widehat{F}_{\bar{D},\mathbf{Z}}(\cdot)$. Similar arguments to those given in Appendices B and C can be used to show the asymptotic normality of $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{h}(u)$. Derivation of the large-sample distribution theory is omitted here.

One nice feature of this estimator is that it can be obtained from standard software such as `Splus` (function `coxph`) once the placement values are estimated. From our simulation studies, we also find that $\widetilde{\boldsymbol{\beta}}$ is even more efficient than the estimator proposed in Section 3 when $g(x) = 1 - \exp\{-\exp(x)\}$. For example, in one of our simulation studies, we generate data from the same models as described in Section 5.2, except that ϵ_{Dik} now follows the extreme value distribution. We let $\sigma_{\bar{D}} = 1, K = 1$. In Table 4, we show the bias and mean square error of these estimators based on 500 realizations. When $\beta_t = -1$ and $\beta_z = 1$, we find that the empirical efficiency of the estimator given in Section 3 relative to the maximum marginal partial likelihood estimator (MMPLE) is about 58% for β_t and 83% for β_z at sample size 200. The efficiency of the MMPLE is expected since the partial likelihood is the profile likelihood when the response variable is observed and the cluster size is 1.

7. REMARKS

In this paper, we study a class of semi-parametric regression models for the ROC curve. The estimator for the regression parameter proposed here does not require estimating the baseline ROC function h. Eliminating the estimation of nuisance parameters results in improved efficiency. As shown in our simulation studies, the new estimator under the semi-parametric model, while always being more robust, has efficiency that is comparable to or better than the parametric estimator from the parametric model. Asymptotic distribution theory was derived. The re-sampling method was used to acquire confidence bands for the ROC curves.

Estimation procedures for $\boldsymbol{\beta}_0$ described in Section 3 can be carried out with standard software, `Splus`

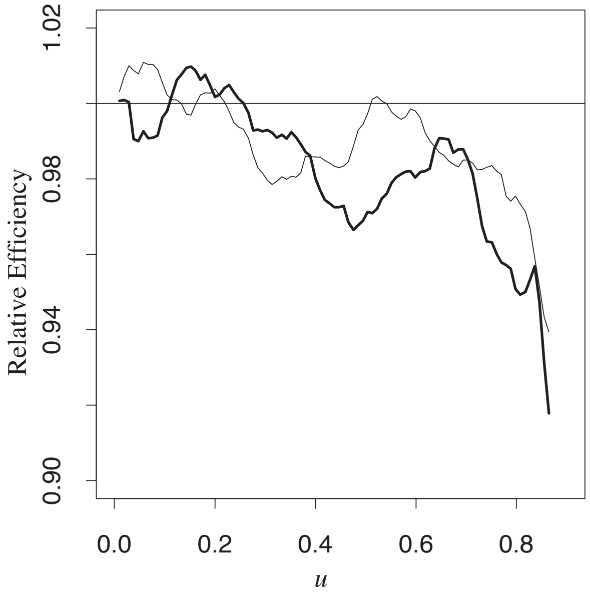


Fig. 3. Empirical efficiency of $\widehat{h}_1(\cdot)$ relative to $\widehat{h}(\cdot)$ at sample size $n_{\mathrm{D}} = n_{\bar{\mathrm{D}}} = 100$ (thinner line) and 200 (thicker line).

or SAS, for popular link functions such as probit and logit once $\widehat{U}_{\mathrm{D}ik}$ are obtained. The choice of weight function w_g for estimating h ensures the uniqueness of the solution to (4.1) and allows us to obtain the solution through standard software. Other weight functions could be used as well. For example, any positive bounded uniformly continuous weight function $v(u; \mathbf{X}_{ik})$ that does not involve the unknown parameter $h(\cdot)$ will ensure the uniqueness of the solution to (4.1). Based on our simulation studies, we find that the efficiency of the estimator for h is not very sensitive to the choice of weight functions. Figure 3 shows that the estimator for h using constant weight has comparable efficiency to $\widehat{h}(\cdot)$.

