Statistica Sinica **23** (2013), 829-851
doi:http://dx.doi.org/10.5705/ss.2011.279

SEMIPARAMETRIC ROC ANALYSIS USING ACCELERATED REGRESSION MODELS

Eunhee Kim and Donglin Zeng

Brown University and University of North Carolina at Chapel Hill

Abstract: The Receiver Operating Characteristic (ROC) curve is a widely used measure to assess the diagnostic accuracy of biomarkers for diseases. Biomarker tests can be affected by subject characteristics, the experience of testers, or the environment in which tests are carried out, so it is important to understand and determine the conditions for evaluating biomarkers. In this paper, we focus on assessing the effects of covariates on the performance of the ROC curves. In particular, we develop an accelerated ROC model by assuming that the effect of covariates relates to rescaling a baseline ROC curve. The proposed model generalizes the accelerated failure time model in the survival context to ROC analysis. An innovative method is developed to construct estimation and inference for model parameters. The obtained parameter estimators are shown to be asymptotically normal. We demonstrate the proposed method via a number of simulation studies, and apply it to analyze data from a prostate cancer study.

Key words and phrases: Accelerated failure time model, asymptotic normality, receiver operating characteristic curve, regression models.

1. Introduction

In medical studies, noninvasive and accurate biomarkers are widely used for evaluating patients' disease status or their responses to treatments. Examples include the use of prostate-specific antigen and CA-125 to detect the presence of prostate cancer and ovarian cancer, respectively. To assess the accuracy of biomarkers for diagnosis and prognosis of disease, one of the most popular tools is the analysis of the Receiver Operating Characteristic (ROC) curve (Swets and Pickett (1982); Hanley (1989)). The definition of a ROC curve is as follows: let Y_1 denote the biomarker for a diseased subject and Y_0 denote the biomarker for a non-diseased subject. For any threshold value c for which any test results greater than c are considered to be positive, the true positive and false positive rates are defined as $S_1(c) = P(Y_1 \geq c)$ and $S_0(c) = P(Y_0 \geq c)$. The ROC curve is defined as the plot of the true positive rate versus the false positive rate, $(S_0(c), S_1(c))$, when the threshold value c varies from $-\infty$ to ∞. Equivalently, the ROC curve is a function

$$ROC(t) = S_1(S_0^{-1}(t)), \quad t \in (0, 1),$$

where S_0^{-1} denotes the inverse of S_0.

In practice, the diagnostic performance of biomarkers can vary under different conditions. They may be accurate for predicting diseases for certain patients, but may not perform well for others. Biomarker performance may also depend on the particular conditions under which biomarker tests are carried out, including the level of experience of the tester. In order to evaluate the diagnostic performance of biomarkers, it is important to understand how the performance depends on patient characteristics or test conditions.

In the existing literature, three approaches to incorporate covariate effects into the ROC analysis have been suggested (c.f., Pepe (1998, 2003); Zhou, Obuchowski, and McClish (2002)). The first approach is to model the ROC curve summary indices as a function of covariates. Particularly, Dorfman, Berbaum, and Metz (1992) and Obuchowski and Rockette (1995) suggested modeling the area under the curve (AUC), while Thompson and Zucchini (1989) recommended modeling the partial area under the curve (pAUC). This approach is feasible only when covariates are discrete, and there are enough patients in each covariate combination to permit the reliable calculation of the summary accuracy measure. The second approach is to model the distributions of test results as a function of disease status and covariates. Tosteson and Begg (1998) described the use of an ordinal regression model to induce the regression models for the ROC curve for tests with ordinal outcomes. Their method has been extended to random effects models (Beam (1995); Gatsonis (1995)) and Bayesian methods (Peng and Hall (1996); Hellmich et al. (1998); Ishwaran and Gatsonis (2000)). However, in this approach, the parameter estimates do not reflect the covariate effects on the ROC curve, so it is difficult to examine how the ROC curves can vary over different covariates. Instead, the third approach directly models covariate effects on the ROC curve (Pepe (1997, 2000); Alonzo and Pepe (2002)). Sometimes, this approach is called a parametric distribution free approach since it only assumes a parametric model for the ROC curve, but is distribution-free regarding the distribution of the test results. The most important advantage of this approach is that the interpretation of model parameters pertains directly to the covariate effects on the ROC curves. Specifically, in this approach, Pepe (1997, 2000) proposed parametric ROC regression models of the generalized linear model (GLM) form by assuming,

$$ROC_X(t) = g(h(t) + \beta^T X), \quad t \in (0,1), \tag{1.1}$$

where $ROC_X(t)$ denotes the ROC curve at a false positive rate t associated with covariates X, $g(\cdot)$ is a known link function, and $h(\cdot)$ is a baseline function specified up to some finite parameters. Here, the baseline function h defines the location and shape of the ROC curve, and β quantifies covariate effects.

Pepe used the estimating equations for β based on the binary indicator variable $I(Y_1 \geq S_0^{-1}(t|X))$. Later, Cai and Pepe (2002) extended this parametric ROC regression model to a semiparametric approach by allowing an arbitrary nonparametric baseline function for h. They assumed a semiparametric location model for $S_0(y|X)$ (Pepe (1998); Heagerty and Pepe (1999)), and constructed high-dimensional estimating equations for estimating β and h. We emphasize that the last two models both assume that the effects of covariates are related to the location shift of a baseline ROC curve. This may not be true in some situations.

In this article, we develop an alternative regression model, namely the accelerated ROC model, by adjusting for covariates that can influence the performance of biomarkers. We consider modeling covariates directly on the ROC curve and our model generalizes the usual accelerated failure time model (Kalbfleisch and Prentice (2002)) in the survival context to the ROC analysis. A practical advantage of the proposed approach is that the estimation for the regression parameters only requires solving a small number of equations compared to the estimation techniques by Cai and Pepe (2002). In Section 2, we describe an accelerated ROC model and the procedures for estimating parameters of covariates β as well as the ROC function. The asymptotic properties of β and the ROC function are given in Section 3, and simulation studies are provided in Section 4. As an example, we apply our method to a prostate cancer dataset in Section 5, and a discussion is given in Section 6. All technical proofs are given in the Appendix.

2. Models and Inference Procedure

To model the covariate effects on the ROC curve, we propose the semiparametric ROC model

$$ROC_X(t) = G(e^{\beta^T X} \log t), \quad t \in (0,1), \tag{2.1}$$

where $ROC_X(t)$ denotes the ROC curve at a false positive rate t associated with covariates X and $G(\cdot)$ is an unknown and increasing function satisfying $G(-\infty) = 0$ and $G(0) = 1$, this is because $ROC_X(1) = 1$ and $ROC_X(0) = 0$ for any fixed value X. It is noted that (2.1) becomes one ROC-GLM model if a link function g in (1.1) takes $G(\exp(t))$ with a baseline function $h(t)$ equal to $\log(\log(t))$.

In model (2.1), a negative value for β indicates that discrimination improves as X increases since $\log(t), t \in (0,1)$ is negative and $G(\cdot)$ is an increasing function. For example, if $G(t) = \exp(\alpha t)$, then $ROC_X(t) = \exp\{\alpha e^{\beta^T X} \log t\},\ 0 < t < 1$. If $\alpha = 1$ and $\beta = 0$, the ROC curve is the 45 degree line indicating that a biomarker has no discriminatory ability, while if $0 < \alpha < 1$ and $\beta = 0$, the ROC curve is

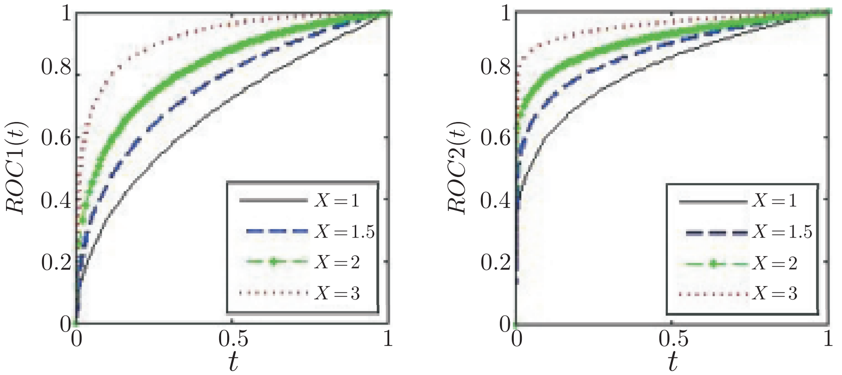

Figure 1. (Left) Parametric ROC-GLM, $ROC1(t) = \Phi(0.6X + 0.8\Phi^{-1}(t))$; (Right) Accelerated ROC model, $ROC2(t) = \exp(0.5\mathrm{e}^{-0.8X} \log(t))$

above the diagonal line, and a biomarker is considered to have reasonable discriminatory ability to diagnose patients with and without the disease. If $\alpha = 0.5$ and $\beta = -0.8$, then $ROC_X(t) = \exp\{0.5e^{-0.8X} \log t\}$; in this case, discrimination improves as X increases, see Figure 1.