Throughout our numerical studies, we use a semi-parametric location-scale model (Heagerty and Pepe, 1999) for the non-diseased population:

$$F_{\bar{\mathrm{D}},\mathbf{Z}}(y) = F_0\left\{\frac{y - \mu(\mathbf{Z})}{\sigma(\mathbf{Z})}\right\} \tag{7.1}$$

with $\mu(\mathbf{Z}) = \gamma^{\mathsf{T}}\mathbf{Z}$, $\sigma(\mathbf{Z}) = \exp\{\boldsymbol{\eta}^{\mathsf{T}}\mathbf{Z}\}$. In practice, other models for $\mu(\mathbf{Z})$ and $\sigma(\mathbf{Z})$ can be used, for example, spline functions. Note that when the location-scale model (7.1) holds, the indicator variable $I(\widehat{U}_{\mathrm{D}ik} \leqslant \widehat{U}_{\mathrm{D}\iota\kappa})$ is equal to $I\left\{\frac{Y_{\mathrm{D}ik}-\widehat{\gamma}^{\mathsf{T}}\mathbf{Z}_{ik}}{\exp(\widehat{\boldsymbol{\eta}}^{\mathsf{T}}\mathbf{Z}_{ik})} \geqslant \frac{Y_{\mathrm{D}\iota\kappa}-\widehat{\gamma}^{\mathsf{T}}\mathbf{Z}_{\iota\kappa}}{\exp(\widehat{\boldsymbol{\eta}}^{\mathsf{T}}\mathbf{Z}_{\iota\kappa})}\right\}$ since F_0 is an increasing function. Therefore estimating $\boldsymbol{\beta}_0$ only requires knowledge of $\widehat{\gamma}$ and $\widehat{\boldsymbol{\eta}}$ from the non-diseased population and does not involve estimation of the infinite-dimensional parameter, F_0.

Often interest does not lie in the entire range of FPRs and, consequently, one may only model a portion of the ROC curve. For example, very low false positive rates such as FPR $\leqslant 0.05$ have been advocated in settings such as cancer screening (Baker and Pinsky, 2001) and hence analysis should be restricted to the portion of the curve corresponding to FPR $\leqslant 0.05$. To accommodate this, regression model (2.2) may be specified for $u \leqslant f$ with some $f < 1$. This is equivalent to assuming that (2.1) holds when $U_{\mathrm{D}ik} \leqslant f$.

We can extend our proposed method to this restricted model by artificially censoring observations $\widehat{U}_{{\rm D}ik}$ by f if $\widehat{U}_{{\rm D}ik} > f$, thus obtaining new data for analysis: $\{(\widehat{T}_{{\rm D}ik} \equiv \widehat{U}_{{\rm D}ik} \wedge f, \widehat{\Delta}_{{\rm D}ik} \equiv I(\widehat{U}_{{\rm D}ik} \leqslant f), \mathbf{X}_{ik}) : k = 1, \ldots, K_i; i = 1, \ldots, n_{\rm D}\}$. Now, letting $\alpha = h(f)$ and $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$ we can obtain an estimator of $\boldsymbol{\theta}$ by minimizing $\sum_{i,k,\iota,\kappa} \left\{\widehat{\Delta}_{{\rm D}ik} I(\widehat{T}_{{\rm D}ik} \leqslant \widehat{T}_{{\rm D}\iota\kappa}) - \eta_{{\rm D}^{ik}_{\iota\kappa}}(\boldsymbol{\theta})\right\}^2$, where $\eta_{{\rm D}^{ik}_{\iota\kappa}}(\boldsymbol{\theta}) = P(\epsilon_{ik} - \epsilon_{\iota\kappa} \leqslant \boldsymbol{\beta}^{\sf T}\mathbf{X}_{{\rm D}^{ik}_{\iota\kappa}},\ \epsilon_{ik} \leqslant \boldsymbol{\beta}^{\sf T}\mathbf{X}_{ik} + \alpha \mid \mathbf{X}_{ik}, \mathbf{X}_{\iota\kappa})$. For the special case when $g(\cdot) = \log\{-\log(\cdot)\}$, we can easily incorporate censoring into these estimation procedures developed for the Cox proportional hazards model.

ACKNOWLEDGEMENTS

The author thanks Dr Margaret Pepe and Dr Mary Lou Thompson for helpful discussions and comments. Support for this research was provided by grant 5 R37 AI024643-15 from the National Institute of Health.