The parameter β characterizes the shape of the ROC curve for X, where the effects of X in the proposed ROC model relate to rescaling a baseline ROC curve. To see how this is different from the parametric ROC regression model in Pepe (1997, 2000), we plot the ROC curves using the two models in Figure 1. Clearly, the covariate affects true positive rates more dramatically for low false positive rates based on our model.

Suppose we observe n_1 biomarker measurements from diseased subjects and n_0 biomarker measurements from non-diseased subjects. Let Y_{i1} $(i = 1, \ldots, n_1)$ denote the biomarker measurement for diseased subject i and Y_{j0} $(j = 1, \ldots, n_0)$ denote the biomarker measurement for non-diseased subject j. We assume that each subject may have one or more covariates and denote them as X_{i1} and X_{j0} for diseased subject i and non-diseased subject j, respectively. In many applications, the measurements of a biomarker are subject to a finite upper detection limit, denoted by τ, where the test results above τ are not quantifiable and are considered to be censored. Thus, the observed data can be represented as

$$\{(\min(Y_{i1} \wedge \tau), X_{i1}, \Delta_{i1}),\ i = 1, \ldots, n_1\}$$

for diseased subjects and

$$\{(\min(Y_{j0} \wedge \tau), X_{j0}, \Delta_{j0}),\ j = 1, \ldots, n_0\}$$

for non-diseased subjects, where $\Delta_{i1} = I(Y_{i1} \leq \tau)$ and $\Delta_{j0} = I(Y_{j0} \leq \tau)$.

By the definition of $ROC_X(t)$, the model (2.1) can be rewritten as

$$S_1(t|X) = G(e^{\beta^T X} \log S_0(t|X)). \tag{2.2}$$

It is to be noted that we make no assumptions on the model for $S_0(t|X)$. To estimate β, we take $Z_{i1} = -\log S_0(Y_{i1}|X_{i1})$. Using (2.2), it can be shown that

$$\begin{aligned} P(Z_{i1} \leq z|X_{i1}) &= 1 - P(-\log S_0(Y_{i1}) > z|X_{i1}) = 1 - G(e^{\beta^T X_{i1}} \log e^{-z}) \\ &= 1 - G(-ze^{\beta^T X_{i1}}) \equiv F(ze^{\beta^T X_{i1}}), \end{aligned}$$

with $F(x) = 1 - G(-x)$. Hence, Z_{i1} satisfies the accelerated failure time (AFT) model, so inference for β can be conducted by solving the log-rank estimating equation, that is commonly used for the estimation in the AFT model. Specifically, the log-rank estimating equation is

$$\sum_{i=1}^{n_1} \Delta_{i1} \left\{ X_{i1} - \frac{\sum_j I(\log Z_{j1} + \beta^T X_{j1} \geq \log Z_{i1} + \beta^T X_{i1}) X_{j1}}{\sum_j I(\log Z_{j1} + \beta^T X_{j1} \geq \log Z_{i1} + \beta^T X_{i1})} \right\} = 0. \tag{2.3}$$

Alternatively, other methods, such as the the Gehan-rank estimation equation (Jin et al. (2003)) or the nonparametric maximum likelihood estimation (Zeng and Lin (2007)) can be applied. Because of the AFT model implication on the $Z_{i1} = -\log S_0(Y_1|X)$, we call the proposed ROC function (2.1) the accelerated ROC model.

Since S_0 is unknown, we estimate S_0 nonparametrically using the smoothed Breslow estimator as follows:

$$\hat{S}_0(y|x) = \exp\left\{ -\sum_{j=1}^{n_0} I(Y_{j0} \leq y) \frac{\triangle_{j0} K_{a_n}(X_{j0} - x)}{\sum_{k=1}^{n_0} I(Y_{k0} \geq Y_{j0}) K_{a_n}(X_{k0} - x)} \right\}, \tag{2.4}$$

where $K_{a_n}(x) = K(x/a_n)/a_n^d$ with a_n the bandwidth and d the dimension of X. Alternatively, one may use the Kaplan-Meier type estimator.

We suggest using an optimal bandwidth selection method in Wang and Shen (2008) for a_n in (2.4). First, we obtain a smoothed Breslow estimator $\hat{S}(y|x, a_n)$ using a reasonable initial bandwidth a_n. We also generate repeatedly B bootstrap samples $\{(x_{i0}^*, y_{i0}^*), i = 1, \ldots, n_0\}$ from $\{(x_{i0}, y_{i0}), i = 1, \ldots, n_0\}$ and get bootstrapped smoothed Breslow estimators,

$$\hat{S}_0^b(y|x, a_n^*) = \exp\left\{ -\sum_{j=1}^{n_0} I(Y_{j0}^* \leq y) \frac{K_{a_n^*}(X_{j0}^* - x)}{\sum_{k=1}^{n_0} I(Y_{k0}^* \geq Y_{j0}^*) K_{a_n^*}(X_{k0}^* - x)} \right\}, b = 1, \ldots, B.$$

Then, we select $\hat{a}_n$ by minimizing the bootstrapped mean integrated squared error (MISE)

$$\frac{1}{B} \sum_{b=1}^{B} \int \left(\hat{S}_0^b(y|x, a_n^*) - \hat{S}_0(y|x, a_n) \right)^2 dy \tag{2.5}$$

over possible bandwidths a_n^*. If the difference between $\hat{a}_n$ and the initial value a_n is small enough, the process stops and the optimal bandwidth is set to $\hat{a}_n$. Otherwise, we replace the initial bandwidth and repeat similar procedures until the process converges to an optimal value.

Z_{i1} is estimated by $\hat{Z}_{i1} = -\log \hat{S}_0(Y_{i1}|X_{i1})$ using (2.4) and, after plugging $\hat{Z}_{i1}$ into (2.3), $\hat{\beta}$ is calculated by solving

$$\sum_{i=1}^{n_1} \Delta_{i1}\Big\{X_{i1} - \frac{\sum_j I(\log \hat{Z}_{j1} + \hat{\beta}^T X_{j1} \geq \log \hat{Z}_{i1} + \hat{\beta}^T X_{i1})X_{j1}}{\sum_j I(\log \hat{Z}_{j1} + \hat{\beta}^T X_{j1} \geq \log \hat{Z}_{i1} + \hat{\beta}^T X_{i1})}\Big\} = 0.$$

Remark 1. When X is discrete, the estimator for S_0, $\hat{S}_0(y|x)$ in (2.4) can be replaced by the Breslow estimator using the data with $X_{j0} = x$. i.e.,

$$\hat{S}_0(y|x) = \exp\Big\{ -\sum_{j=1}^{n_0} I(Y_{j0} \leq y)\frac{\triangle_{j0} I(X_{j0} = x)}{\sum_{k=1}^{n_0} I(Y_{k0} \geq Y_{j0}) I(X_{k0} = x)}\Big\}.$$

Remark 2. When X has more than one continuous covariate, the kernel estimate $\hat{S}_0$ may not perform well with a moderate sample size. In this case, we suggest estimating $S_0(y|x)$ based on the Cox regression model using the non-diseased data. That is,

$$\hat{S}_0(y|x) = \exp\Big[-\hat{\Lambda}(y)\exp(\hat{\gamma}^T x)\Big],$$

where $\hat{\Lambda}(y)$ is the estimate of the cumulative baseline function and $\hat{\gamma}$ is the regression parameter estimate. An alternative approach is to use the single index model, which is more flexible than the Cox regression model. The estimators from the latter model, however, can be computed easily.

We next describe the procedures for estimating G and the ROC function specified in (2.1). Clearly, $P(Z_{i1}e^{\beta^T X_{i1}} \leq z|X_{i1}) = 1 - G(-z)$. Therefore, $Z_{i1}e^{\beta^T X_{i1}}$ is independent of X_{i1} and has distribution function $1 - G(-z)$. This implies that we can estimate G consistently by using the empirical distribution of $W_{i1} \equiv Z_{i1}e^{\beta^T X_{i1}}$. In light of possible upper limit detection in practice, we specifically use the Kaplan-Meier estimator to estimate the survival function of W_{i1}. After replacing W_{i1} with its estimate

$$\hat{W}_{i1} = -e^{\hat{\beta}^T X_{i1}} \log \hat{S}_0(Y_{i1}|X_{i1}), \quad i = 1, \ldots, n_1,$$

we estimate $G(\cdot)$ using

$$\hat{G}(t) = \prod_{i=1}^{n_1}\Big[1 - \frac{\Delta_{i1} I(\hat{W}_{i1} \leq -t)}{\sum_{j=1}^{n_1} I(\hat{W}_{j1} \geq \hat{W}_{i1})}\Big]. \tag{2.6}$$

Finally, the ROC curve for fixed covariates X is estimated by

$$\widehat{ROC}_X(t) = \widehat{G}(e^{\hat{\beta}^T X} \log t), \quad t \in (0,1), \tag{2.7}$$

and the corresponding AUC estimate is

$$\widehat{AUC}_X = \int_0^1 \widehat{ROC}_X(t)dt = \int_0^1 \widehat{G}(e^{\hat{\beta}^T X} \log t)dt, \tag{2.8}$$

which can be calculated via the trapezoidal numerical integration.

Although the asymptotic variance of $\hat{\beta}$ has an analytic expression (see the Appendix), directly estimating its variance involves estimating some derivatives and can be computationally tedious. Thus, we propose to estimate the variances of $\hat{\beta}$ and $\hat{G}$ using the bootstrap method in order to make inferences. Bootstrap samples are drawn repeatedly with replacement from the dataset, and β and G are estimated for each bootstrap sample. We then use the variances of these $\hat{\beta}$'s and $\hat{G}$'s as our estimates.

The confidence region for $\widehat{ROC}_X(t)$ in (2.7) can be calculated in the following manner. For $0 < \xi < 1$ and $0 \leq a < b \leq 1$, we first find C_ξ such that

$$Pr\left\{ \sup_{t \in (a,b)} \frac{|\hat{G}(e^{\hat{\beta}^T X} \log t)|}{S_G(e^{\hat{\beta}^T X} \log t)} \leq C_\xi \right\} = 1 - \xi,$$

where $S_G(e^{\hat{\beta}^T X} \log t)$ is the estimated standard deviation of $\hat{G}(e^{\hat{\beta}^T X} \log t)$. Then, a $100(1-\xi)\%$ confidence region for $ROC_X(t)$ over $[a,b]$ is

$$\hat{G}(e^{\hat{\beta}^T X} \log t) \pm C_\xi n_1^{-1/2} S_G(e^{\hat{\beta}^T X} \log t).$$

Specifically, we generate K (e.g. $K = 500$) samples consisting of biomarker measurements and corresponding covariates in the diseased and nondiseased groups and compute $\hat{G}_k$ $(k = 1, \ldots, K)$ for each sample. Then, $S_G(e^{\hat{\beta}^T X} \log t)$ can be calculated by the sample standard deviation of $\hat{G}_k$'s, and C_ξ can be computed as the 100(1-ξ)% percentile of the $\sup_{t \in (a,b)} \dfrac{|\hat{G}_k(e^{\hat{\beta}^T X} \log t)|}{S_G(e^{\hat{\beta}^T X} \log t)}$, $k = 1, \ldots, K$.

Remark 3. The proposed approach can be generalized to handle the situation in which each subject may have multiple or repeated biomarkers. We assume the marginal ROC model of such multivariate biomarkers. In this case, the estimating equation for β is replaced by

$$\sum_{i=1}^{n_1}\sum_{k=1}^{n_{i1}} \Delta_{ik1} \left\{ X_{ik1} - \frac{\sum_j \sum_{l=1}^{n_{j1}} I(\log \hat{Z}_{jl1} + \hat{\beta}^T X_{jl1} \geq \log \hat{Z}_{ik1} + \hat{\beta}^T X_{ik1}) X_{jl1}}{\sum_j \sum_{l=1}^{n_{j1}} I(\log \hat{Z}_{jl1} + \hat{\beta}^T X_{jl1} \geq \log \hat{Z}_{ik1} + \hat{\beta}^T X_{ik1})} \right\} = 0,$$

where Δ_{ij1}, $\hat{Z}_{ij1}$, and X_{ij1} are the observations of jth measurement for subject i in the diseased group, and $\hat{Z}_{ij1}$ can be estimated similarly as $\hat{Z}_{i1}$. The bootstrapping method can still be used for inference by randomly selecting subjects for each bootstrap sample.

Remark 4. We suggest using the following procedure to check model adequacy. First, we stratify the data based on covariates X to obtain L groups. Let Y_{i1}^l $(i = 1, \ldots, n_1^l)$ be the biomarker measurement for diseased subject i, and Y_{j0}^l $(j = 1, \ldots, n_0^l)$ be the biomarker measurement for non-diseased subject j in group l. Next, we compare the empirical ROC curve with the proposed ROC curve $\widehat{ROC}_X(t) = G(e^{\hat{\beta}} X \log t)$ in each group, where X takes the mean value from the group. A good-fitting ROC model is reasonably consistent with the empirical ROC curve. Finally, we check if the proposed AUC estimate (2.8) is close to the empirical AUC within each stratum

$$\widehat{AUC}_{Group=l} = \sum_{j=1}^{n_0^l} \sum_{i=1}^{n_1^l} \Big\{ I(Y_{i1}^l > Y_{j0}^l) + \frac{1}{2} I(Y_{i1}^l = Y_{j0}^l) \Big\} (n_1^l n_0^l)^{-1}, \quad l = 1, \ldots, L.$$

3. Asymptotic Properties

In this section, we derive the asymptotic properties of $\hat{\beta}$ and $\hat{G}$. Consider the following conditions.

(C.1) The true parameter value, β_0, belongs to a compact set $\mathcal{B}$.

(C.2) The true densities with respective to a dominating measure for (Y_1, X_1) and (Y_0, X_0) are $(\chi + 1)$-continuously differentiable, where $\chi > d/2$ with d the dimension of X_0. Additionally, X_1 and X_0 have bounded support.

(C.3) The matrix $[1, X_1]$ is linearly independent with positive probability.

(C.4) The kernel function $K(\cdot)$ is differentiable with bounded symmetric support and first $(\chi - 1)$ moments begin zero. Moreover, $na_n^d \to \infty$ and $na_n^{2\chi} \to 0$.

(C.5) $n_0/n \to \nu \in (0, 1)$, where $n = n_0 + n_1$.

(C.1) and (C.5) are standard conditions for this type of problem. (C.3) ensures the identifiability of the regression parameters, and (C.4) states the restrictions on the choice of possible kernel functions. For example, when $d = 2$, the kernel function can be chosen to be the Gaussian kernel or the Epanechnikov kernel. Both (C.2) and (C.4) are necessary conditions to prove the asymptotic distribution of $\hat{\beta}$. Obviously, if S_0 is estimated using the Breslow method with discrete X_1 or from the Cox regression method, (C.4) is not needed.

Theorem 1. *Under Conditions* (C1)−(C5), $\|\widehat{\beta} - \beta_0\| \to_{a.s.} 0$.

Table 1. Estimates of optimal bandwidths h_{opt} and bootstrapped MISEs when $X = 0.54$ in Simulation 2 based on 1,000 simulations.

Scenario	$n_1 = n_0$	h_{opt}	$MISE$
Simulation 2	100	0.1706	0.619
	200	0.1212	0.624

Theorem 2. *Under Conditions* (C1)−(C5), $\sqrt{n}(\hat{\beta}-\beta_0)$ *converges in distribution to a mean zero normal random vector as* $n \to \infty$*.*

Theorem 3. *Under Conditions* (C1)−(C5), $\sqrt{n}(\hat{G}(\log t) - G_0(\log t))$ *converges weakly to a zero mean Gaussian process in* $l^\infty([0,1])$*.*

The proofs of Theorems 1−3 are provided in the Appendix. For the proof of Theorem 1, we use the fact that $\hat{S}_0(y;x)$ converges uniformly in (y,x) to $S_0(y;x)$ as n goes to ∞, which is given in Zeng (2004). We then apply Theorem 2.10.3 of van der Vaart and Wellner (1996) and Theorem 5.9 of van der Vaart (1998). The proofs of Theorems 2 and 3 follow the same arguments as in Zeng (2004), and we use the central limit theorems for the empirical process indexed by classes depending on samples (Theorem 2.11.23, van der Vaart and Wellner (1996)).

4. Simulation Studies

Simulation studies were conducted to examine the performance of the proposed method. First, we took as true $G(x) = \exp(\alpha x)$, so from (2.1),

$$ROC_X(t) = \exp[\alpha e^{\beta^T X}\log(t)]. \qquad (4.1)$$

The biomarker values for diseased and non-diseased subjects, Y_1 and Y_0, were generated by

$$Y_0 = -\frac{\log(U_0)}{(\lambda\exp(\gamma^T X))} \text{ and } Y_1 = -\frac{\log(U_1)}{(\lambda\alpha\exp(\gamma^T X + \beta^T X))}, \qquad (4.2)$$

where U_0 and U_1 are Uniform(0, 1) random variables. It is easy to check such (Y_0, Y_1) gives the ROC function specified in (4.1). We used an equal number of diseased and non-diseased subjects but varied the total sample size n from 200 to 400. Additionally, we set the upper detection limit τ as the 95th percentile of the biomarker in the non-diseased group.

We conducted three simulations with different types of covariates. For the first simulation, a binary covariate X was generated from a Bernoulli distribution with probability 0.5, and true parameters in (4.2) were set to $\beta = 0.5$, $\gamma = -0.5$, $\lambda = 1$, and $\alpha = 1.2$. Because X was discrete, we estimated $S_0(y|x)$ using the Breslow estimator given in Remark 1. In the second simulation, we used a

Table 2. Summary results from simulation studies.