APPENDIX A

Uniqueness and Consistency of $\widehat{\boldsymbol{\beta}}$

For technical reasons, we assume that potentially every diseased subject has $K_{\rm D} = \max(K_1, \ldots, K_{n_{\rm D}})$ records and the $n_{\rm D}$ sets of clustered observations $\{(Y_{\rm D}, \mathbf{Z}, \mathbf{Z}_{\rm D})\}$ are independent and identically distributed. Although not every diseased subject has $K_{\rm D}$ records, the presence or absence of individual records in a cluster does not depend on the observations. Corresponding assumptions are made for observations $\{(Y_{\bar{\rm D}}, \mathbf{Z})\}$ from non-diseased subjects.

To show the consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}$, we assume that the estimators of $F_{\bar{\rm D},\mathbf{Z}}(y)$ are uniformly consistent and $n_{\bar{\rm D}}^{\frac{1}{2}} \left\{\widehat{F}_{\bar{\rm D},\mathbf{Z}}(y) - F_{\bar{\rm D},\mathbf{Z}}(y)\right\}$ converges to a Gaussian process uniformly in y and $\mathbf{Z}$. Without loss of generality, we also assume that $n_{\bar{\rm D}}^{\frac{1}{2}} \left\{\widehat{I}_{\bar{\rm D},\mathbf{Z}}(u) - u\right\}$ can be approximated by a sum of independent terms:

$$\sup_{u,\mathbf{Z}} \left| n_{\bar{\rm D}}^{\frac{1}{2}} \left\{\widehat{I}_{\bar{\rm D},\mathbf{Z}}(u) - u\right\} - n_{\bar{\rm D}}^{-\frac{1}{2}} \sum_{j=n_{\rm D}+1}^{n_{\rm D}+n_{\bar{\rm D}}} I_{\bar{\rm D}j}(u, \mathbf{Z}) \right| \to 0 \tag{A.1}$$

in probability, where $\widehat{I}_{\bar{\rm D},\mathbf{Z}}(u) = \widehat{F}_{\bar{\rm D},\mathbf{Z}} \left\{F^{-1}_{\bar{\rm D},\mathbf{Z}}(1-u)\right\}$. The estimator based on the location-scale model of Heagerty and Pepe (1999) satisfies these conditions.

Suppose that $\boldsymbol{\beta}_0$ lies in a compact set $\Re_\beta$ and the covariate vector $\mathbf{X}$ is bounded. We also assume that $g(\cdot)$ and $h(\cdot)$ are twice continuously differentiable. Let

$$\widehat{Q}(\boldsymbol{\beta}) = -\sum_{i,k,\iota,\kappa} \left\{\widehat{\mathcal{I}}_{{\rm D}^{ik}_{\iota\kappa}} \log \xi(\boldsymbol{\beta}^{\sf T}\mathbf{X}_{{\rm D}^{ik}_{\iota\kappa}}) + (1 - \widehat{\mathcal{I}}_{{\rm D}^{ik}_{\iota\kappa}}) \log(1 - \xi(\boldsymbol{\beta}^{\sf T}\mathbf{X}_{{\rm D}^{ik}_{\iota\kappa}}))\right\},$$

where the suppressed notation $\sum_{i,k,\iota,\kappa}$ denotes $\sum_{i=1}^{n_{\rm D}} \sum_{k=1}^{K_i} \sum_{\iota<i} \sum_{\kappa=1}^{K_\iota}$ and $\widehat{\mathcal{I}}_{{\rm D}^{ik}_{\iota\kappa}} = I(\widehat{U}_{{\rm D}ik} \leqslant \widehat{U}_{{\rm D}\iota\kappa})$. Then $\widehat{U}(\boldsymbol{\beta}) = -\partial\widehat{Q}(\boldsymbol{\beta})\partial\boldsymbol{\beta}$. To show the consistency of $\widehat{\boldsymbol{\beta}}$, it is sufficient to show that $n_{\rm D}^{-2}\widehat{Q}(\boldsymbol{\beta})$ converges uniformly to a deterministic function which has a unique minimizer at $\boldsymbol{\beta}_0$ (Newey and McFadden, 1994). To this end, let $Q(\boldsymbol{\beta})$ be obtained by replacing $\widehat{\mathcal{I}}_{{\rm D}^{ik}_{\iota\kappa}}$ in $\widehat{Q}(\boldsymbol{\theta})$ with $\mathcal{I}_{{\rm D}^{ik}_{\iota\kappa}} = I(U_{{\rm D}ik} \leqslant U_{{\rm D}\iota\kappa})$. The uniform convergence of $\widehat{F}_{\bar{\rm D},\mathbf{Z}}(y)$ ensures that $n_{\rm D}^{-2}\left\{\widehat{Q}(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0)\right\} \to 0$ almost surely. Since the covariate vectors $\mathbf{X}_{ik}$ are bounded, $n_{\rm D}^{-2}\{\widehat{Q}(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0)\}$ is equicontinuous and the above convergence is uniform in $\boldsymbol{\beta} \in \Re_\beta$.

Since $n_{\mathrm{D}}^{-2}Q(\boldsymbol{\beta})$ is a U-process, it follows from Honore and Powell (1994, p. 249) that $n_{\mathrm{D}}^{-2}Q(\boldsymbol{\beta})$ converges to a deterministic function $q(\boldsymbol{\beta})$ uniformly in $\boldsymbol{\beta} \in \mathfrak{R}_{\beta}$. To show $q(\boldsymbol{\beta})$ has a unique minimizer at $\boldsymbol{\beta}_0$, consider the quantity $\lim_{n\to\infty} n_{\mathrm{D}}^{-2}\left\{Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0)\right\}$ which can be written as

$$\begin{aligned} q(\boldsymbol{\beta}) - q(\boldsymbol{\beta}_0) &= -n_{\mathrm{D}}^{-2} \lim_{n\to\infty} \sum_{i,k,\iota,\kappa} \left\{ \mathcal{I}_{\mathrm{D}^{ik}_{\iota\kappa}} \log \frac{\xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})}{\xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})} + (1 - \mathcal{I}_{\mathrm{D}^{ik}_{\iota\kappa}}) \log \frac{1 - \xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})}{1 - \xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})} \right\} \\ &= \int \left\{ \xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{x}_{12}) \log \frac{\xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{x}_{12})}{\xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{12})} + (1 - \xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{x}_{12})) \log \frac{1 - \xi(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{x}_{12})}{1 - \xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{12})} \right\} \mathrm{d}\mathcal{L}(\mathbf{x}_1)\mathcal{L}(\mathbf{x}_2) \end{aligned}$$

where $c = Um \sum_{i-1}^{n_y} \sum_{\iota<i} K_i K_\iota / n_D^2$ $\mathbf{x}_{12} = \mathbf{x}_1 - \mathbf{x}_2$, and $\mathcal{L}(\cdot)$ is the marginal distribution function of $\mathbf{X}_{11}$. Since function $f(p) = p_0 \log(p_0/p) + (1 - p_0)\log\{(1 - p_0)/(1 - p)\}$ has a unique minimizer at $p = p_0$ for any given $p_0 \in (0, 1)$, we have $q(\boldsymbol{\beta}) - q(\boldsymbol{\beta}_0) \geqslant 0$. The uniqueness of the minimizer then follows from the monotonicity of $\xi(\cdot)$.

APPENDIX B

Asymptotic distribution of $\widehat{\boldsymbol{\beta}}$

By the consistency of $\widehat{\boldsymbol{\beta}}$ and a Taylor series expansion of $\widehat{U}(\widehat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}_0$, we obtain $n_{\mathrm{D}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \widehat{\mathbb{A}}^{-1}(\boldsymbol{\beta}_0) n_{\mathrm{D}}^{-\frac{3}{2}} \widehat{U}(\boldsymbol{\beta}_0) + o_p(1)$, where $\widehat{\mathbb{A}}(\boldsymbol{\beta}) = n_{\mathrm{D}}^{-2} \sum_{i,k,\iota,\kappa} w_{\xi}^2(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}}\mathbf{X}^{\mathsf{T}}_{\mathrm{D}^{ik}_{\iota\kappa}}$. Since $\widehat{\mathbb{A}}^{-1}(\boldsymbol{\beta}_0)$ is a U-statistic (Serfling, 1980), we have

$$n_{\mathrm{D}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbb{A}^{-1} n_{\mathrm{D}}^{-\frac{3}{2}} \widehat{U}(\boldsymbol{\beta}_0) + o_p(1), \tag{B.1}$$

where $\mathbb{A}$ is the limit of $\widehat{\mathbb{A}}^{-1}(\boldsymbol{\beta}_0)$. Note that