		$n_1 = n_0 = 100$				$n_1 = n_0 = 200$			
Par.	True	Est	ASE	SE	CP	Est	ASE	SE	CP
		Simulation Study 1. X: 0 or 1							
β	0.5	0.479	0.309	0.302	95.7	0.517	0.219	0.210	94.9
$G(-2.7)$	0.259	0.255	0.081	0.080	95.6	0.262	0.060	0.060	93.0
$G(-1.3)$	0.522	0.519	0.084	0.082	94.8	0.524	0.060	0.059	94.0
$G(-0.5)$	0.779	0.774	0.063	0.060	96.0	0.778	0.044	0.042	94.3
		Simulation Study 2. X $\sim$ Beta(4,2)							
β	-1	-1.080	0.967	0.936	94.9	-1.057	0.703	0.694	95.8
$G(-0.89)$	0.663	0.630	0.160	0.165	92.2	0.641	0.125	0.128	93.4
$G(-0.542)$	0.763	0.727	0.141	0.142	94.3	0.741	0.106	0.106	94.2
$G(-0.23)$	0.891	0.865	0.098	0.091	95.0	0.877	0.066	0.061	95.6
		Simulation Study 3. $X_1 \sim$ Uniform(0,1), $X_2 \sim$ Uniform(0,1)							
β_1	-1.3	-1.300	0.571	0.543	95.6	-1.297	0.384	0.367	94.7
β_2	-1.8	-1.847	0.587	0.582	94.5	-1.817	0.395	0.391	95.2
$G(-0.162)$	0.321	0.319	0.144	0.150	92.7	0.325	0.108	0.107	94.1
$G(-0.116)$	0.442	0.430	0.151	0.158	93.2	0.441	0.112	0.110	94.2
$G(-0.056)$	0.675	0.649	0.137	0.137	95.2	0.668	0.096	0.091	94.9

continuous covariate generated from Beta(4, 2) distribution, and true parameters were set to $\beta = -1$, $\gamma = -0.2$, $\lambda = 0.5$, and $\alpha = 0.5$. In this simulation, $S_0(y|x)$ was estimated using the smoothed Breslow estimator in (2.4), where the Gaussian kernel function $K(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ was applied. The initial bandwidth was set to $a_n = {n_1}^{-1/3}$, and optimal bandwidths a_n were chosen such that the bootstrapped MISE (2.5) attained is minimum. Specifically, we used the optimal bandwidths at the average values of the covariates as shown in Table 1. Our simulation studies showed, however, that the optimal bandwidths and initial bandwidth $a_n = {n_1}^{-1/3}$ resulted in very similar β estimates. For the third simulation, two continuous covariates, generated as Uniform(0,1) were used with $\beta = (-1.3, -1.8)^T$, $\gamma = (-0.2, -0.25)^T$, $\lambda = (1, 1)^T$, and $\alpha = 7$. We then fit the Cox model to estimate S_0 as described in Remark 2. In all the simulation studies, we obtained $\hat{\beta}$ by solving the log-rank estimating equation (2.3) through bisection search.

Table 2 summarizes the simulation results based on 1,000 replicates. Column "Est" is the average value of the estimates from 1,000 replicates; column "ASE" is the average of the estimated standard errors by the bootstrap method with 1,000 replicates; column "SE" is the standard deviation of the estimates; column "CP" gives the (100×) coverage proportion of the 95% confidence intervals based on asymptotic normality. Overall, the estimates for β are very close to the actual values, and the estimated standard errors using the bootstrap method

Table 3. Estimates of β under the misspecified model for S_0.

		$n_1 = n_0 = 100$			$n_1 = n_0 = 200$		
Par	True	Est	SE	MSE	Est	SE	MSE
β_1	-1.3	-1.271	0.688	0.448	-1.342	0.462	0.215
β_2	-1.8	-1.788	0.724	0.524	-1.840	0.503	0.254

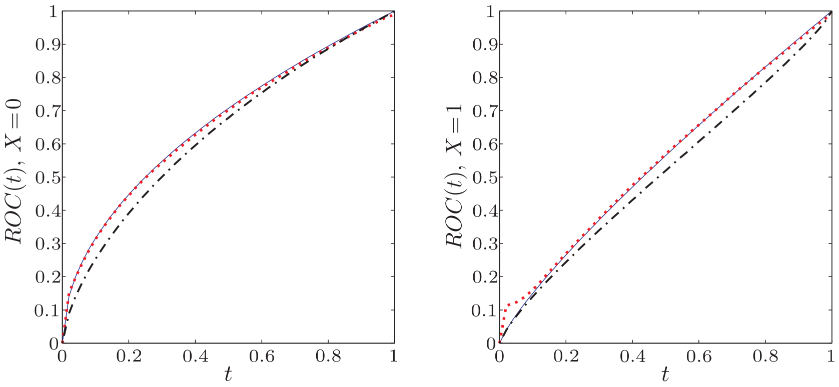

Figure 2. Semiparametric ROC curve ($\cdots$), misspecified parametric ROC curve ($- \cdot - \cdot$), and true ROC curve (—)

approximate the empirical standard errors fairly well. In addition, the coverage proportions of 95% CIs are close to the nominal level of 95% across sample sizes. In the same table, we present the true and estimated G at three fixed points chosen to be the quartiles of the true distribution of $-W_1$ $(= e^{\beta^T X_1} \log S_0(Y_1; X_1))$. For all simulations, the estimated values of G are very close to the actual values at all three points. Moreover, the fitted semiparametric ROC curves obtained from the three simulation studies were extremely close to the true ROC curves (e.g., Figure 2).

We next investigated the robustness of the β estimates conducted in the third simulation study by misspecifying the model for S_0. Specifically, Y_0 was generated from the log-normal model of the form

$$Y_0 = \exp(-1.25 + 0.37X_1 + 0.5X_2 + Z)$$

with $Z \sim \text{Normal}(0, 0.6^2)$, and Y_1 was generated as at (4.2),

$$Y_1 = -\frac{\log(U_1)}{\alpha \exp(\gamma^T X + \beta^T X)},$$

with $\beta = (-1.2, -2)^T$, $\gamma = (-2, 1.95)^T$, and $\alpha = 7$. As shown in Table 3, the estimates of β using the accelerated ROC model are fairly robust to the choice of distributions for S_0.

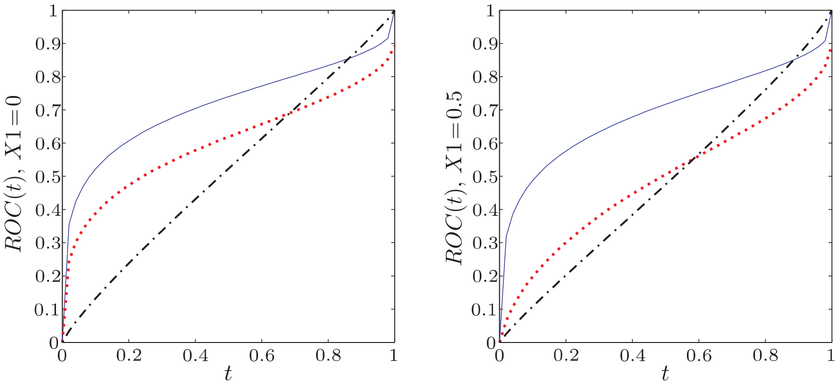

Figure 3. Misspecified semiparametric ROC curve $(\cdots)$, misspecified parametric ROC Curve $(- \cdot - \cdot)$, and true ROC curve (—)

Furthermore, we compared the performance of the proposed semiparametric approach, based on the accelerated ROC model, to the parametric ROC-GLM approach (Alonzo and Pepe (2002))

$$ROC_X(t) = \Phi(\gamma_0 + \gamma_1\Phi^{-1}(t) + \gamma_2 X).$$

Specifically, the three simulation scenarios were considered ($n_1 = n_0 = 100$) and the two approaches were compared with respect to MISEs shown in Table 4. First, we used the data obtained from Simulation 1 and the semiparametric and parametric ROC curves were estimated at $X = 0, 1$. Figure 2 and the MISEs in Table 4 indicate that the fitted semiparametric ROC curve $\hat{G}(e^{0.479X}\log t)$ is slightly closer to the true ROC curve $\exp(0.5e^{0.5X}\log(t))$ than the misspecified parametric ROC curve $\Phi(-0.463 + 1.449\Phi^{-1}(t) + 0.56X)$ at both points.

Second, we simulated the biomarker values from

$$Y_0 = \exp(-1 + 1.4X + \epsilon),\ \ Y_1 = -\frac{\log(U)}{\exp(-0.2 + 0.2X)},$$

where X and U were generated from Uniform$(0,1)$ and the ϵ were from $N(0, 0.6^2)$. The induced ROC curve is $ROC_X(t) = \exp[-\exp(-1.2 + 1.6X - 0.6\Phi^{-1}(t))]$, which is neither parametric ROC-GLM nor our accelerated ROC model. Figure 3 suggests that the proposed semiparametric ROC curve $\hat{G}(e^{1.39X}\log t)$ is closer to the truth than is the parametric curve $\Phi(0.058 + 0.916\Phi^{-1}(t) - 0.24X)$ at both points. Interestingly, the ROC-GLM approach gives very different estimates from the truth.