$$n_{\mathrm{D}}^{-\frac{3}{2}} \widehat{U}(\boldsymbol{\beta}_0) = n_{\mathrm{D}}^{-\frac{3}{2}} U(\boldsymbol{\beta}_0) + n_{\mathrm{D}}^{-\frac{3}{2}} \sum_{i,k,\iota,\kappa} w_{\xi}(\boldsymbol{\beta}_0\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}} \left\{ \widehat{\mathcal{I}}_{\mathrm{D}^{ik}_{\iota\kappa}} - \mathcal{I}_{\mathrm{D}^{ik}_{\iota\kappa}} \right\} + o_p(1) \tag{B.2}$$

where $U(\boldsymbol{\beta}) = \sum_{i=1}^{n_{\mathrm{D}}} \sum_{k=1}^{K_i} \sum_{\iota<i} \sum_{\kappa=1}^{K_\iota} w_{\xi}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}}\{\mathcal{I}_{\mathrm{D}^{ik}_{\iota\kappa}} - \xi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\}$. Assume that $g(\cdot)$ is twice continuously differentiable. It follows from the Functional Limit theorems for U-processes (Nolan and Pollard, 1987, 1988) that

$$n_{\mathrm{D}}^{-\frac{3}{2}} \sum_{i,k,\iota,\kappa} w_{\xi}(\boldsymbol{\beta}_0\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}} \left\{ I(U_{\mathrm{D}ik} \leqslant U_{\mathrm{D}\iota\kappa} + \varepsilon) - P_{\mathbf{X}_{ik},\mathbf{X}_{\iota\kappa}}(\epsilon) \right\} \tag{B.3}$$

converges weakly to a Gaussian process in ε, where

$$\begin{aligned} P_{\mathbf{X}_{ik},\mathbf{X}_{\iota\kappa}}(\varepsilon) &= P(U_{\mathrm{D}ik} \leqslant U_{\mathrm{D}\iota\kappa} + \varepsilon \mid \mathbf{X}_{ik}, \mathbf{X}_{\iota\kappa}) \\ &= \int g\left\{\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik} + h_0(u + \varepsilon)\right\} \mathrm{d}g\left\{h_0(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{\iota\kappa}\right\}. \end{aligned}$$

The convergence ensures the equicontinuity of (B.3) in ϵ (Nolan and Pollard, 1988). It then follows from (A.1) and a Taylor series expansion that

$$n_{\mathrm{D}}^{-\frac{3}{2}} \sum_{i,k,\iota,\kappa} w_{\xi}(\boldsymbol{\beta}_0\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}})\mathbf{X}_{\mathrm{D}^{ik}_{\iota\kappa}} \left(\widehat{\mathcal{I}}_{\mathrm{D}^{ik}_{\iota\kappa}} - \mathcal{I}_{\mathrm{D}^{ik}_{\iota\kappa}} \right) = n_{\bar{\mathrm{D}}}^{-\frac{1}{2}} \sum_{j=n_{\mathrm{D}}+1}^{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}} \int \mathbf{W}_{\bar{\mathrm{D}}j}(u)\,\mathrm{d}u \tag{B.4}$$

where $\mathbf{W}_{\bar{\mathrm{D}}j}(u)$ is the limit of

$$n_{\mathrm{D}}^{-2} p_{10}^{\frac{1}{2}} \sum_{i,k,\iota,\kappa} w_{\xi}(\boldsymbol{\beta}_0 \mathbf{X}_{\mathrm{D}_{\iota\kappa}^{ik}}) \mathbf{X}_{\mathrm{D}_{\iota\kappa}^{ik}} f_{ik}(u) f_{\iota\kappa}(u) \left\{ I_{\bar{\mathrm{D}}j}(u; \mathbf{Z}_{ik}) - I_{\bar{\mathrm{D}}j}(u; \mathbf{Z}_{\iota\kappa}) \right\},$$

$p_{10} = n_{\mathrm{D}}/n_{\bar{\mathrm{D}}}$, and $f_{ik}(u) = \dot{g}\left\{h_0(u) + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{ik}\right\} \dot{h}_0(u)$ This, coupled with (B.1), (B.2) and (B.4), entails that

$$n_{\mathrm{D}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approx n_{\mathrm{D}}^{-\frac{3}{2}} \sum_{i=1}^{n_{\mathrm{D}}} \sum_{\iota<i} \mathcal{U}_{\mathrm{D}i\iota} + n_{\bar{\mathrm{D}}}^{-\frac{1}{2}} \sum_{j=n_{\mathrm{D}}+1}^{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}} \mathcal{U}_{\bar{\mathrm{D}}j}. \tag{B.5}$$

It then follows from the properties of U-statistics (Serfling, 1980) that the distribution of $n_{\mathrm{D}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ can be approximated by a normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

APPENDIX C

Large-sample properties for $\widehat{h}(t)$ and $\widehat{ROC}_{\mathbf{x}}(t)$

To show that $R(u; \mathbf{x}) = n_{\mathrm{D}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}}\mathbf{x} + n_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{h}(u) - h_0(u)\right\}$ converges to a Gaussian process, it remains to show the large-sample property of $n_{\mathrm{D}}^{-\frac{1}{2}}\left\{\widehat{h}(u) - h_0(u)\right\}$. First, the same argument as in Appendix C of Cai and Pepe (2002) gives that

$$n_{\mathrm{D}}^{-\frac{1}{2}} \sum_{i=1}^{n_{\mathrm{D}}} \sum_{k=1}^{K_i} w_{g_{ik}}(u) \left[I(\widehat{U}_{\mathrm{D}ik} \leqslant u) - g\left\{h_0(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\right\}\right]$$

is asymptotically equivalent to $n_{\mathrm{D}}^{-\frac{1}{2}} \sum_{i=1}^{n_{\mathrm{D}}} e_i(u) + n_{\bar{\mathrm{D}}}^{-\frac{1}{2}} \sum_{j=n_{\mathrm{D}}+1}^{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}} \mathcal{P}_j(u)$, where $\mathcal{P}_j(u)$ is the limit of

$$\widehat{\mathcal{P}}_j(u) = p_{10}^{\frac{1}{2}} n_{\mathrm{D}}^{-1} \sum_{i=1}^{n_{\mathrm{D}}} \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} w_{g_{ik}}(u) f_{ik}(u) \left\{ I_{\bar{\mathrm{D}}j}(u, \mathbf{Z}_{ik}) - u \right\}.$$

It then follows from Appendix B, (B.5) and a Taylor series expansion of (4.1) around $h_0(u)$ that

$$\begin{aligned} n_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{h}(u) - h_0(u)\right\} &\approx \frac{n_{\mathrm{D}}^{-\frac{1}{2}}}{\boldsymbol{a}_0(u)} \sum_{i=1}^{n_{\mathrm{D}}} \sum_{k=1}^{K_i} \left[I\left\{U_{\mathrm{D}ik} \leqslant 1 - \widehat{I}_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}^{-1}(u)\right\} - g\left\{h_0(u) + \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_{ik}\right\}\right] \\ &\approx n_{\mathrm{D}}^{-\frac{1}{2}} \sum_{i=1}^{n_{\mathrm{D}}} \frac{e_i(u)}{\boldsymbol{a}_0(u)} - n_{\mathrm{D}}^{-\frac{3}{2}} \sum_{i=1}^{n_{\mathrm{D}}} \sum_{\iota<i} \frac{\boldsymbol{a}_1^{\mathsf{T}}(u)}{\boldsymbol{a}_0(u)} \mathcal{U}_{\mathrm{D}i\iota} + n_{\bar{\mathrm{D}}}^{-\frac{1}{2}} \sum_{j=n_{\mathrm{D}}+1}^{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}} \left\{ \frac{\mathcal{P}_j(u)}{\boldsymbol{a}_0(u)} - \frac{\boldsymbol{a}_1^{\mathsf{T}}(u)}{\boldsymbol{a}_0(u)} \mathcal{U}_{\bar{\mathrm{D}}j} \right\}, \end{aligned}$$

where $\boldsymbol{a}_k(u)$ is the limit of $n_{\mathrm{D}}^{-1} \sum_{i=1}^{n_{\mathrm{D}}} \sum_{k=1}^{K_i} w_{g_{ik}}(u) \dot{g}\left\{h_0(u) + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{ik}\right\} \mathbf{X}_{ik}^{\otimes k}$, and for any vector $\mathbf{x}$, $\mathbf{x}^{\otimes 0} = 1$, and $\mathbf{x}^{\otimes 1} = \mathbf{x}$. This, combined with (B.5) implies that $R(u; \mathbf{x})$ is asymptotically equivalent to $\widetilde{R}(u; \mathbf{x})$. It follows from the Functional Limit theorem in Nolan and Pollard (1988) that $\{\widetilde{R}(u; \mathbf{x}),\ u \in [a, b]\}$ converges to a Gaussian process.