Finally, we simulated the biomarker values from

$$Y_1 = 3 + 3.85X + \epsilon_1, \quad Y_0 = 1.5 + 2X + \epsilon_0,$$

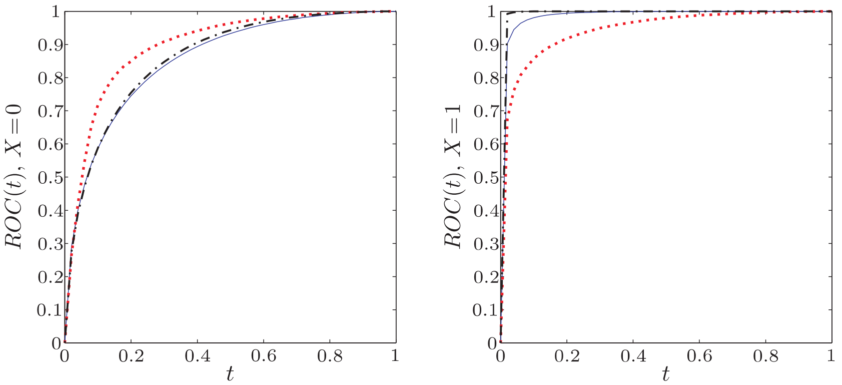

Figure 4. Misspecified semiparametric ROC curve ($\cdots$), parametric ROC curve ($- \cdot - \cdot$), and true ROC curve (—)

Table 4. Average of estimated MISE based on 1,000 simulated datasets by the proposed semiparametric and parametric ROC-GLM approaches.

Figure	Model	MISE (Covariate)	
1	Semiparametric	2.32×10^{-5} $(X = 0)$	1.94×10^{-4} $(X = 1)$
	Misspecified ROC-GLM	0.0013 $(X = 0)$	0.0014 $(X = 1)$
2	Misspecified semiparametric	0.0131 $(X = 0)$	0.0451 $(X = 0.5)$
	Misspecified ROC-GLM	0.0599 $(X = 0)$	0.0635 $(X = 0.5)$
3	Misspecified semiparametric	0.0035 $(X = 0)$	0.0044 $(X = 1)$
	ROC-GLM	6.75×10^{-5} $(X = 0)$	2.72×10^{-4} $(X = 1)$

where X is a Bernoulli random variable with probability 0.5, and ϵ_1 and ϵ_0 have the standard normal distributions. Then the corresponding covariate-specific curve is

$$ROC_X(t) = \Phi(1.5 + \Phi^{-1}(t) + 1.85X),$$

which is exactly the form in the ROC-GLM approach. Undoubtedly, the ROC-GLM approach fits data well but our accelerated ROC model, even though biased, is still not far from the truth; see Figure 4.

5. Application

We illustrate our approach by utilizing a prostate cancer dataset. Prostate-specific antigen (PSA) is a protein produced by the prostate gland, and the PSA test measures the level of PSA in the blood. Most healthy men have PSA levels under 4 nanograms per milliliter (ng/mL) of blood, and the chance of having prostate cancer rises as the PSA level increases. PSA occurs in two major forms in the blood. One form is attached to blood proteins while the other freely circulates. The free PSA is the ratio of how much PSA circulates freely compared to the total PSA level. Low free PSA may indicate prostate cancer, and most men with prostate cancer have a free PSA below 15%. According to the American Cancer Society and National Cancer Institute, men with free PSA at 7% or lower should undergo a biopsy as a precaution. We used a dataset of 71 prostate cancer subjects and 68 controls, all of whom participated in the Beta-Carotene and Retinol Efficacy Trial (CARET), a randomized lung cancer prevention study including 12,025 men (Goodman et al. (1993); Etzioni et al. (1999)).

The objective of this analysis was to evaluate the capacity of free PSA levels to discriminate men with prostate cancer from those with no malignancy prior to the onset of clinical symptoms. Subjects who participated in CARET had serum samples drawn at baseline and at two-year intervals thereafter. Blood samples drawn after a diagnosis of prostate cancer were excluded from this analysis, leaving 1-7 blood samples per subject. Previous studies have suggested that age and the time at which PSA was measured prior to diagnosis may affect the detection of prostate cancer. Let X be the age PSA was measured, and T be the time (years) from the onset of symptoms to the time at which the serum sample was drawn, so that time is negative and increases to 0, the time of clinical diagnosis. Accuracy would be expected to increase with increasing values of T. The average age of participants was 63.7 (range from 46.7 to 80.8), and the average time was -3.06 years (range from -9.008 to -0.003 yrs). We fitted an accelerated ROC model adjusting for age X and time T,

$$ROC_{T,X}(u) = G(\mathrm{e}^{\beta_x X + \beta_t T} \log u).$$

Since each subject may have more than one measurement, the estimating equations in Remark 3 were solved for estimating β's.

We found $\hat{\beta}_x = 0.0485$ with SE 0.0248 (p-value 0.0505) and $\hat{\beta}_t = -0.0587$ with SE 0.0442 (p-value 0.1841). The positive coefficient for age suggested that discrimination of disease is more efficient in younger men, and the negative coefficient for time implies that discrimination improves as PSA levels are measured closer to actual diagnosis although the time T was not found to be significant.

The estimated AUCs based on the accelerated ROC model were 0.8579, 0.8103, and 0.7623 using the median time $T = -2.82$ and the respective mean ages for groups age ≤ 61, $61<$ age ≤ 65, and age ≥ 65. On the other hand, the estimated parametric binormal model was $ROC_{T,X}(u) = \Phi(4.82 + 0.715\Phi^{-1}(t) - 0.051X + 0.166T)$ and the corresponding estimated AUCs were 0.8788, 0.8220, and 0.7634. Our AUC estimates turned out to be closer to the empirical AUCs which were calculated as 0.8575, 0.8062, and 0.7527. Additionally, Figure 5 shows that the fitted ROC curves based on the accelerated ROC model matches well with the empirical curves demonstrating the model adequacy of the proposed method; refer to Remark 4.

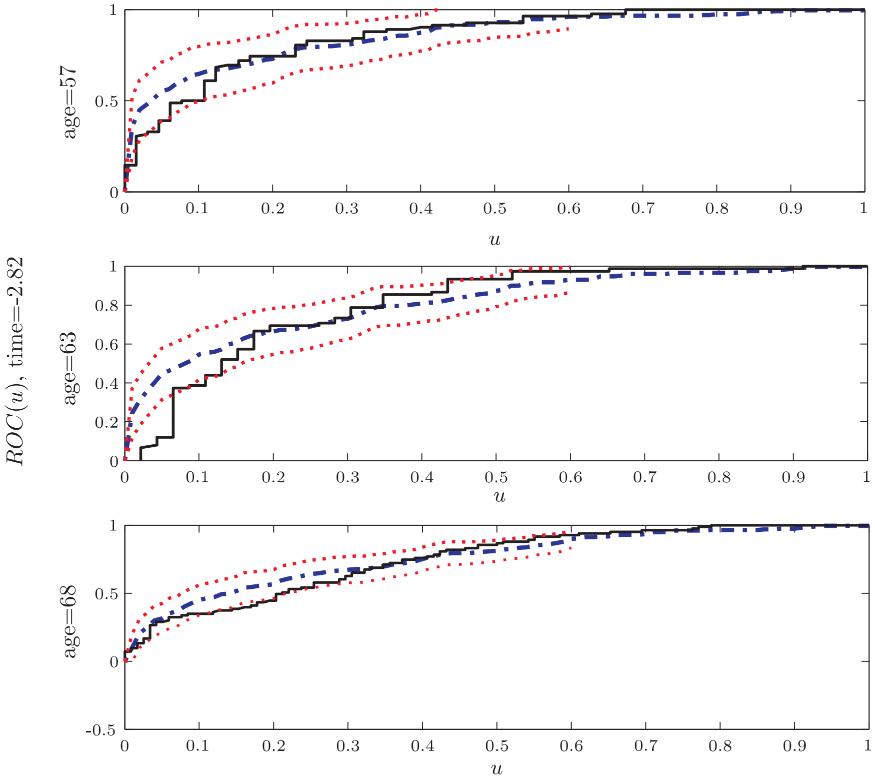

Figure 5. Estimated Semiparametric ROC Curves for PSA Adjusted for Age and Time ($- \cdot - \cdot$) and their 95 % Confidence Regions over [0, 0.6] ($\cdot\cdot\cdot$), Corresponding Empirical ROC Curves (—).

6. Discussion

We have proposed a semiparametric method to assess the accuracy of biomarkers by adjusting for covariates that could influence their performance. We developed an accelerated ROC model by employing the properties of the AFT model and showed that the parameter estimate of β can be conducted by solving the log-rank estimating equation. The function G was estimated using the empirical distribution of $Z_{ij}e^{\beta^T X_{i1}}$ without making any assumptions about the distribution of G. We demonstrated that $\hat{G}$ derived using the Kaplan-Meier estimator of $Z_{ij}e^{\beta^T X_{i1}}$ is a good fit to the true function G. The bootstrapping method was used for inference, and the asymptotic properties of $\hat{\beta}$ and $\hat{G}$ were presented.

In our proposed method, the parameter estimates of covariates based on the log-rank estimating equations may not be efficient. Other estimation approaches such as described by Jin et al. (2003) and Jin, Lin, and Ying (2006) can be applied. For future work, we will examine whether a semiparametrically efficient estimator can be obtained.

Both our model and Pepe's (1997; 2000) directly model the effects of covariates on the ROC curves. These two models are in parallel to the AFT model and the proportional hazards model in the survival context. In survival analysis, there have been a number of approaches developed for model diagnostics and model checking. It is interesting to see how those approaches can be extended to the ROC regression models. Another possibility is to consider an even more general model by assuming

$$ROC_X(t) = G(\exp\{X^T\beta\}h(t)),$$

where both G and h are unknown functions. This general model includes our model and Pepe's model as special cases with $h(t) = \log t$ and $G(\cdot)$ a known link, respectively. However, it is unclear how reliably the model parameters can be estimated in practice.

Appendix

A.1. Proof of Theorem 1

With direct calculations, (C.3) implies that

$$\sigma_1 \equiv -\frac{\partial}{\partial\beta} E\left[\frac{Q_1(\log Z_1 + \beta^T X_1)}{Q_0(\log Z_1 + \beta^T X_1)}\right]$$

is positive for $\beta \in \mathcal{B}$, where $Q_1(x) = E[X_1 I(\log Z_1 + \beta^T X_1 \geq x)]$ and $Q_0 = E[I(\log Z_1 + \beta^T X_1 \geq x)]$. Therefore, β_0 must be the unique solution to

$$E\left[\left(X_1 - \frac{Q_1(\log Z_1 + \beta^T X_1)}{Q_0(\log Z_1 + \beta^T X_1)}\right)\right] = 0,$$

We introduce some notation. We use $\mathcal{P}_{n1}$ and $\mathcal{P}_1$ to denote the empirical measure and expectation based on i.i.d. observations in the diseased group, $(Y_{i1}, X_{i1}), i = 1, \ldots, n_1$. Similarly, we use $\mathcal{P}_{n0}$ and $\mathcal{P}_0$ to denote the empirical measure and expectation based on i.i.d. observations in the non-diseased group, $(Y_{j0}, X_{j0}), j = 1, \ldots, n_0$. Moreover, $\mathcal{G}_{n1}$ and $\mathcal{G}_{n0}$ denote the empirical processes $\sqrt{n_1}(\mathcal{P}_{n1} - \mathcal{P}_1)$ and $\sqrt{n_0}(\mathcal{P}_{n0} - \mathcal{P}_0)$, respectively. Thus, by definition, $\hat{\beta}$ should solve

$$0 = \mathcal{P}_{n1}\left[\left\{X_1 - \frac{\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \hat{\beta}^T X_{i1} \geq \log \hat{Z}_1 + \hat{\beta}^T X_1)X_{i1}}{\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \hat{\beta}^T X_{i1} \geq \log \hat{Z}_1 + \hat{\beta}^T X_1)}\right\}\right].$$

We show the consistency of $\hat{\beta}$. First, conditional on non-diseased data, $(\hat{Z}_{i1}, X_{i1})$ are i.i.d. Therefore, the class

$$\mathcal{F} \equiv \left\{I(x \geq \log \hat{Z}_1 + \beta^T X_1) : x \in (-\infty, \infty), \beta \in \mathcal{B}\right\}$$

is the VC-class, so is Donsker. Note that the random functions

$$n^{-1}\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \beta^T X_{i1} \geq \log \hat{Z}_1 + \beta^T X_1)X_{i1},$$

$$n^{-1}\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \beta^T X_{i1} \geq \log \hat{Z}_1 + \beta^T X_1),$$

$$E^*\left[\frac{I(\log \hat{Z}_1 + \beta^T X_1 \geq \log \hat{Z}_1^* + \beta^T X_1^*)X_1}{n^{-1}\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \beta^T X_{i1} \geq \log \hat{Z}_1^* + \beta^T X_1^*)}\right],$$

where here and later, E^* and E^{**} denote the expectation with respect to those random variables with asterisk and double asterisk respectively, can be expressed as the limit of the convex combinations of $\mathcal{F}$ and are bounded from above. Thus, they belong to *sconv*$\mathcal{F}$, which is a Donsker class by Theorem 2.10.3 of van der Vaart and Wellner (1996). Therefore, by the Glivenko-Cantelli Theorem, it is easy to see that

$$\sup_\beta \Bigg| \mathcal{P}_{n1}\left[\left\{X_1 - \frac{\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \beta^T X_{i1} \geq \log \hat{Z}_1 + \beta^T X_1)X_{i1}}{\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \beta^T X_{i1} \geq \log \hat{Z}_1 + \beta^T X_1)}\right\}\right]$$
$$-E\left[\left\{X_1 - \frac{E^*[I(\log \hat{Z}_1^* + \beta^T X_1^* \geq \log \hat{Z}_1 + \beta^T X_1)X_1^*]}{E^*[I(\log \hat{Z}_1^* + \beta^T X_1^* \geq \log \hat{Z}_1 + \beta^T X_1)]}\right\}\right] \Bigg| \to_{a.s.} 0.$$

Furthermore, as n goes to ∞, $\hat{S}_0(y;x)$ converges uniformly in (y,x) to $S_0(y;x)$, as shown in Zeng (2004). Thus, the limit function

$$E\left[\left\{X_1 - \frac{E^*[I(\log \hat{Z}_1^* + \beta^T X_1^* \geq \log \hat{Z}_1 + \beta^T X_1)X_1^*]}{E^*[I(\log \hat{Z}_1^* + \beta^T X_1^* \geq \log \hat{Z}_1 + \beta^T X_1)]}\right\}\right]$$

converges uniformly in β to

$$E\left[\left\{X_1 - \frac{Q_1(\log Z_1 + \beta^T X_1)}{Q_0(\log Z_1 + \beta^T X_1)}\right\}\right].$$

The latter has a unique minimum zero at β_0 by (C.1). Additionally, it satisfies the separability at β_0 by (C.3). Therefore, by Theorem 5.9 of van der Vaart (1998), $\hat{\beta}$ converges almost surely to β_0.

A.2. Proof of Theorem 2

We derive the asymptotic distribution of $\hat{\beta}$. From

$$\mathcal{P}_{n1}\left[\left\{X_1 - \frac{\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \hat{\beta}^T X_{i1} \geq \log \hat{Z}_1 + \beta^T X_1)X_{i1}}{\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \hat{\beta}^T X_{i1} \geq \log \hat{Z}_1 + \hat{\beta}^T X_1)}\right\}\right] = 0 \qquad \text{(A.1)}$$

if we define

$$\hat{Q}_1(x) = E[X_1 I(\log \hat{Z}_1 + \hat{\beta}^T X_1 \geq x)], \quad \hat{Q}_0(x) = E[I(\log \hat{Z}_1 + \hat{\beta}^T X_1 \geq x)],$$

then we obtain

$$\begin{aligned}
&\mathcal{G}_{n1}\left[\left\{X_1 - \frac{\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \hat{\beta}^T X_{i1} \geq \log \hat{Z}_1 + \hat{\beta}^T X_1) X_{i1}}{\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \hat{\beta}^T X_{i1} \geq \log \hat{Z}_1 + \hat{\beta}^T X_1)}\right\}\right] \\
&\quad -\mathcal{G}_{n1} E^*\left[\frac{I(\log \hat{Z}_1 + \hat{\beta}^T X_1 \geq \log \hat{Z}_1^* + \hat{\beta}^T X_1^*) X_1}{n^{-1}\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \hat{\beta}^T X_{i1} \geq \log \hat{Z}_1^* + \hat{\beta}^T X_1^*)}\right] \\
&\quad +\mathcal{G}_{n1} E^*\left[\frac{I(\log \hat{Z}_1 + \hat{\beta}^T X_1 \geq \log \hat{Z}_1^* + \hat{\beta}^T X_1^*)\hat{Q}_1(\log \hat{Z}_1^* + \hat{\beta}^T X_1^*)}{n^{-1}\sum_{i=1}^{n_1} I(\log \hat{Z}_{i1} + \hat{\beta}^T X_{i1} \geq \log \hat{Z}_1^* + \hat{\beta}^T X_1^*)\hat{Q}_0(\log \hat{Z}_1^* + \hat{\beta}^T X_1^*)]}\right] \\
&= -\sqrt{n_1} E\left[\left\{X_1 - \frac{E^*[I(\log \hat{Z}_1^* + \hat{\beta}^T X_1^* \geq \log \hat{Z}_1 + \hat{\beta}^T X_1) X_1^*]}{E^*[I(\log \hat{Z}_1^* + \hat{\beta}^T X_1^* \geq \log \hat{Z}_1 + \hat{\beta}^T X_1)]}\right\}\right].
\end{aligned}$$