References

ALONZO, T. A. AND PEPE, M. S. (2002). Distribution-free analysis using binary regression techniques. *Biostatistics* **3**, 421–432.

ANDERSEN, P. K., BORGAN, Ø, GILL, R. D. AND KEIDING, N. (1995). *Statistical Models Based on Counting Processes*. Berlin: Springer.

BAKER, S. G. AND PINSKY, P. F. (2001). A proposed design and analysis for comparing digital and analog mammography: Special receiver operating characteristic methods for cancer screening. *Journal of the American Statistical Association* **96**, 421–428.

BEGG, C. B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine* **10**, 1887–1895.

BRESLOW, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

CAI, T. AND PEPE, M. S. (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* **97**, 1099–1107.

CAI, T., WEI, L. J. AND WILCOX, M. (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika* **87**, 867–878.

CHENG, S. C., WEI, L. J. AND YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.

CHENG, S. C., WEI, L. J. AND YING, Z. (1997). Prediction of survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association* **92**, 227–235.

COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **34**, 187–220.

DABROWSKA, D. AND DOKSUM, K. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* **15**, 1–23.

ETZIONI, R., PEPE, M., LONGTON, G., HU, C. AND GOODMAN, G (1999). Incorporating the time dimension in Receiver Operating Characteristic curves: A case study of prostate cancer. *Medical Decision Making* **19**, 242–251.

GATSONIS, C. A., BEGG, C. B. AND WIEAND, S. A. (1995). Advances in statistical methods for diagnostic radiology: Introduction to a symposium. *Acadamic Radiology* **2**, S1–S3.

HAN, A. K. (1987). A non-parametric analysis of transformations. *Journal of Econometrics* **35**, 191–209.

HANLEY, H. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Clinical Reviews in Diagnostic Imaging* **29**, 307–335.

HEAGERTY, P. J. AND PEPE, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics* **48**, 533–551.

HONERE, B. E. AND POWELL, J. L. (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics* **64**, 241–278.

LEE, E. W., WEI, L. J. AND AMATO, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations (disc:P247). *Survival Analysis: State of the Art*, Dordrecht: Kluwer, pp. 237–247.

METZ, C. E. (1986). ROC methodology in radiologic imaging. *Investigative Radiology* **21**, 720–733.

NEWEY, W. AND MCFADDEN, D. (1994). Estimation in large samples. In McFadden, D. and Engler, R. (eds), *Handbook of Econometrics*, Vol. 4. Amsterdam: North-Holland.

NOLAN, D. AND POLLARD, D. (1987). u-processes: rates of convergence. *The Annals of Statistics* **15**, 780–799.

NOLAN, D. AND POLLARD, D. (1988). Functional limit theorems for U-processes. *The Annals of Probability* **16**, 1291–1298.

Parzen, M. I., Wei, L. J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341–350.

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

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **56**, 352–359.

Pepe, M. S. and Cai, T. (2002). The analysis of placement values for evaluating discriminatory measures. *Technical Report*, University of Washington.

Scharfstein, D. O., Tsiatis, A. A. and Gilbert, P. (1998). Semi-parametric efficient estimation in the generalized odds-rates class of regression models for right-censored time to event data. *Lifetime Data Analysis* **4**, 355–391.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic.

Thornquist, M. D., Omenn, G. S., Goodman, G. E., *et al*. (1993). Statistical design and monitoring of the carotene and retinol efficacy trial. *Controlled Clinical Trials* **14**, 308–324.

Toledano, A. and Gatsonis, C. (1995). Regression analysis of correlated receiver operating characteristic data. *Acadamic Radiology* **2**, 530–536.

Tosteson, A. N. and Begg, C. B. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making* **8**, 204–215.

[*Received April 17, 2003; accepted for publication June 30, 2003*]