From the Donsker theorem, we have

$$-\sqrt{n_1} E\left[\left\{X_1 - \frac{\hat{Q}_1(\log \hat{Z}_1 + \hat{\beta}^T X_1)}{\hat{Q}_0(\log \hat{Z}_1 + \hat{\beta}^T X_1)}\right\}\right] = \mathcal{G}_{n1} g(X_1, Z_1; \beta_0) + o_p(1), \qquad \text{(A.2)}$$

where

$$\begin{aligned}
g(X_1, Z_1; \beta_0) = &\left\{X_1 - \frac{Q_1(\log Z_1 + \beta_0^T X_1)}{Q_0(\log Z_1 + \beta_0^T X_1)}\right\} \\
&-E^*\left[\frac{I(\log Z_1 + \beta^T X_1 \geq \log Z_1^* + \beta^T X_1^*) X_1}{Q_0(\log Z_1^* + \beta_0^T X_1^*)}\right] \\
&+E^*\left[\frac{I(\log Z_1 + \beta^T X_1 \geq \log Z_1^* + \beta^T X_1^*) Q_1(\log Z_1^* + \beta_0^T X_1^*)}{Q_0(\log Z_1^* + \beta_0^T X_1^*)^2}\right].
\end{aligned}$$

On the other hand, from (C.2),

$$\hat{Q}_0(x) = E\left[P\left(Y_1 \geq \hat{H}_0^{-1}(e^{x-\hat{\beta}^T X_1}; X_1)\middle| X_1\right)\right],$$

where $\hat{H}_0^{-1}(y; x)$ denotes the inverse of $H_0(y; x) \equiv -\log S_0(y; x)$ for given x. Thus, if $f_1(y|x)$ is the conditional density of Y_1 given X_1, then

$$\begin{aligned}
&\hat{Q}_0(x) \\
&= -E\left[f_1\left(H_0^{-1}(e^{x-\hat{\beta}^T X_1}; X_1)\middle| X_1\right)\left(\hat{H}_0^{-1}(e^{x-\hat{\beta}^T X_1}; X_1) - H_0^{-1}(e^{x-\hat{\beta}^T X_1}; X_1)\right)\right] \\
&\quad +E\left[P\left(Y_1 \leq H_0^{-1}(e^{x-\hat{\beta}^T X_1}); X_1)\middle| X_1\right)\right] + o(1).
\end{aligned}$$

By slightly modifying Lemma 3.9.20 of van der Vaart and Wellner (1996), we can show

$$\begin{aligned}
&\hat{H}_0^{-1}(e^{x-\hat{\beta}^T X_1}; X_1) - H_0^{-1}(e^{x-\hat{\beta}^T X_1}; X_1) \\
&\quad = -\frac{\hat{H}_0(H_0^{-1}(e^{x-\hat{\beta}^T X_1}; X_1); X_1) - H_0(H_0^{-1}(e^{x-\hat{\beta}^T X_1}; X_1); X_1)}{H_0'(H_0^{-1}(e^{x-\beta_0^T X_1}; X_1); X_0 = X_1)} + o(1),
\end{aligned}$$

and that it holds uniformly in $x, \hat{\beta}$, and X_1. Moreover, since $\hat{H}_0(\cdot; x)$ converges to $H_0(\cdot; x)$ in $D[0,\tau]$ uniformly in x, we obtain

$$\begin{aligned}
&\hat{Q}_0(\log \hat{Z}_1^* + \beta_1^T X_1^*) \\
&= E\Bigg[\frac{f_1\left(H_0^{-1}(Z_1^* e^{\beta_0^T X_1^* - \beta_0^T X_1}; X_1)|X_1\right)}{H_0'(H_0^{-1}(Z_1^* e^{\beta_0^T X_1^* - \beta_0^T X_1}; X_1); X_0 = X_1)} \\
&\quad \times \left(\hat{H}_0(H_0^{-1}(Z_1^* e^{\beta_0^T X_1^* - \beta_0^T X_1}; X_1); X_1) - H_0(H_0^{-1}(Z_1^* e^{\beta_0^T X_1^* - \beta_0^T X_1}; X_1); X_1)\right)\Bigg] \\
&\quad + E\left[P\left(Y_1 \geq H_0^{-1}(\hat{Z}_1^* e^{\hat{\beta}_0^T X_1^* - \hat{\beta}^T X_1}; X_1)\middle| X_1\right)\right] + o(1).
\end{aligned}$$

The last term on the right-hand side can be further approximated by

$$\begin{aligned}
&E\left[P\left(Y_1 \geq H_0^{-1}(Z_1^* e^{\beta_0^T X_1^* - \beta_0^T X_1}; X_1)\middle| X_1\right)\right] \\
&\qquad E\Bigg[\frac{f_1(H_0^{-1}(Z_1^* e^{\beta_0^T X_1^* - \beta_0^T X_1}; X_1)|X_1)}{H_0'(H_0^{-1}(Z_1^* e^{\beta_0^T X_1^* - \beta_0^T X_1}; X_1); X_0 = X_1)} Z_1^* e^{\beta_0^T X_1^* - \beta_0^T X_1} \\
&\qquad \times \left\{\frac{\hat{H}_0(Y_1^*; X_1^*) - H_0(Y_1^*; X_1^*)}{H_0(Y_1^*; X_1^*)} + (\hat{\beta} - \beta_0)(X_1^* - X_1)\right\}\Bigg].
\end{aligned}$$

Similarly, we can expand the numerator term in the left-hand side of (A.2) to eventually obtain that (A.2) is equivalent to

$$\begin{aligned}
&\mathcal{G}_{n1} g(X_1, Z_1; \beta_0) + o_p(1) \\
&= \sqrt{n_1}\sigma_1(\hat{\beta} - \beta_0) + \sqrt{n_1} E^*\Big[\sigma_2(Z_1, X_1, X_1^*)\Big(\hat{H}_0(H_0^{-1}(Z_1 e^{\beta_0^T X_1 - \beta_0^T X_1^*}; X_1^*); X_1^*) \\
&\quad - H_0(H_0^{-1}(Z_1 e^{\beta_0^T X_1 - \beta_0^T X_1^*}; X_1^*)\Big)\Big] \\
&\quad + \sqrt{n_1} E\left[\sigma_3(Y_1, X_1)\left(\hat{H}_0(Y_1; X_1) - H_0(Y_1; X_1)\right)\right], \qquad \text{(A.3)}
\end{aligned}$$

for some differentiable functions σ_2 and σ_3. Particularly, σ_1 has the same expression as given in (C.3) with $\beta = \beta_0$, so σ_1 is non-singular.

Using the same arguments as in Zeng (2004) and (C.4), we can show that, uniformly in x and $y \in [0, \tau]$,

$$
\begin{aligned}
&(\hat{H}_0(y; x) - H_0(y, x)) \\
&= \left\{ (\mathcal{P}_{n0} - \mathcal{P}_0) \left[\frac{I(Y_0 \le y)(n_0 a_n^d)^{-1} K_{a_n}(X_0 - x)}{(n_0 a_n^d)^{-1} \sum_{k=1}^{n_0} I(Y_{k0} \ge Y_0) K_{a_n}(X_{k0} - x)} \right] - (\mathcal{P}_{n0} - \mathcal{P}_0) \right. \\
&\left. E^* \left[\frac{I(Y_0 \ge Y_0^*)(n_0 a_n)^{-1} K_{a_n}(X_0^* - x)}{(n_0 a_n^d)^{-1} \sum_{k=1}^{n_0} I(Y_{k0} \ge Y_0^*) K_{a_n}(X_{k0} - x) E^{**}[I(Y_0^{**} \ge Y_0^*) K_{a_n}(X_0^{**} - x)]} \right] \right\} \\
&\quad + O(a_n^{\chi}) \\
&\equiv (\mathcal{P}_{n0} - \mathcal{P}_0) q_n(y, x, Y_0, X_0) + o_p(a_n^{\chi}).
\end{aligned}
$$

We plug the above expression into (A.4), then (A.2). From (C.4), we obtain

$$
\begin{aligned}
&\mathcal{G}_{n1} g(X_1, Z_1; \beta_0) + o_p(1) = \sqrt{n_1} \sigma_1 (\hat{\beta} - \beta_0) \\
&\quad + \sqrt{n_1} (\mathcal{P}_{n0} - \mathcal{P}_0) E \left[\sigma_2(Z_1, X_1, X_1^*) q_n(H_0^{-1}(Z_1^* e^{-\beta_0^T X_1 + \beta_0^T X_1^*}; X_1^*), X_1^*, Y_0, X_0) \right] \\
&\quad + \sqrt{n_1} (\mathcal{P}_{n0} - \mathcal{P}_0) E \left[\sigma_3(Y_1, X_1) q_n(Y_1, X_1, Y_0, X_0) \right]. \qquad \text{(A.4)}
\end{aligned}
$$

Finally, we apply Theorem 2.11.23 in van der Vaart and Wellner (1996) to the last two terms in the right-hand side of (A.4). Particularly, their conditions are satisfied by observing that after integration by parts, both

$$
E \left[\sigma_2(Z_1, X_1, X_1^*) q_n(H_0^{-1}(Z_1 e^{\beta_0^T X_1 - \beta_0^T X_1^*}; X_1^*), X_1^*, Y_0, X_0) \right]
$$

and $E[\sigma_3(Y_1, X_1) q_n(Y_1, X_1, Y_0, X_0)]$ converge uniformly in (Y_0, X_0) to

$$
E \left[\sigma_2(Z_1, X_1, X_1^*) q(H_0^{-1}(Z_1 e^{\beta_0^T X_1 - \beta_0^T X_1^*}; X_1^*), X_1^*, Y_0, X_0) \middle| X_1^* = X_0 \right]
$$

and $E[\sigma_3(Y_1, X_0) q(Y_1, X_0, Y_0, X_0) \Big| X_1 = X_0]$, respectively, where $q(Y_1, X_1, Y_0, X_0)$ $= I(Y_0 \le Y_1)/S_0(Y_0 | X = x)$. Furthermore, they have bounded total variation in Y_0 uniformly in X_0, and are Lipschitz continuous in X_0. The latter implies the entropy condition in Theorem 2.11.23. Therefore, combining the above results and the non-singularity of σ_1 in (A.4), we obtain

$$
\sqrt{n}(\hat{\beta} - \beta_0) = \mathcal{G}_{n1} \sigma_1^{-1} g_1(Y_1, X_1; \beta_0) + \mathcal{G}_{n2} \sigma_1^{-1} g_2(Y_0, X_0) + o_p(1),
$$

where

$$
\begin{aligned}
&g_2(Y_0, X_0) \\
&\quad = -E \left[\sigma_2(Z_1, X_1, X_1^*) q(H_0^{-1}(Z_1 e^{\beta_0^T X_1 - \beta_0^T X_1^*}; X_1^*), X_1^*, Y_0, X_0) \middle| X_1^* = X_0 \right] \\
&\qquad - E \left[\sigma_3(Y_1, X_0) q(Y_1, X_0, Y_0, X_0) \middle| X_1 = X_0 \right].
\end{aligned}
$$

Hence, $\sqrt{n}(\hat{\beta}-\beta_0)$ converges in distribution to a normal distribution with mean zero and variance $(1-\nu)\mathrm{Var}\,(g_1)+\nu\mathrm{Var}\,(g_2)$.

Remark A2.1 When X's take discrete values, the proof can be much simplified. Particulary, we can set $a_n = 1/n$ and $K_{a_n}(x) = I(x=0)$ in the above arguments.

Remark A2.2 When $S_0(y|x)$ is estimated by the Cox model, the only difference is in the expressions of $\hat{H}_0(y;x) - H_0(y,x)$; the influence function $q_n(y,x,Y_0,X_0)$ is given by the influence function of $\exp[-\hat{\Lambda}(y)\exp(\hat{\gamma}^T x)]$, where $(\hat{\Lambda},\hat{\gamma})$ is the nonparametric maximum likelihood estimator in the Cox model.

A.3. Proof of Theorem 3

The asymptotic property of $\hat{G}(t)$ follows the same expansion as the proof of Theorem 2 but we utilize the differentiability of the product-limit function. Let S_W denote the survival function for W_1, and H_W the cumulative hazard function of W_1. We have

$$\begin{aligned}
&\hat{G}(t) - G_0(t)\\
&\quad = -G_0(t)(\mathcal{P}_{n1}-\mathcal{P}_1)\left[\frac{I(W_1\le -t)}{E^*[I(W_1^*\ge W_1)]} - E^*\left\{\frac{I(W_1\ge W_1^*)I(W_1^*\le -t)}{S_W(W_1^*)^2}\right\}\right]\\
&\qquad -G_0(t)\left\{E\left[\frac{I(\hat{W}_1\le -t)}{E^*[I(\hat{W}_1^*\ge \hat{W}_1)]}\right] - H_W(t)\right\} + o_p(n^{-1/2}).
\end{aligned}$$

We further expand the second term in the right-hand side as in the previous section to obtain

$$\begin{aligned}
&\tilde{\sigma}_1(\hat{\beta}-\beta_0) + E\Big[\tilde{\sigma}_2(Z_1,X_1,X_1^*)\Big(\hat{H}_0(H_0^{-1}(Z_1e^{\beta_0^T X_1-\beta_0^T X_1^*};X_1^*);X_1^*)\\
&\qquad -H_0(H_0^{-1}(-Z_1e^{\beta_0^T X_1-\beta_0^T X_1^*};X_1^*)\Big)\Big]\\
&\qquad +E\Big[\tilde{\sigma}_3(Y_1,X_1)\Big(\hat{H}_0(Y_1;X_1)-H_0(Y_1;X_1)\Big)\Big] + o_p(n^{-1/2}).
\end{aligned}$$

Hence, from the same arguments as in Theorem 2, we obtain

$$\begin{aligned}
&\hat{G}(t) - G_0(t)\\
&\quad = -G_0(t)(\mathcal{P}_{n1}-\mathcal{P}_1)\left[\frac{I(W_1\le -t)}{E^*[I(W_1^*\ge W_1)]} - E^*\left\{\frac{I(W_1\ge W_1^*)I(W_1^*\le -t)}{S_W(W_1^*)^2}\right\}\right]\\
&\qquad +\tilde{\sigma}_1(\hat{\beta}-\beta_0) + (\mathcal{P}_{n0}-\mathcal{P}_0)g_3(Y_0,X_0) + o_p(n^{-1/2})
\end{aligned}$$

for some $g_3(Y_0,X_0)$. Therefore, $\sqrt{n}(\hat{G}(t)-G_0(t))$ converges in distribution to a Gaussian process in $l^\infty[0,\tau]$, and the covariance function is equal to the covariance of

$$\sqrt{1-\nu}\left[\frac{I(W_1 \le -t)}{E^*[I(W_1^* \ge W_1)]} - E^*\left\{\frac{I(W_1 \ge W_1^*)I(W_1^* \le -t)}{S_W(W_1^*)^2}\right\}\right]$$
$$+\sqrt{1-\nu}\tilde{\sigma}_1\sigma_1^{-1}g_1(Y_1, X_1; \beta_0) + \sqrt{\nu}\tilde{\sigma}_1\sigma_1^{-1}g_2(Y_0, X_0) + \sqrt{\nu}g_3(Y_0, X_0).$$

References

Alonzo, T. A. and Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics* **3**, 421-433.

Beam, C. A. (1995). Random-effects models in the receiver operating characteristic curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches, and issues. *Academic Radiology* **2**, 4-13.

Cai, T. and Pepe, M. S. (2002). Semi-parametric ROC analysis to evaluate biomarkers for disease. *J. Amer. Statist. Assoc.* **97**, 1099-1107.

Dorfman, D. D., Berbaum, K. S., and Metz, C. E. (1992). ROC rating analysis: generalization to the population of readers and cases with the jackknife method. *Invest. Radiol.* **27**, 723-731.

Etzioni, R., Pepe. M. S., Longton, G., Hu, C. and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Med. Decis. Making.* **19**, 242-251.

Gatsonis, C. A. (1995). Random effects models for diagnostic test accuracy. *Academic Radiology* **2**, S14-S21.

Goodman, G., Omenn, G. S., Thornquist, M., Lund, B., Metch, B., and Gylys-Colwell, I. (1993). The Carotene and Retinol Efficacy Trial (CARET) to prevent lung cancer in high-risk populations: pilot study with cigarette smokers. *Cancer Epidemiol. Biomarkers Prev.* **2**, 389-396.

Hanley, H. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Crit. Rev. Diagn. Imag.* **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. *J Roy. Statist. Soc. Ser. C* **48**, 533-551.

Hellmich, M., Abrams, K. R., Jones, D. R., and Lambert, P. C. (1998). A Bayesian approach to a general regression model for ROC curves. *Med. Decis. Making.* **18**, 436-443.

Ishwaran, H. and Gatsonis, C. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canad. J. Statist.* **28**, 731-750.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341-353.

Jin, Z., Lin, D. Y. and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika* **93**, 147-161.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

Obuchowski, N. A. and Rockette, H. E. (1995). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. *Commun. Stat. B.-Simul.* **24**, 285-308.

Peng, F. and Hall, W. J. (1996). Bayesian analysis of ROC curves using Markov-Chain Monte Carlo methods. *Med. Decis. Making.* **16**, 404-411.

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

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

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

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

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

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of ROC curves. *Statist. Medicine* **8**, 1277-1290.

Tosteson, A. and Begg, C. (1998). A general regression methodology for ROC curve estimations. *Med. Decis. Making.* **8**, 204-215.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Wang, Q. and Shen, J. (2008). Estimation and confidence bands of a conditional survival function with censoring indicators missing at random. *J. Multivariate Anal.* **99**, 928-948.

Zeng, D. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *Ann. Statist.* **32**, 1533-1555.

Zeng, D. and Lin, D. Y.(2007). Efficient estimation for the accelerated failure time model. *J. Amer. Statist. Assoc.* **102**, 1387-1396.

Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine.* Wiley, New York.

Department of Biostatistics and Center for Statistical Sciences, Brown University, Providence, Rhode Island 02912, U.S.A.

E-mail: ekim@stat.brown.edu

Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina 27599, U.S.A.

E-mail: dzeng@bios.unc.edu

(Received November 2011; accepted August 2012)